Thesis:

Quantifying Oxidative Folding

by Lukas Alexander Rettenbacher, Msc

Supervisor:

Dr. Tobias von der Haar

Host Institute:

University of Kent, School of Biosciences

Year of submission:

2023

Word Count / Page Numbers:

42300 / 143

Abstract:

Proteins are involved in almost every known biological process. Their immensely diverse pool of functionalities is largely determined by their amino acid sequence, their three-dimensional structure and by additional modifications known as post translational modifications (PTMs). One of these PTMs is the disulfide bond (DSB) which can be formed between two cysteine amino acids via covalent bonding of two sulfur atoms. DSBs can be found in many important proteins such as antibodies and the processes involved in their formation have been intensely studied for decades. During this time, many – mostly qualitative - aspects of their formation and prevalence have been researched. The research described in this thesis combine parts of this immense existing knowledge and elevate our understanding of the quantitative aspects of DSB formation. The thesis introduction summarises much of this existing knowledge and current research trends are further outlined in a published review. In Chapter 1 the quantitative formation of DSBs in *Escherichia coli* has been modelled in order to describe and predict both host proteome and recombinant protein DSB formation. Chapter 2 expands our understanding of the DSB forming machinery in the important recombinant protein production host *Komagataella phaffii* (syn. *Pichia pastoris*). In the final Chapter, protein structure predictions by AlphaFold are used for predicting both qualitative and quantitative DSB levels in several model organisms.

Table of Contents

Foreword

The title of my PhD project as advertised in 2019 was “Modelling recombinant protein folding and secretion” as part of the Marie Curie ITN Grant SECRETERS with the overall title “A new generation of microbial expression hosts and tools for the production of biotherapeutics and high-value enzymes”. Within this consortium my goal was to investigate the cellular processes which are central to protein folding and overall protein production in microbial cells. More specifically in *Escherichia coli*, *Bacillus subtilis* and *Komagataella phaffii* which are the three microbes of interest in the SECRETERS consortium.

As both the wording of my own project title as well as the SECRETERS title suggest, there was a lot of freedom for developing my project into different directions. However, within the context of the other consortia projects, it was clear that focusing on the oxidative folding aspect of the overall topic of protein folding would be best suited for my work.

Within the SECRETERS consortia it was envisioned that experimental data would be obtained by other early stage researchers (ESRs) and subsequently handed over to me for further computational and modelling based research. But generating clean and usable data in the laboratory takes time and therefore I started out by finding alternative, readily available data sources for the purpose of investigating oxidative protein folding. This decision turned out to be an important one. Seven months into my project the SARS-CoV-2 pandemic paralysed the world and laboratory-based research was shut down. Not only did this further delay any inter-consortia collaboration, it also prevented me from obtaining my own data in the laboratory as I was in home-office for the next 15 months. As such, the initial idea to utilise existing data sets for investigating oxidative folding capacities in *E. coli* was developed much further than initially envisioned and became the central piece of work of this thesis.

One and a half years later the pandemic had cooled off and scientific collaborations were possible again. However, at this point both my own project as well as my fellow ESRs projects had changed significantly and the initial ideas for collaborations were no longer feasible. Nevertheless, new ideas arose. In early 2022 I was invited to go to the University of Oulu, Finland and work on an oxidative folding project in *K. phaffii*. This secondment lasted for a total of 5 months and not only allowed me to investigate oxidative folding in this yeast but also to go back to the laboratory and generate my own data.

It was during this secondment that the idea for the third part of this thesis was developed. Initially envisioned for investigation of the disulfide proteome of *K. phaffii*, the protein structure prediction algorithm “AlphaFold” became a central source of information for disulfide bonds in many more biotechnologically relevant organisms. AlphaFold allowed me to also include not only the last remaining SECRETERS organisms of interest *B. subtilis* into my work, but also investigate most other biotechnologically relevant organisms such as mammalian cells.

Looking beyond the capabilities of a single organism became a core aspect of my work and was further highlighted by the review written by the "SECRETERS" ECRs and for which I acted as corresponding author titled “Microbial protein cell factories fight back?”. For many of the ESR involved, myself included, writing this review was exceedingly challenging as it was done completely remotely during the height of the SARS-CoV-2 pandemic. And as such it is a great source of pride for me and my fellow ESRs.

In summary, this thesis investigates oxidative folding in several different organisms and combines both different techniques and research approaches. It provides insight into the quantitative oxidative folding capacities of *E. coli* and many other organisms and sheds further light into the oxidative folding machinery of the increasingly biotechnologically relevant yeast *K. phaffii*. The new insights gained during the course of this work will provide the scientific community with valuable new understanding of the capabilities and limitations of oxidative folding in these organisms.

Although the last three years have certainly not been easy, I am immensely grateful for the opportunity to work on this project and collaborate with so many fantastic people.

Acknowledgments

I want to thank the People Programme (Marie Skłodowska-Curie Actions) of the European Union's Horizon 2020 Programme under REA grant agreement no. 813979 (SECRETERS) for funding this project.

Selecting the right PhD project supervisor is a pivotal decision. In this regard, I feel deeply fortunate to have Tobias as my supervisor. Over the past three years, especially during the challenging periods of the pandemic and remote work, his unwavering patience and support have been invaluable to me.

Moving to a new country without any family or friends can be a daunting task which is why I am beyond grateful for all the friends I made during my time in Canterbury. Here I want to give a very special thanks to Charlotte Bilsby who has been an amazing friend and a fantastic housemate.

What makes ITN projects like SECRETERS truly remarkable are the collaborations and connections forged among students, PIs, and organizations. I'm immensely grateful for the enriching scientific meetings and discussions we've had. Within this wonderful community, I'd like to extend a heartfelt thanks to Professor Lloyd Ruddock, whose pivotal role in our review and generous offer to join his working group in Oulu during my secondment have been particularly noteworthy.

This thesis project started with my arrival in Canterbury in 2019, there were many important professors and teachers on my way to get there. I extend a heartfelt appreciation to Dr. Oliver Spadiut, who mentored me through my master's and bachelor's theses, igniting my passion for biotechnology and introducing me to the world of scientific research. Also, to Mag. Renate Untner, Mag. Ulrika Notdurfter and Mag. Ulrike Riedel, my high school teachers whose guidance and teachings were pivotal in shaping my academic journey to its current stage.

Finally, I want to thank my family for their love and support. They have always been there for me and helped me become the man I am today. Here I need to give the most special of thanks to my mother, Marianne Rettenbacher, who has given me the most amazing support any son could hope for, and I want to dedicate this thesis to her.

Abbreviations

0S	Fully reduced BPTI with no disulfide bonds
1S	BPTI with one disulfide bond
2S	BPTI with two disulfide bonds
3S	Fully oxidised BPTI with three disulfide bonds
Å	Angstrom (same unit as 0.1 nm)
ACN	Acetonitrile
AF	AlphaFold 2
AOX1	Alcohol oxidase 1
ARBA	Association-rule-based annotator
ATP	Adenosine triphosphate
BPTI	Bovine or basic pancreatic trypsin inhibitor
BSE	Bovine spongiform encephalopathy
CASP	Critical assessment of protein structure prediction
CC-BY-4.0	Creative Commons License - Attribution 4.0 International
CD1	Cluster of differentiation 1
CHO	Chinese hamster ovary
CryoEM	Cryogenic electron microscopy
CV	Column volume
CyDisCo	Cytoplasmic disulfide bond formation in *E. coli*
DARPins	Designed ankyrin repeat proteins
DNA	Deoxyribonucleic acid
DSB	Disulfide bond
EC	Enzyme classification
EDTA	Ethylenediaminetetraacetic acid
ER	Endoplasmic reticulum
ESI	Electrospray ionisation
EU	European Union
FAD	Flavin-adenine-dinukleotide
FP	Folded protein
GC	Gas chromatography
G-CSF	Granulocyte colony stimulating factor
GPCRs	G protein-coupled receptors
GSH	Glutathione reduced form

GSSG	Glutathione oxidised form
hG-CSF	Human granulocyte colony stimulating factor
hGH	Human growth hormone
hEGF	Human epidermal growth factor
hIFN-α	Human interferon alpha
hIL-6	Human interleukin 6
IFN	Interferon
IL	Interleukin
IMAC	Immobilized metal ion affinity chromatography
IPTG	Isopropyl ß-D-1-thiogalactopyranoside
ITN	Innovative Training Network
kDa	Kilo Dalton (same unit as kg/mol)
LB	Lysogeny broth
LC	Liquid chromatography
mAb	Monoclonal antibody
MALDI	Matrix-assisted laser desorption ionisation
MFP	Misfolded protein
MHC	Major histocompatibility complex
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
mV	Milli volt
NEM	N-Ethylmaleimide
NMR	Nuclear magnetic resonance
OD_{600}	Optical density at 600 nm
ODE	Ordinary differential equation
PaxDB	Protein abundance database
PDB	Protein data bank
PDI	Protein disulfide isomerase
POI	Protein of interest
ppm	Parts per million
PRIDE	Proteomics identification database
PTM	Post translational modification
Q	Quinone
QSOX	Quiescin sulfhydryl oxidase
rHSA	Recombinant human serum albumin

R^2	Coefficient of determination
RNA	Ribonucleic acid
RP	Reverse phase (chromatography)
rpm	Rounds per minute
RT	Retention time
SDS-PAGE	Sodium dodecyl sulfate–polyacrylamide gel electrophoresis
scFV	Single chain variable fragment
TFA	Trifluoroacetic acid
TNF	Tumour necrosis factor
UB	Ubiquitin
UFP	Unfolded protein
UPR	Unfolded protein response
vHH	Variable domains of (camelid) heavy chain-only antibody or Single-domain antibody

Introduction

Oxidative folding has been of immense scientific and biotechnological interest for decades [1], [2]. Cellular pathways, electron transfer chains and kinetic mechanisms have been extensively investigated for many different organisms [3], [4]. This introduction will serve to summarise this mostly qualitative knowledge and combine it with modern sources of quantitative data. The current developments for microbial systems in biotechnology have been extensively described in the review titled “Microbial protein cell factories fight back?” written by myself and my Marie Curie Grant colleagues [5]. Parts of this introduction were used in that published review (The relevant paragraphs are marked with floating commas).

Why we study microbes

Microbial organisms such as bacteria and yeasts have a long history of human utilisation. For hundreds of years, they were used for curing and preserving food and once these organisms became genetically tractable, the knowledge gained from decades of brewing, was adapted to the production of proteins and metabolites such as citric acid, penicillin and insulin [6]. During the second half of the 20th century the number of available products and processes increased sharply, and biotechnology as a whole became a major industry [7]. During the last decades microbial based production processes have gained increasing competition. Mammalian cells such as the widely used Chinese Hamster Ovary (CHO) cells have significant advantages over microbes when it comes to producing proteins with post-translational modifications (PTMs) such as antibodies [8]. Considering this increasing competition, it has become more important than ever to promote a more holistic understanding of microbial organisms. We rely on these microorganisms for scientific research and the production of life saving medications, yet there is much about them that remains unexplored and unknown. In Chapter 1 the quantitative capabilities of *Escherichia coli*, one of the central organisms not only in research but also in protein production, will be investigated. In Chapter 2 one of the central enzymes in oxidative folding in eukaryotes will be kinetically characterised for one of the most promising organisms in protein production, *Pichia pastoris*. Therefore, the next two paragraphs will introduce these two organisms in more detail before the next sections will summarize the current state of microbial protein production as a whole.

Escherichia coli

Many modern research techniques such as recombinant protein expression and genetic approaches were originally developed in *E. coli* [9]. Consequently, *E. coli* itself has been extensively studied and became the most completely understood organism in all of science. Researchers value the simplicity of its cultivation and the wide range of tools available for cloning and genetic manipulations of any kind. A wide range of *E. coli* strains engineered to perform specific tasks have been established, broadening the applicability of this versatile organism even further [10]. Its genome was one of the first to be completely described by genomics and subsequently transcriptomics, proteomics, metabolomics and fluxomics studies have all been completed for *E. coli* [11], [12], [13]. In 2006 the Keio collection was published, providing the researchers worldwide with a library containing every possible *E. coli* single-gene knockout [14].

With all this available information and tools, it comes as no surprise that *E. coli* is also widely used in biotechnology and the biopharmaceutical industry for the production of recombinant proteins.

Pichia pastoris (syn. Komagataella spp.)

The methylotrophic yeast *P. pastoris* has established itself as one of the most widely used organisms in the biotechnological industry. It was first isolated in 1920 in France by Alexandre Guilliermond and isolated again in 1950 by Herman Phaff in California. Following phylogenetic analysis performed by Yamada et al. in 1995 the genus was reclassified as Komagataella in 2005 and separated into two species: *K. pastoris* (which includes the France isolate) and *K. phaffi* (which includes the Californian isolate) [15]. However, the old name stuck around and both strains collectively are still commonly referred to as *P. pastoris*. This yeast has established itself at the forefront of biotechnology thanks to its strongly-inducible *AOX1* promotor which evolved as a fast cellular response to the availability of the cell-toxin and carbon source methanol [16]. Cloning recombinant proteins under this promotor can achieve high yields.

The current recombinant protein market

“

Small cells in an expanding market

Recombinant proteins have dramatically changed our lives, and their market size and impact are projected to keep expanding (Figure 1). Despite the widespread use of recombinant proteins in many industrial sectors, one of the main driving forces for continuous market expansion are biopharmaceuticals, the fastest growing group in the pharmaceutical industry [17]. This has triggered the development of a large spectrum of industrial expression platforms for their production, including both microbial and mammalian cell hosts [18].

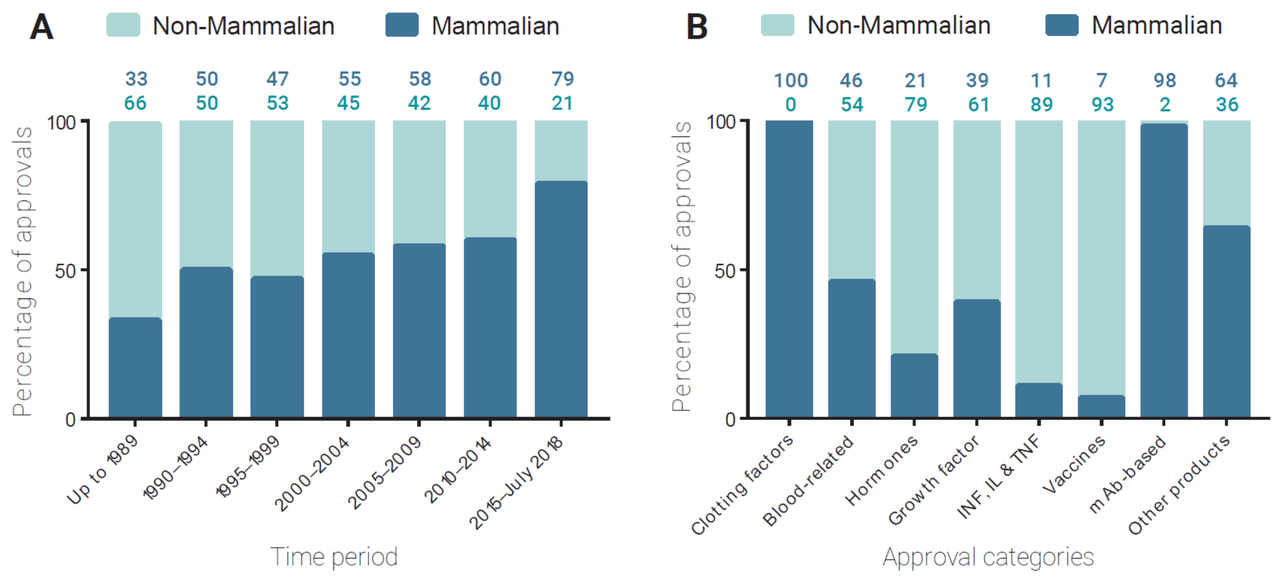


Figure 1 Market share comparison between biotherapeutic production hosts. (A) Approval trend of mammalian versus non-mammalian biopharmaceuticals from 1989 to 2018. (B) Percentage of biopharmaceuticals approved in the USA and EU up until mid-2018, categorized by product type and compared between mammalian and non-mammalian hosts. Both graphs are built on data collected by Gary Walsh [19]. Abbreviations: IFN, inteferon, IL, interleukin; mAb, monoclonal antibody; TNF, tumor necrosis factor.

The trend in recent years has seen mammalian cell lines increasingly outcompete their microbial counterparts (Figure 1A). From 2014 to mid-2018, more than 87% of the genuinely new biopharmaceutical active ingredients that were released to the market were proteins [19]. Of these, 84% were expressed in mammalian expression systems, with Chinese hamster ovary (CHO) cell-based systems being the most widely used. This surge in the biotherapeutics sector can be explained mainly by the increasing dominance of monoclonal antibodies (mAbs), which require humanised post-translational modifications (PTMs) [19].

Despite the numerous advantages of mammalian-based protein production, there are also drawbacks. Compared with mammalian hosts, microbial expression systems are characterised by being easy to work with, robust, and cost-effective, all highly desirable features in the context of biopharmaceutical production. In fact, microbial platforms are capable of delivering in a scalable and affordable manner a range of functional recombinant therapeutic proteins [20], such as vaccines, hormones, interferons, and growth factors (Figure 1B), as well as non-pharma products, such as industrial enzymes. Nevertheless, these platforms also come with some disadvantages, particularly with respect to their PTM requirements and secretion.

In the diverse area of scientific developments and engineering strategies the SECRETERS grant network follows an integrative approach to summarise and compare the major innovations in the field of microbial recombinant protein production, with a focus on four microbial platforms: *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae*, and *Pichia pastoris* (syn. Komagataella spp.). The following section is a discussion on recent scientific progress made in innovation-rich research areas, such as omics, systems biology, protein secretion, and PTMs, and place them into the wider context of the biopharmaceutical market. By examining not only recent improvements, but also competing production methods and organisms, I predict the future roles of microbes in the pharmaceutical market.

"

What can be achieved with microbial based production

"

Antibody formats and mimetics

Full-length mAbs represent a large part of the global biopharmaceutical market. Due to their high complexity, they are mostly produced in mammalian cell lines, which require long processing times and elevated production costs. More recently, we have seen a shift toward the production of antibody fragments, among which single-chain variable fragments (scFvs) and fragmented antigen-binding (Fabs) are the most exploited [21]. These small antibody-based formats have several advantages and, due to the lack of requirement for glycosylation, can be expressed in microbial platforms. In 2017, Gupta and Shukla reviewed an exhaustive survey of expression technologies for the production of antibody-like molecules in microbes, particularly in *E. coli* [22]. In the same year, another review covered antibody fragments and, more in general, biopharmaceuticals production in Bacillus strains [23]. Yeast species, notably *S. cerevisiae* and *P. pastoris*, are also used as hosts for the expression of antibody formats, such as scFvs, single-domain antibodies (vHHs), and Fabs [24], [25]. In addition to the four hosts discussed here, some promising alternative microorganisms have also made significant attempts to enter the antibody formats markets.

Antibody-like scaffolds, also known as antibody mimetics, are considered to be future alternatives to mAbs. These include adnectins, affibodies, anticalins, avimers, DARPins, fynomers, and Kunitz domains (Figure 2) [26], [27]. Similar to antibodies, these compounds have high target-binding specificity and affinity, but are characterized by additional benefits, such as a much smaller size, increased thermostability, and low immunogenic potential. Moreover, they can be easily and efficiently produced in microbial host cells: Binz and colleagues reported yields of 15 g/l of DARPins expressed solubly in *E. coli* [28].

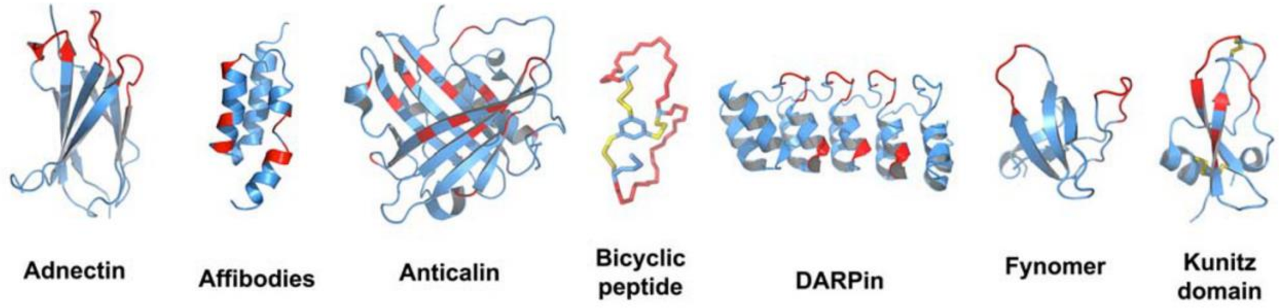


Figure 2 Common formats of antibody mimetics. Figure from [26] with extra credit mentioned to Daniel Christ.

Non-antibodies

Some categories of biopharmaceuticals are still heavily dominated by microbial hosts (Figure 1B). These include interferons (e.g., IFNα-2b), cytokines [e.g., granulocyte colony stimulating factor (G-CSF) and tumour necrosis factors (TNFs)], hormones [e.g., insulin glargine and human growth hormone (hGH)], and interleukins (IL-2 and IL-11), all of which can be produced in *E. coli* and yeasts[29], [30], [31]. High recombinant human serum albumin (rHSA) yields were recently reported (17.5 g/l) in *P. pastoris* by means of medium optimization [32].

Industrial proteins

Due to its efficient secretory production, *P. pastoris* is widely used also for the recombinant production of protein-based polymers [33] and enzymes. Some examples are reviewed by Vieira Gomes and colleagues [34] and Gifre and colleagues [35], the latter of which illustrates an extensive array of enzymes with interest in the feed industry. These include phytases, which are produced recombinantly mainly in *P. pastoris* and *E. coli*. Among the most important industrial enzyme producers are Bacillus spp., which are capable of secreting 20–25 g/l of proteins into the culture medium [36]. These enzymes can be applied in numerous industrial applications, such as food, feed, detergent, textile, and waste treatment processes [37].

Titer comparison of selected proteins of interest

Comparing titer numbers between organisms can be difficult because of significant differences both in terms of production process (e.g., fermentation mode, and media composition and optimization) and product analytics (including product quality, correct folding, biological activity, and quantification of titers). Despite this difficulty in benchmarking, it is clear that, once microbial systems can produce a protein, they generally outperform CHO cells in terms of volumetric titres (Table 1). For individual proteins of interest (POIs), there can be pronounced differences between the four compared microbes, but there is no clear overall winner.

Table 1: Comparative list of reported titres of therapeutic proteins 1

Protein	Production system	Cell culture process	Site of accumulation	Titer[2]	Ref
hEGF 3x DSBs	*Bacillus spp.*	Shake flask	Culture medium	360 mg/L [3]	[38]
	E. coli	Fed-batch bioreactor	Culture medium	250 mg/L	[39]
	S. cerevisiae	Fed-batch bioreactor	Culture medium	259.2 mg/L	[40]
	P. pastoris	Baffled shake flask	Culture medium	2.5 mg/L	[41]
	CHO	-	-	-	-
hGH 1x DSB	*Bacillus spp.*	Fed-batch bioreactor	Culture medium	497 mg /L	[42]
	E. coli	Fed-batch bioreactor	Periplasm	2390 mg/L [5]	[43]
	S. cerevisiae	Shake flask	Culture medium	0.9 mg/L	[44]
	P. pastoris	Fed-batch bioreactor	Culture medium	640 mg/L	[45]
	CHO	Semi-continuous batch culture in spinner flasks	Culture medium	75 mg/L	[46]
hIFN-α 2x DSBs 1x O-glycos. site	*Bacillus spp.*	Shake flask	Culture medium	15 mg /L	[47]
	E. coli	Fed-batch bioreactor	Cytoplasm	300 mg/L [5,4]	[48]
	S. cerevisiae	Fed-batch bioreactor	Culture medium	276 mg/L	[49]
	P. pastoris	Batch bioreactor	Culture medium	436 mg/L	[50]
	CHO	-	-	-	-
hIL-6 2x DSBs 1x N-glycos. site	*Bacillus spp.*	Shake flask	Culture medium	200 mg/L [5]	[51]
	E. coli	Fed-batch bioreactor	Cytoplasm	1100 mg/L [5]	[52]
	S. cerevisiae	Batch bioreactor	Culture medium	30 mg/L	[53]
	P. pastoris	Fed-batch bioreactor	Culture medium	170 mg/L [5]	[54]
	CHO	35 mm dishes	Culture medium	1.4 µg/ 10^6cells/day	[55]
hG-CSF	*Bacillus spp.*	Batch bioreactor	Culture medium	120 mg/L	[56]

[1] For improved comparison, high-density CHO cultures have average cell-densities of 2-4 x 10^6 cells/mL in a suspension-batch mode and 10-15 x 10^6 cells/mL in a suspension-perfusion mode using stirred-tank bioreactors. Disulfide bonds (**DSBs**) and N- and O-glycosylation sites for each protein are also listed.

[2] Unless indicated ([6]), all proteins were obtained in a soluble form.

[3] Purified titres

[4] Produced as inclusion bodies. For a state-of-the-art overview on how IBs are processed please refer to the “In vitro Refolding of Proteins” Chapter by Haslbeck and Buchner [236].

2x DSBs 1x O-glycos. site	*E. coli*	Fed-batch bioreactor	Cytoplasm	4200 mg/L [6]	[57]
	S. cerevisiae	Fed-batch bioreactor	Culture medium	98 mg/L	[58]
	P. pastoris	Fed-batch bioreactor	Culture medium	35 mg/L	[59]
	CHO	100 mm dishes	Culture medium	90 µg/ 10^6cells/day	[60]
Insulin 3x DSBs	*Bacillus spp.*	Batch bioreactor	Culture medium	1000 mg/L	[61]
	E. coli	Fed-batch bioreactor	Cytoplasm	4340 mg/L [6]	[62]
	S. cerevisiae	Shake flask	Culture medium	79 mg/L	[63]
	P. pastoris	Fed-batch bioreactor	Culture medium	3075 mg/L	[64]
	CHO	T-75 flasks	Culture medium	1.98 ng/ 10^6cells/day	[65]

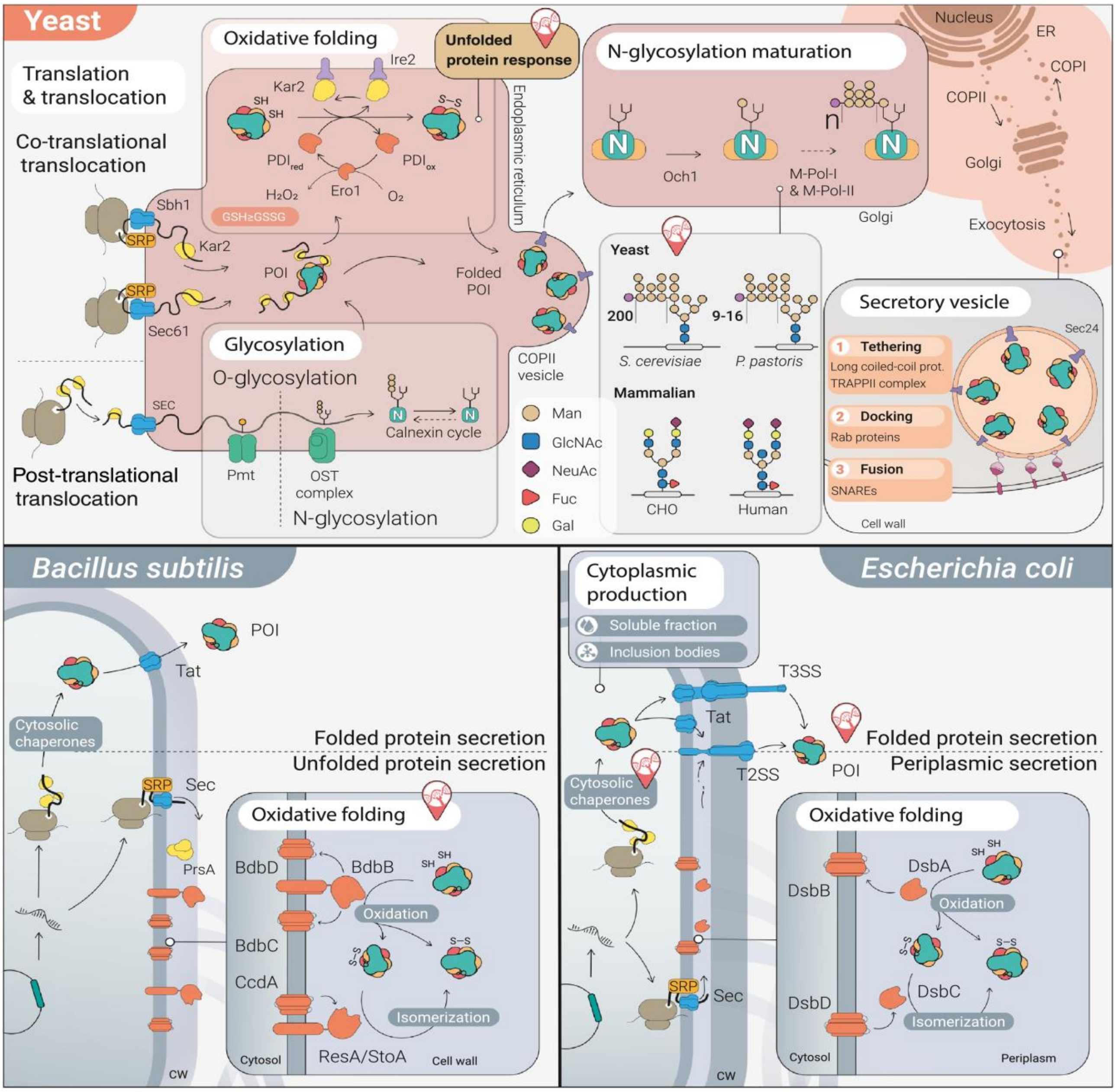
“

What do microbial systems struggle to produce

“

What is holding microbes back? PTMs

PTMs occur during or after protein synthesis. They change the physicochemical properties and potentially the activity of the protein. They range from small chemical modifications to the amino acid chain, to the addition of complex branching structures on the protein backbone [66], [67]. Such modifications can vary immensely in terms of form and function. Many PTMs are involved in metabolic crosstalk, such as phosphorylation and glycosylation, or in improved folding and stability, in which glycosylation and disulfide bonds have a major role, or in signaling for aberrations such as nitrosylation and deamination. The most commonly found PTMs in recombinant proteins are glycosylation and disulfide bond (DSB) formation. The endogenous mechanisms carrying out these two processes are illustrated in Figure 3 . While other PTMs, such as methylation and carboxylation exist, glycosylation and DSB formation are frequently required for correct protein folding and biological activity. However, only disulfide bond formation will be discussed in more detail below.



Trends in Biotechnology

Figure 3 Key cellular machineries involved in protein production in yeasts, E. coli and B. subtilis. Image taken from the SECRETERS review [5]. More detailed description of the different (here not relevant) elements can be found there.

Disulfide bonds

Many pharmaceutically relevant proteins are disulfide bonded. While *E. coli* and *B. subtilis* both have endogenous systems for disulfide bond formation in the periplasm and cell wall, respectively (Figure 3), they can struggle with disulfide-rich proteins with complex folding patterns. Nevertheless, it is possible to obtain g/l yields in some cases [43]. By contrast, yeasts have a DSB-forming system located in the endoplasmic reticulum (ER), an environment that is capable of pairing up more complex DSBs for their endogenous proteins. However, a strong unfolded protein response (UPR) when expressing heterologous disulfide-bonded proteins can cause issues [68]. When a cell detects an excess amount of unfolded or misfolded proteins in its ER, the UPR gets activated to mitigate this cell-stress [69]. The three main responses are an overexpression of folding factors and chaperones, the slowing down of protein translation as well as an increased degradation of unfolded proteins. Several attempts at mitigating the UPR response in both *S. cerevisiae* and *P. pastoris* have been made [70].

A common approach to increase rates of DSB formation in all four microbial systems is the overexpression of chaperones and isomerases, such as Kar2 and PDI, in yeast, the extracellular chaperone PrsA in *B. subtilis*, and periplasmic chaperons DsbC, FkpA, DsbA, and SurA in *E. coli* [71], [72], [73]. All of these have resulted in improved protein titres. Furthermore, these overexpressed folding catalysts not only improve protein solubility, but may also facilitate improved quality control because misfolded proteins are either refolded or tagged for proteolytic destruction [74].

Over the past decade, several different approaches aiming to improve DSB formation in *E. coli* have been researched. Disruption of the native reducing pathways resulted in the Origami strain, capable of forming DSBs in the cytoplasm [75]. Later, the Shuffle strain also incorporated the native periplasmic isomerase DsbC (Figure 3) in the cytoplasm, further improving cytoplasmic DSB formation [76]. Both strains have historically been used in research labs. However, while they both facilitate improved DSB formation in some heterologous proteins, they lack growth fitness and sufficient protein yields for industrial relevance [77]. Similar attempts have been made to enhance DSB formation in proteins secreted by *B. subtilis* but, despite proof of principle, major breakthroughs in terms of industrial bioproduction have not yet been achieved [78].

Owing to its reducing nature, the cytoplasm of *E. coli* does not have any endogenous mechanisms for structural DSB formation (Figure 3). However, a different approach toward improving DSB formation is based on the cytoplasmic expression of recombinant oxidases and isomerases. This was first established in the CyDisCo strain, which expresses eukaryotic Erv1p and PDI from a plasmid [79]. Over the past decade, this approach has yielded promising results both in terms of DSB complexity and yields. An example of the former is the production of a 44-DSB containing extracellular matrix protein by Sohail and colleagues [52], [80]. A note-worthy yield achievement facilitated by the CyDisCo system is the reported production of around 1 g/l of both hGH1 and IL6 [52].

“

Protein Folding

Ribosomes translate mRNA into its corresponding / encoded peptide chain. The mRNA contains the base pair sequence transcribed from a gene and sets of three base pairs code for one of the 20 native amino acids according to an (almost) universal translation code. These individual amino acids are then linked via a peptide bond in the ribosome during the translation process. The resulting chain of amino acids is referred to as a polypeptide. Before these poly peptides can be considered proteins, they first must obtain their native confirmation, i.e., their correct three-dimensional structure.

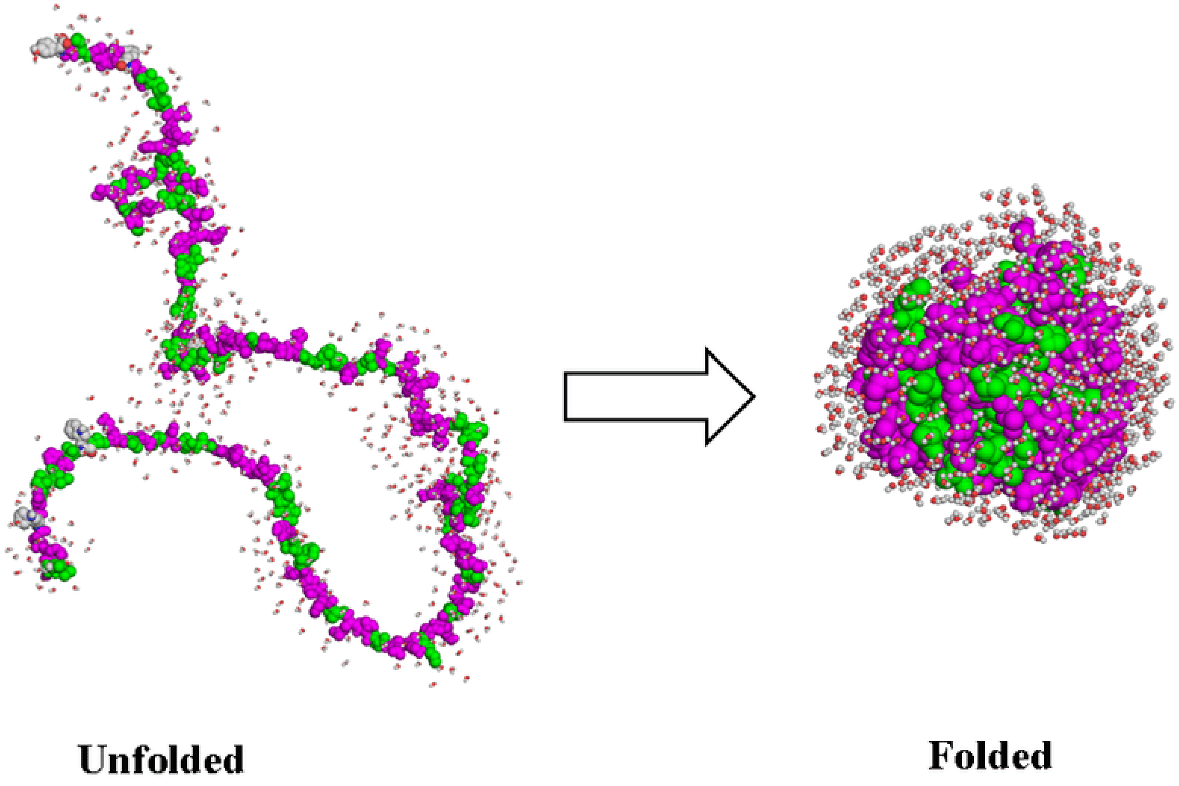


Figure 4 Schematic visualisation of the protein collapse from unfolded to folded protein structure. Hydrophobic amino acids are displayed in green and hydrophilic ones in purple. After the collapse the hydrophilic amino acids are predominantly in the core of the protein structure. Figure from Guidi et al. [81]

In its simplest form the folding process already begins during the translation of the peptide chain as it exits the ribosome. The side chains of the amino acids interact with each other and gradually the unordered peptide chain assembles into its native secondary and tertiary structure. This 'ideal' folding behaviour was famously championed by Christian B. Anfinsen in 1961 and subsequently named the 'Anfinsen dogma' [82]. This dogma states that for simple globular proteins, the native folding state is unique and under native conditions solely determined by the primary structure of the protein. And while this dogma has been widely challenged since its first postulation, it still holds true for many 'simple' proteins.

Although in many cases the correct folding of proteins by themselves would be achieved eventually, it is often important for the cells to assemble and fold new proteins more quickly to react to environmental changes or to avoid aggregation of unfolded proteins in the cell [83].

Therefore, the underlying self-assembly of the peptide chain is assisted by a wide range of folding factors and enzymes specifically designed to stabilize unfolded proteins (i.e., chaperones), add or cut molecules to or from the peptide (e.g., proteases and transferases), connect amino acid side chain together (e.g., oxidases) or isomerise between different folding confirmations (i.e., isomerases) [84].

Failure to correctly fold proteins can have severe impact on the fitness of the cell or the health of the organism [85]. Several human diseases are linked to misfolded proteins. Mutation in genes can leading to changes in the amino acids which in turn changes the ‘correct’ folding of the protein to one that might not be able to perform it’s intended function. In other cases, changes to the protein folding apparatus itself can lead to misfolded proteins which are then unable to perform their intended tasks correctly or sufficiently [86]. Misfolded proteins themselves can also be the cause of diseases when they aggregate and resist cellular attempts for their digestion and removal. The most well-known example of this is Alzheimer’s disease which is caused by the aggregation of amyloid proteins in the brain [87].

A special case of disease caused by misfolded proteins is BSE which is caused by prions [88]. Prions are a direct contradiction to Anfinsens dogma because they are proteins with different (‘wrong’) confirmations. These prions can interact with other ‘correctly’ folded versions of themselves and cause them to change their confirmation to the ‘wrong’ folding state as well. Because of this, prions are often described as the virus version of a protein because of their predatory way of self-propagation.

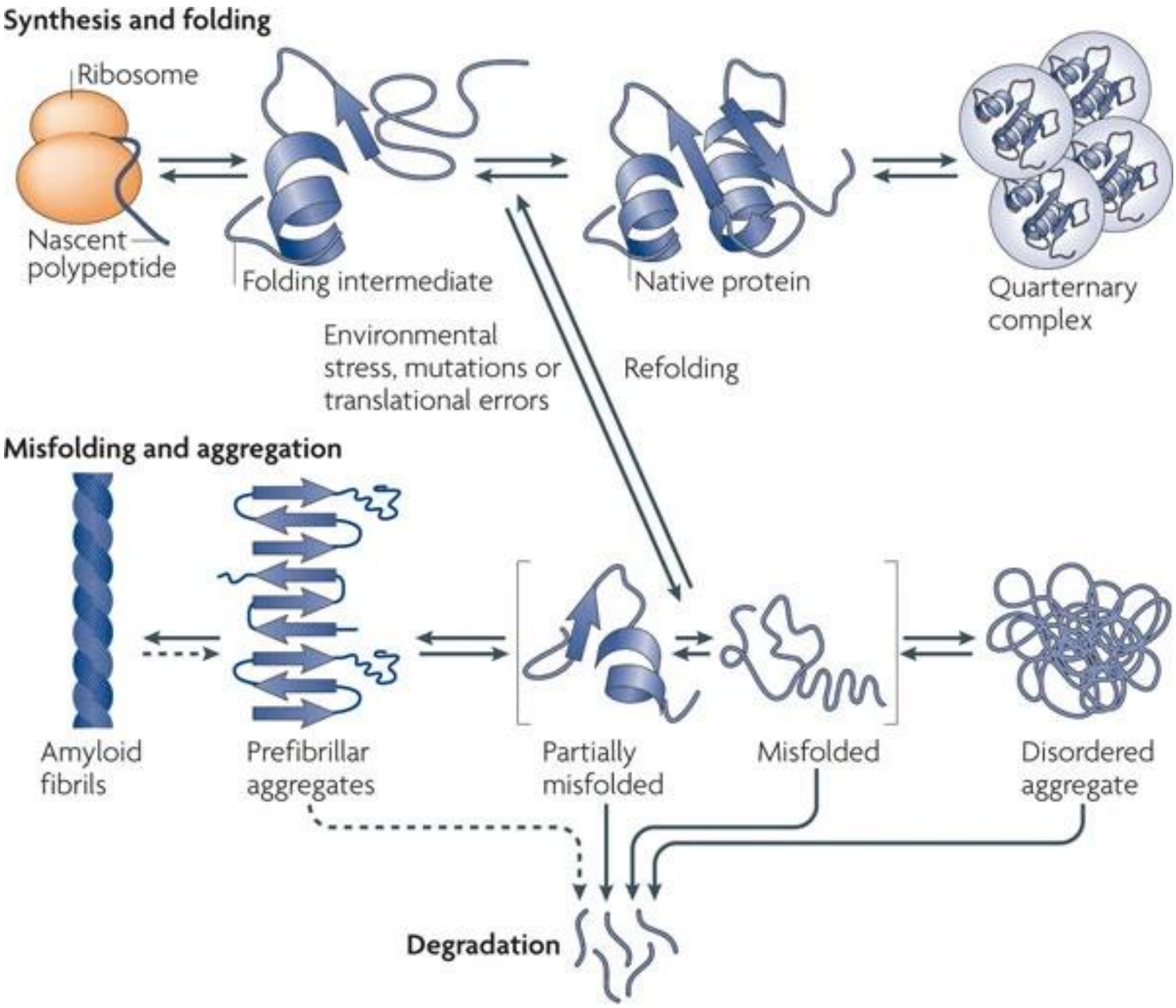


Figure 5 Possible protein folding pathways. Top row displays protein synthesis followed by correct protein folding and assembly. Bottom row displays possibilities for misfolding and unwanted protein aggregation. Figure from Tyedmers et al. [89]

Protein misfolding and aggregation is not only an issue in human health, it also has strong implications for the biotech industry [90]. As stated above, cells have developed a wide range of enzymes and even cellular compartments with the main purpose of correctly folding proteins. And even though the translational apparatus is largely compatible between different organisms, the enzymes and environments that are involved in protein folding can differ significantly.

When attempting to produce a recombinant protein in an organism that has not evolved to correctly fold proteins similar in nature to the desired protein, the cell will struggle to produce it [91]. The classical example of this is the production of complex eukaryotic enzymes (often from humans) in prokaryotic cells such as *E. coli*. If the cell fails to fold the peptide chains correctly, they can either be digested by the cell or (when there are too many unfolded peptide chains to digest) the unfolded proteins will aggregate and form inclusion bodies within the cell. In the case of *E. coli* this phenomenon of inclusion body formation has been successfully used to produce recombinant proteins by harvesting the aggregated and unfolded protein and folding them in-vitro [92]. However, in most protein production strategies this aggregation of proteins in undesirable.

Due to these limitations many research and development attempts have been made to improve protein folding, particularly in microbes such as *E. coli* and *B. subtilis* which are widely used for protein production in part due to their favourable fermentation conditions [5]. Common approaches include the co-expression of chaperones or isomerases, moderation of the target proteins expression levels and expression of the target proteins as fusion-constructs. However, there are many more strategies and in most cases the approach is uniquely designed for each protein of interest.

Oxidative Folding

The chemistry of the disulfide bond

Oxidative protein folding is the process which links two cysteine amino acids together via a covalent disulfide bond between the two sulfur atoms of the cysteines. Chemically speaking this constitutes a redox reaction which oxidises the sulfur atoms from their reduced form (-SH) to an oxidised form (-S-S-). In enzymes, disulfide bonds can be classified in one of two categories, functional or structural [93]. The former is usually found in the active centres of proteins and are involved in the functionality of the enzyme in some shape or form. The structural disulfide bonds on the other hand are required by a protein to attain its correct 3D structure. However, since structure and function of proteins are heavily linked, the distinction can be blurry at times. A better distinction might therefore be between temporary disulfide bonds and permanent disulfide bonds.

When a reduced pair of cysteines interacts with a redox partner with a lower redox potential disulfide bond compared to itself, it can transfer its electrons to that partner. In biological environments the redox partner of reduced cysteines is most commonly another DSB. When such a ‘donating’ disulfide bond is brought into proximity with the reduced cysteine it can be nucleophilic attacked by the cysteine and form a mixed disulfide bond with it instead [94]. This creates a temporary covalent disulfide bond between the enzymes. When a second reduced cysteine then comes into proximity with this temporary bond another nucleophilic attack can occur which releases the temporary link and forms the inner-protein disulfide bond. An important sidenote: this description as a transfer of electrons is often less intuitive to understand than the inverse description of this redox reaction which is the transfer of the disulfide bond in the opposite direction. While the first is more chemically ‘correct’ the latter is more easily understandable which is why it is preferentially used in this work.

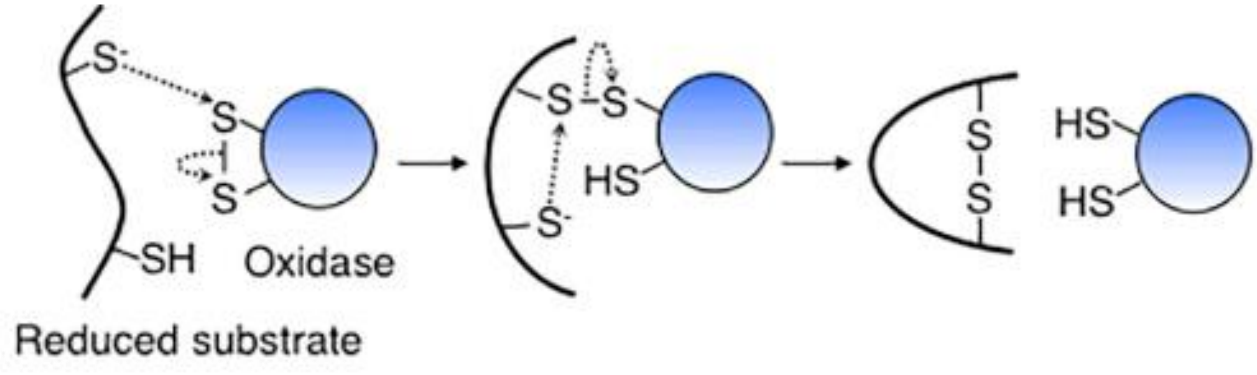


Figure 6 Oxidative folding reaction of a disulfide bond. The DSB acceptor (e.g. two reduced cysteine residues) enters a redox reaction with the donor (e.g. an oxidoreductase enzyme). The reaction constitutes an electron transfer from the DSB acceptor to the DSB donor. Figure from Fujimoto et al. [95]

This leaves the disulfide bond donating protein in a reduced state and the receiving protein with the disulfide bond and in an oxidised state. The donating enzyme can then be regenerated by interacting with another enzyme or molecule that is even more electron negative than itself. In aerobic conditions, molecular oxygen is usually the terminal electron acceptor in this oxidation chain which creates H_2O_2 and eventually H_2O.

The stability of a cysteine-cysteine disulfide bond is in a very useful range for biological systems. The theoretical dissociation energy of a S-S bond is roughly 250 kJ mol^{-1}. This energy level makes for a stable covalent connection between to atoms but also a relatively weak link compared to other common bonds such as the C-C bond or the C-H bond (347 and 413 respectively) [96]. Furthermore, the electron negativity of sulfur (2.58 on the Pauling scale) is not as high compared to the more common representative of the Chalcogen periodic table group, oxygen (3.44 on the Pauling scale). This makes for an atom which is more easily reduced or oxidised in a biological environment. The associated redox potential of sulfur atoms in sulfur-organic compounds such as enzymes, can be adapted based on the surrounding structure of the sulfur atom. Neighbouring electron-donating groups can stabilize disulfide bonds, while electron-withdrawing groups can destabilize them (or stabilize reduced cysteines). Based on this adaptability of the disulfide bond, evolution has created a range of enzymes which can stabilize, inhibit or modulate each other via the transfer of disulfide bonds [97].

The thioredoxin fold

The thioredoxin fold is the central motif for cysteine-linked enzymes [98]. As the name suggests, it was first described in thioredoxin a cytoplasmic protein found in many prokaryotes such as *E. coli* where its job is to keep the cytoplasmic environment reducing. The thioredoxin fold, however, has subsequently been found conserved in many different enzymes in every kingdom of life, usually involved in disulfide redox activities [99]. The structure of the fold consists of a four-stranded antiparallel beta sheet core surrounded by three alpha helices and a CXXC active site motif. The roughly 80 amino acids long thioredoxin fold can be further classified into an N-terminal beta-alpha-beta motif and a C-terminal beta-beta-alpha motif connected via a loop which contains the third alpha helix [100]. The remaining structure of thioredoxin fold containing proteins shows very little homology [101]. As mentioned in the above section, the redox potential of a thioredoxin fold (i.e. the active centre disulfide) can be modulated by modifications to the common scaffold. For example, the *E. coli* thioredoxin (Figure 7) has a redox potential of -271 mV [102] while the thioredoxin fold containing *E. coli* protein DsbA has a redox potential of -121 mV [103]. Another conserved feature of the thioredoxin fold is a cis-Proline in a loop close to the active centre CXXC motif [104]. This cis-bond has been linked to substrate binding and release as well as preventing of the active centre from binding metal ions [105], [106], [107].

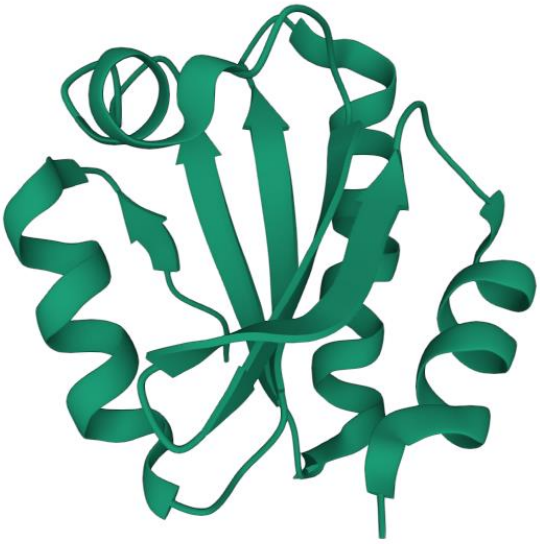


Figure 7 Template protein structure of the thioredoxin fold. Source: PDB structure, accession ID: 1M7T

Isomerisation

When a protein has more than 2 cysteine amino acids it can form more than one potential disulfide bond. In an unfolded protein chain, any two cysteines can be linked; however, in practice this is highly dependent on the folding process and the (temporary) proximity of the respective cysteines [94]. Cysteine is one of the rarest amino acids (together with tryptophan) which helps to keep the number of potential (wrong) binding partners low. Particularly in prokaryotic, periplasmic and disulfide bonded proteins, the number of excess cysteines is kept low [108]. Nevertheless, not every disulfide bond on a substrate is formed correctly. Whenever this happens, the incorrect disulfide bond must be fixed before the protein can be folded correctly and regain its function. One option is to reduce the disulfide bond, bringing the cysteines back to their unbonded state and trying to oxidise them again (this time hopefully correctly). The other approach is to fix the disulfide bond in place via isomerisation. Isomerases, such as DsbC in *E. coli* or PDI in eukaryotes, need to be in their reduced state and can then bind to a substrate with an incorrect disulfide bond. Most of these isomerases can 'detect' unfolded or misfolded proteins [94]. They can interact with exposed hydrophobic parts of the protein which would not be surface exposed in correctly folded proteins. The reduced cysteines of the isomerase can then form a mixed disulfide with the wrong disulfide bond on the substrate and select a different (hopefully correct) cysteine as the binding partner. In such cases the isomerase itself remains in its reduced state and can start another isomerisation process. Alternatively, the isomerase can keep the disulfide bond and reduce the substrate instead of isomerising it. When this happens, the isomerase becomes oxidised and needs to be regenerated before it can function as an isomerase again. This is an important function to remove excess disulfide bonds from the protein pool particularly in cases of oxidative stress.

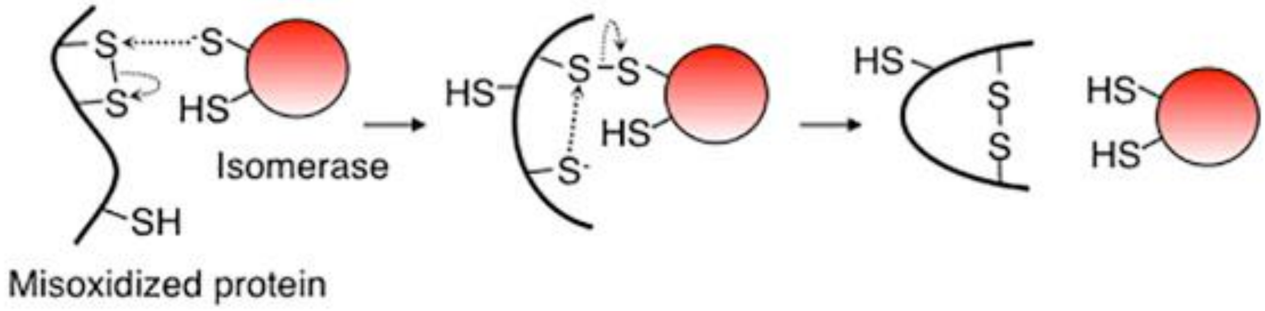


Figure 8 Isomerisation of DSB. The isomerase forms a temporary mixed disulfide bond with the substrate which allows for the subsequent formation of a different (i.e. correct) disulfide bond for the substrate. After a successful isomerisation reaction the oxidative state of both the substrate and the isomerase is unchanged. Figure from Fujimoto et al. [109]

Oxidative Folding – the enzymes involved

Oxidative folding can be split into four distinct functions: 1) creation of novel disulfide bonds, 2) transfer of disulfide bonds onto substrates, 3) isomerisation of incorrectly formed disulfide bonds and 4) removal of excess disulfide bonds. Some of these functions can be covered by a single protein or one function can be covered by a single specialist protein. Most organisms have proteins that cover all four tasks, and their key representatives are described here. While redox reaction can occur anywhere in the cell, the primary creation of disulfide bonds happens in specialized cell compartments for both bacteria and eukaryotes. For cytosol of bacteria is kept in a reducing state and therefore unfavourable for oxidative folding to occur. Therefore, in (Gram-negative) bacteria oxidative folding takes place in the periplasm. In eukaryotic cells oxidative folding happens primarily in the endoplasmic reticulum (ER).

Creating novel disulfide bonds

Oxidative folding starts with the initial creation of a novel disulfide bond between two previously reduced cysteines. In *E. coli* this task is performed by the membrane protein DsbB [110]. DsbB has four membrane-spanning segments with two pairs of disulfide forming cysteines on the periplasmic side. Both pairs are required to create an electron transfer chain designed to transfer excess electron from the periplasm (i.e. DsbA) to membrane associated quinones. Under aerobic conditions, quinones are used as the primary electron acceptor for DsbB which in turn get oxidised by terminal cytochrome oxidases which themselves transfer the electron to molecular oxygen [111]. Under anaerobic conditions molecular oxygen is not available and menaquinones are used as electron acceptor for DsbB. They then transfer the electrons on to either fumarate or nitrate as terminal electron acceptors.

The eukaryotic equivalent of DsbB is ER oxidoreductin 1 (Ero1). This FAD-binding protein can, unlike its prokaryotic counterpart, directly utilize molecular oxygen as terminal electron acceptor [112]. Ero1 (in both humans and *S. cerevisiae*) has several cysteines involved in disulfide bond creation and transfer. Two redox-active motifs are particularly important; a CXXXXC motif near the N-terminus and a CXXCXXC motif near C-terminal. Studies in yeast have determined that the N-terminal motif is primarily interacting with the PDI active side (discussed in the next section) while the C-terminal motif is involved in the electron transfer to FAD and subsequently O_2 [113]. The first cysteine in the CXXCXXC motif was shown to be involved in a long-range intramolecular disulfide bond linked to redox sensing. This regulation is of particular importance since the direct electron transfer between Ero1 and oxygen can easily result in over-oxidation if the enzyme activity is not controlled properly [114], [115].

The second important eukaryotic source of de-novo disulfide bonds is based on the family of proteins called QSOX with Erv2p being the homologous yeast member associated with it. It was identified during the screening of proteins that could rescue Ero1 mutants in yeast [116]. QSOX proteins have been shown to interact with PDI and are also able to interact with unfolded substrates directly [117], [118]

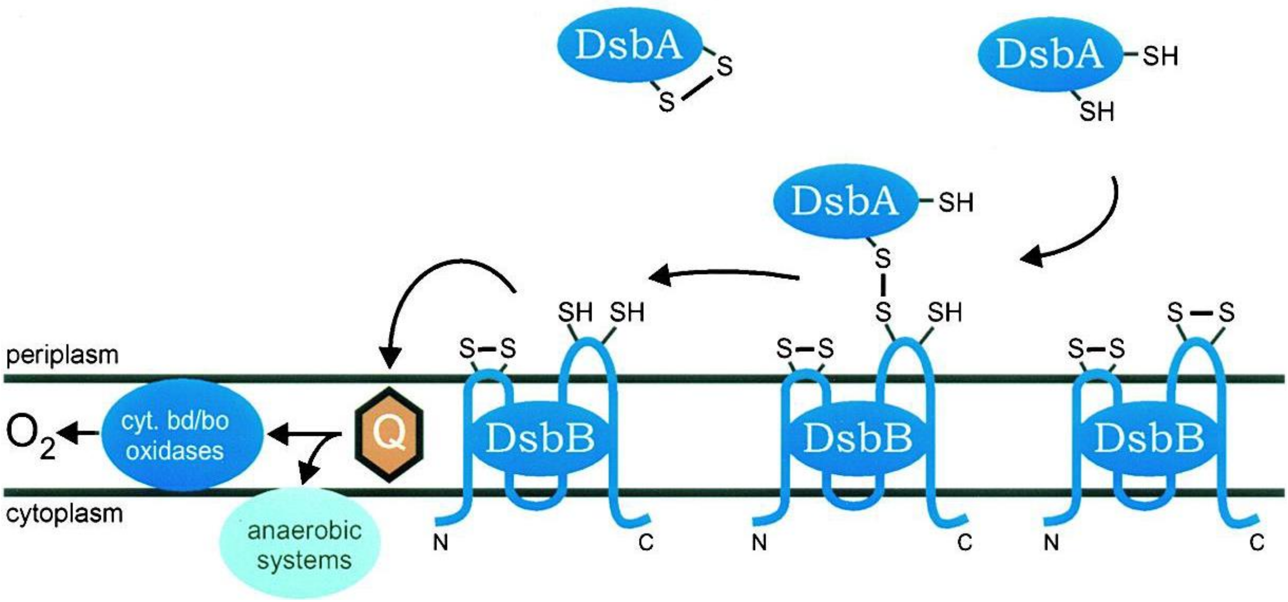


Figure 9 Electron transfer chain for the creation of noval disulfide bonds in E. coli. The electrons are transferred from DsbA onto DsbB which in turn transfers them onto quinones. Under aerobic conditions the terminal electron acceptor is oxygen. The DSB travels the opposite direction to the electrons from DsbB to DsbA. Figure frim Debarbieux and Beckwith [119]

Oxidising substrates

Once a new disulfide bond has been created (step 1) this new disulfide bond needs to be transferred to the right substrates. In *E. coli* this task is performed by DsbA. It can receive a disulfide bond from the membrane bound DsbB and transfer this bond to a reduced substrate. And as far as oxidoreductases go, DsbA is extremely oxidising with a redox potential of -121 mV [120]. These strong oxidative tendencies make DsbA a powerful substrate oxidase, but it also makes it prone to misfolding disulfide bonds in its substrates [120]. Particularly when the disulfides are not sequential. This is linked with the observation that bacteria such as *E. coli* have mostly disulfide bonds between sequential cysteines [121]. When these proteins are translocated into the periplasms via the general secretion (sec) pathway they are translocated beginning from the N-terminus to the C-terminus in an unfolded state. DsbA awaits these new substrates and relatively un-discriminatorily oxidises every pair of cysteines that enter the periplasm. This works well for consecutive disulfide bonds but leaves an increased need for isomerisation on (most likely misfolded) proteins with non-consecutive disulfide bonds.

The main eukaryotic enzyme tasked with transferring disulfide bonds onto reduced substrates is PDI. It is presumably the most thoroughly studied enzyme of all oxidative folding enzymes. PDI primarily receives its catalytic active disulfide bond from the membrane associated protein Ero1. The reduction potential of the active site of PDI is roughly -180 mV which makes it significantly less oxidising compared to its prokaryotic counterpart [122]. Also, PDI, with around 500 amino acids in length (depending on the organism) it is significantly bigger than *E. coli* DsbA with only 208 amino acids. PDI has four distinct domains labelled a, b, b' and a' as well as an extra high-acidity C-terminus (c) and a linker (x), located between domains b' and a'. And while all four domains contain a thioredoxin fold, only the a and a' domains contain the active site. Not every domain is required for PDI to retain its functionality but at least one catalytic domain (a or a') must remain [123].

Isomerisation

Substrates with misfolded disulfide bonds cannot achieve their correct structure and, therefore, function. Such misfolded enzymes would then be either destroyed via proteases or accumulate as inclusion bodies. Both are metabolic wastes of energy, and it is therefore preferential for the cell to employ an isomerisation system that can correct these misfolded disulfide bonds. In *E. coli* two disulfide bond isomerases are known, DsbC and DsbG [124]. Both are located in the periplasm and while they have similar functionality, they are hypothesised to differ in substrate specificity [125]. The isomerases require their active centre cysteines to be reduced so they can bind to wrongly connected disulfide bonds in substrates. Once the mixed disulfide with the misfolded substrate is formed, this state can be resolved in one of two ways. In the isomerisation pathway, the mixed disulfide is resolved by transferring the disulfide bond back to the substrate connecting two substrate cysteine with the disulfide bond. If the correct cysteines are now paired, the isomerisation as successful, if not, the process can be repeated. The other way this mixed disulfide bond can be resolved is for the isomerase to keep the disulfide bond to itself, leaving the substrate in a reduced state while becoming oxidised itself. This grants the substrate a new opportunity to be oxidised correctly. The now oxidised isomerase now needs to be regenerated back to its active reduced state.

Eukaryotic systems employ a different system compared to bacteria. The latter have the tasks of initial oxidation and isomerisation split between two enzymes (DsbA and DsbC/G respectively). Eukaryotic cells on the other hand use PDI as the central enzyme for both tasks. As such PDI is more selective in its initial oxidation and can subsequently isomerise misfolded proteins while already being nearby. However, this combined functionality makes the redox state regulation of PDI more important [126]. Oxidised PDI is required for initial oxidation of substrates (although this functionality is shared with QSOX proteins) but needs to be in the reduced state to perform its isomerisation functionality. An imbalance can lead to an increased accumulation of unfolded or misfolded proteins.

Removing excess disulfide bonds

In a reversed form of the reaction catalysed by DsbB, DsbD can transfer electron from a cytoplasmic donor thioredoxin on to its periplasmic acceptor DsbC (and DsbG) [127]. This reaction regenerates oxidised isomerases and results in the overall removal of a disulfide bond from the periplasm. With this being the reverse reaction to the one performed by DsbA-DsbB, it is essential for the fitness of the organism that the non-functional reactions between DsbC-DsbB and DsbA-DsbD is energetically rare/unfeasible [128].

As mentioned before, PDI functions as both primary oxidase and isomerase and therefore combines the functionality of both DsbA and DsbC. In the prokaryotic case, the two enzymes must not interact with each other's respective redox partner in order to avoid futile oxidation cycles. This is not a problem when both functions are performed by the same protein in the case of PDI. As such, eukaryotic cells do not usually have a homologue for DsbD in their ERs. The redox state of PDI is instead regulated via its interaction with Ero1. As mentioned above, Ero1 has a built in disulfide bond that is involved in redox sensing the ER environment which helps the enzyme to adapt its own oxidation rate of PDI [129].

Data sources relevant to oxidative folding

During the last decades the availability of large quantity and quality biological data sources has increased rapidly [130]. With the widespread adaption of omics tools such as genomics and proteomics our understanding of the diversity of life has entered a new era. Many of these data sources contain information that can also be used to progress our understanding of disulfide bond formation and their various roles in different organisms. The following paragraphs will introduce key data sources utilized in this thesis, particularly in Chapters 1 and 3.

UniProt

The UniProt Knowledge base is the central source of protein information for the scientific community. As of November 2022, the database contains over 227 million protein sequences, most of which are derived from large scale sequencing programs [131]. In addition to providing protein sequence data, UniProt also lists a wide range of expertly and automatically annotated information such as gene names, functions and structural information.

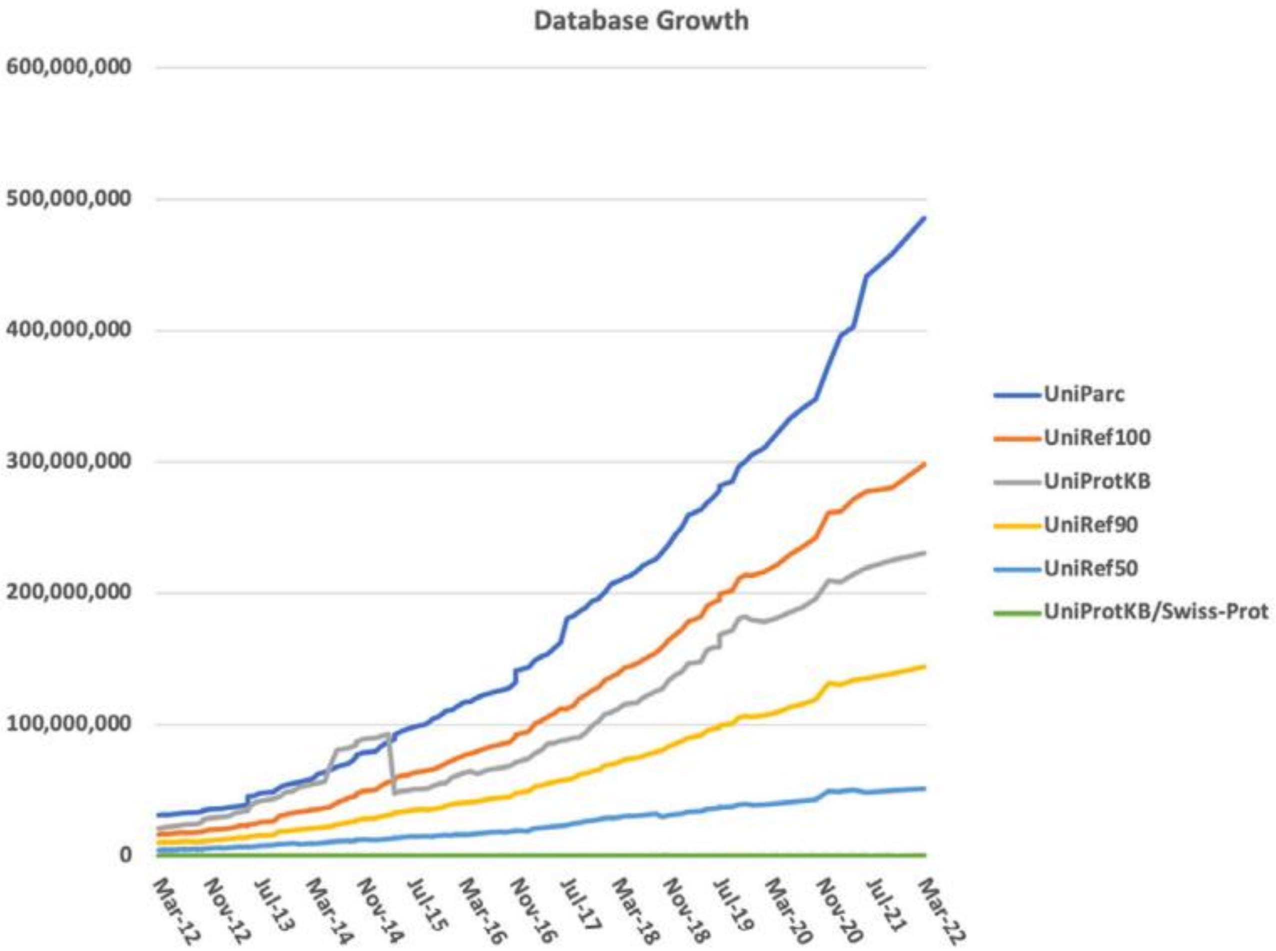


Figure 10 Growth of the UniProt database during the last decade. Y-axis scale is in "total entry counts". UniParc contains the complete collection of sequence filles collected from all UniProt sources. UniRef databases are build on UniParc and reduce the records by clustering and by removing very short length entries. UniProtKB is the central collection of Proteins hosted by UniProt including all available annotations. The Swiss-Prot dataset is the subset of UniprotKB entries which have been manually currated. Figure from [131]

UniProt is also the most commonly available source of information on PTMs. It provides PTM annotations for many proteins, particularly those that are of special interest in particular scientific fields or research areas. Expert-based annotation of proteins has been the central source of annotations in UniProt and over the years this has created the most thorough source of protein annotations available [132]. For well-studied model organisms, the amount of information available (not just PTMs or disulfide bond annotation but annotation in general) is significantly more extensive compared to organisms which are only of niche interest. Regarding the latter, it is the proteins that are linked to that organism's particular sub-area of interest that are expertly annotated while 'standard' proteins are less well studied and therefore less thoroughly curated.

In order to manage the incredible amount of new sequence information being generated by new high-quality sequencing techniques, the UniProt Knowledgebase has put an increased focus on automated systems as the central source of sequence annotations [131]. And together with the continued success and increasing accuracy of prediction tools, has led to a significant increase in the amount of data available on every protein, regardless of its ‘noteworthiness’. UniProt has developed a semi-automated rule-based annotation system which help the annotation process by extrapolating protein information based on better-known similar proteins [131]. Furthermore, UniProt employs a multi-class self-training annotation system (ARBA), which help the expert-based curation process. Nevertheless, the sheer number of proteins requiring annotation and the diversity of their sequences, structures and functions makes it an incredibly difficult job to confidently annotate this huge number of proteins. As such, PTM information is still scarce for most proteins and organisms.

Absolute Protein Quantification

While UniProt is a great source of qualitative protein information, it provides no information on quantity. Organisms often have thousands of different proteins that perform the functions required for the organisms to live, survive and proliferate. However, a single protein is often not enough and therefore some proteins are expressed much more than others. This results in a wide range of protein copy numbers in each cell. Some proteins are not needed in high quantity to perform their - often very specific – function while others, particularly those found in larger multicellular organisms, are required in the millions. As an example, the range of human proteins can range over 10 orders of magnitude from low abundance interleukins (pg/mL) to high abundance albumins (mg/mL) [133]. Some cellular mechanisms such as those linked to protein production itself, are highly linked to the quantitative protein demand of the cell. Oxidative folding is one such mechanism. The qualitative distribution and composition of disulfide bonds on proteins only describes half of the story. The multiplication of each disulfide containing protein with its respective absolute protein abundance is what dictates the demand on the oxidative folding machinery.

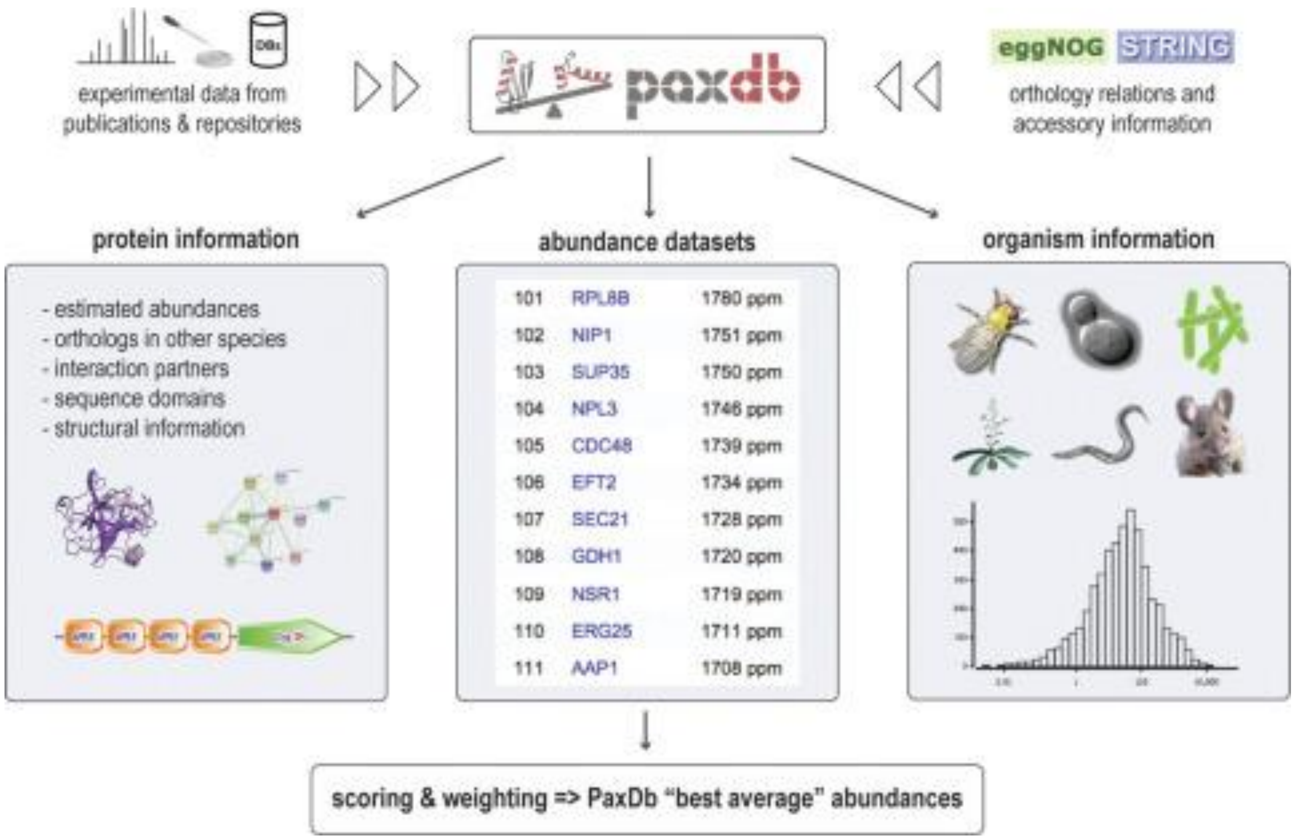
The most widely available source of quantitative protein information are quantitative proteome datasets. Several different techniques have been developed to facilitate the quantification of proteins in samples, all of which are based on GC-MS and use either ESI or MALDI as their ionization source. Relative quantification can be classified as either labelled or label-free methods, whereas absolute quantification always requires some form of labelling [134].

Most available quantitative proteomes provide relative quantitative information, comparing the quantity of proteins between each other, usually a ‘standard’ baseline condition and the condition of interest [135]. This relative quantification is particularly useful when comparing changes in proteome composition in response to outside stimulations such as growth conditions and stress. A key advantage of relative quantitative proteomics (as compared to absolute quantitative proteomics) is lack of need for an internal standard.

When studying protein-aspects such as enzyme kinetics or PTMs, knowledge of the absolute number of proteins in a system is needed. Absolute quantification proteomes also have the advantage of being more comparable between studies since they don't rely on internal comparisons [136]. However, high-quality and high-coverage absolute proteomes require a lot of work and expertise and are still relatively scarce. Nevertheless, when they are available, they provide a great source of information for modelling or other quantitative research approaches.

Many different quantitative proteomes have been published during the last decades and organisations such as the ProteinXchange consortium have worked at standardising the deposition of proteomes and repositories such as PRIDE have established themselves as central sources of proteomics data [137], [138].

With most proteomes investigating the same model organisms of interest it is of particular importance to collect and harmonize this information into a common repository and trying to establish a proteomic baseline for that organism. For absolute quantitative proteomes this work is being done in the PaxDB [139]. This database collects absolute quantification proteomes and also provides an integrated abundance proteome based on all deposited absolute proteomes. These integrated abundance proteomes provide the protein concentrations as part per million (ppm) which make them independent of the cell size and easily comparable between different organisms.


experimental data from publications & repositories
paxdb
eggNOG STRING
orthology relations and accessory information
protein information
abundance datasets
organism information
- estimated abundances
- orthologs in other species
- interaction partners
- sequence domains
- structural information
scoring & weighting => PaxDb "best average" abundances

101	RPL8B	1780 ppm
102	NIP1	1751 ppm
103	SUP35	1750 ppm
104	NPL3	1746 ppm
105	CDC48	1739 ppm
106	EFT2	1734 ppm
107	SEC21	1728 ppm
108	GDH1	1720 ppm
109	NSR1	1719 ppm
110	ERG25	1711 ppm
111	AAP1	1708 ppm

Figure 11 Schematic representation of the Pax Database (PaxDB.). The primary function is to collect and integrate quantitative proteomic data sets and trying to establish a 'consensus' baseline protein abundance. Figure frum Wang et al. [140]

Protein structures

Proteins fold in distinct three-dimensional structures which are essential for their function and stability. And for almost a century, scientists have been able to crystallize proteins and measure their structures with techniques such as X-Ray crystallography or NMR spectroscopy [141]. The resulting structures provide the relative spatial position of the atoms compared to each other and allow scientists to investigate how these proteins perform their function and interact with each other.

The accurate measurement and curation of protein structures is not trivial. Extensive work and experience are required to create protein structures with high levels of accuracy. A central quality attribute of protein structures is the resolution. It describes the smallest observable distance in a structure and is usually given in Angstrom (Å) which is 0.1 nm or 10^{-10} meters. The lower the resolution of a given structure, the more precise are the reported positions of the atoms. High quality structures usually have resolution below 3 Å or even better 2 Å. Another commonly used quality attribute is the R-factor which describes the difference between the observed and the modelled protein structure and 'good' structures usually have R-factors below 20-25 percent [142].

The Protein Data Base

The largest and most widely used repository for protein structures is the Protein Data Base (PDB) [143]. The database contains over 200 000 protein structures (as of September 2023) and also hosts DNA and RNA 3D structures. It provides quality control mechanisms aimed to judge the quality of the deposited structures and in many cases provides links to the corresponding proteins UniProt entries. However, despite great efforts of the data base curators, the large number of structures uploaded by different authors, makes it very challenging to assure the accuracy and quality of each individual protein structure and ensure homogeneous annotation practices.

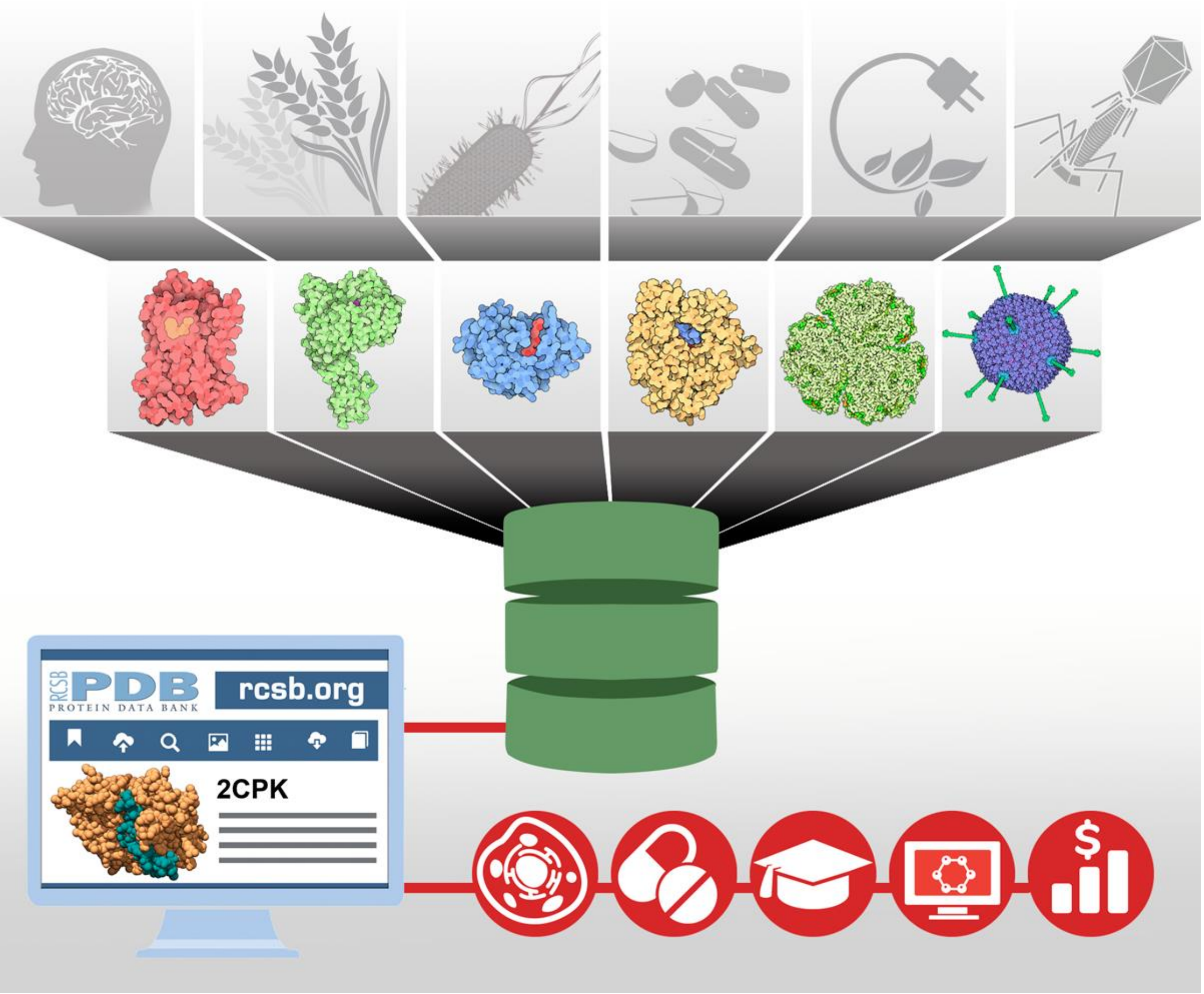


Figure 12 Schematic of the protein database PDB. The PDB is the biggest repository of protein strucutres and provides access to over 200000 protein structures. Figure from the PDB homepage (https://www.rcsb.org/)

AlphaFold

The three-dimensional fold of a protein is determined by its primary structure. And thanks to the ever-increasing number of sequences genomes, more and more protein sequences are discovered every day. However, the traditional practice of determining protein structures by purifying the protein, crystalising it and then measuring the crystal structure via X-Ray crystallography, NMR spectroscopy or Cryogenic electron microscopy (CryoEM) is extremely slow and resource intensive. Therefore, there is a huge discrepancy between the number of known sequences and their corresponding structures [82]. Even though the sequence, in theory, contains all the information it requires to achieve its structure. It should therefore come as no surprise, that efforts to predict 3D structures based on protein sequences have been great over the decades [144]. While chemical interaction simulations do exist, the sheer number of interactions between every amino acid and its many (potential) neighbours and their atoms (and the impact of those interactions on the next neighbours' interactions) are too vast to calculate deterministically. Looking at this from the computational aspect, protein folding is considered a non-polynomial-time problem which are believed to be a set of problems whose computational time increases exponentially with the scale of the problem [145]. In other words, while chemical interaction calculations are feasible for small peptides with relatively few interactions on modern computation clusters, increasing the size of the protein quickly increases the calculation demand to infeasible magnitudes.

AlphaFold 2

The current answer to this dilemma comes in the form of machine learning and neural networks. The power of these tools comes from their ability to solve problems with a high dimensional information-space, usually too high for humans to 'wrap their heads around'. Every two years the Critical Assessment of protein Structure Prediction (CASP) experiment is organised to provide protein sequences together with their newly resolved structures to allow for an independent validation of protein structure prediction software [146]. The last two CASP prediction tournaments in 2018 and 2020 were both won by the artificial intelligence program AlphaFold (AF) by DeepMind [147], [148].

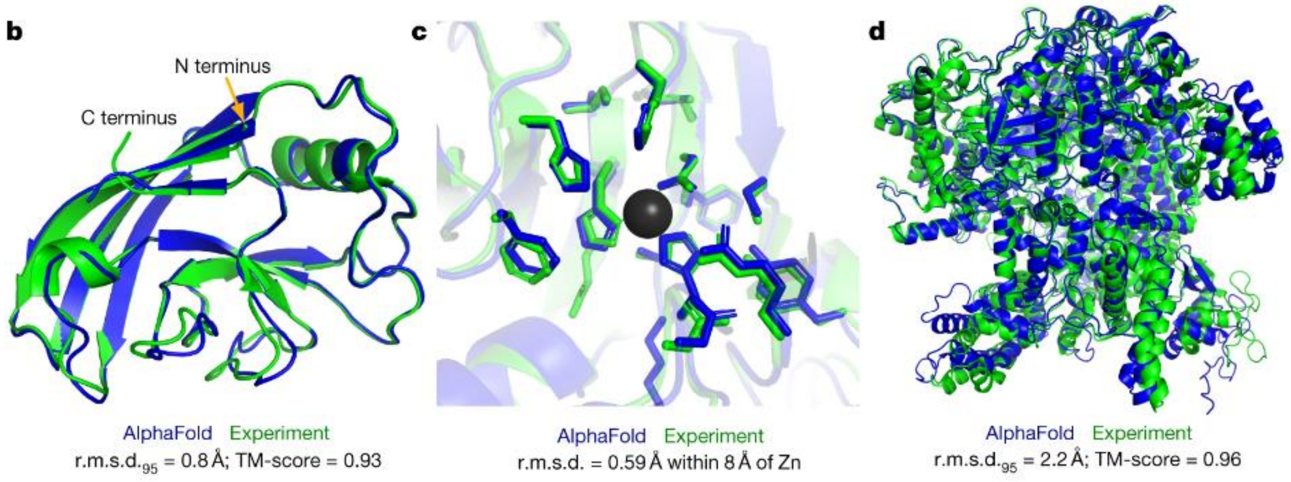

b
N terminus
C terminus
AlphaFold Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93
c
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn
d
AlphaFold Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96

Figure 13 Overlap between the AlphaFold 2 protein structure prediction and the measured protein structure for a selection of different proteins. The proteins were part of the CASP14, the annual protein strucuture preduction competion. Figure from Jumper et al. [148]

AlphaFold 2, which is the current version, uses a combination of multiple sequence alignments and deep neural network architecture to accurately predict protein structures based on protein sequences. In the 2020 CASP prediction tournament, AlphaFold 2 was able to predict protein backbone positions with a median accuracy of 0.96 Å with the second-best performer in the competition reaching 2.8 Å. This jump in prediction accuracy marks a significant increase in usability of the predicted structures.

DeepMind is a subsidiary of Alphabet Inc. (commonly referred to as Google) which is one of the largest providers of computational information processing solutions in the world. And the combination of high-accuracy structure prediction by AlphaFold 2 and the availability of such immense computational capacities has allowed DeepMind to predict the protein structure of over 200 million proteins which is in essence the complete UniProt database and every known protein [131]. However, while this constitutes an incredible new source of information, the accuracy of the predicted structures has to be critically assessed [149]. As mentioned above, AlphaFold is based in part on multiple sequence alignments and extrapolating known structures to unknown ones based on the alignment. As such, the structure predictions for more ‘obscure’ and niche proteins might be less accurate compared to proteins linked to more well-known structures and organisms. Nevertheless, the large quantity of newly available structures constitutes a great new source of information which forms the bases of Chapter 3 of this thesis.

Kinetic Modelling

Together with the development of computation and the computer, biologists have developed computer-based modelling approaches to describe, analyse or predict biological systems. But despite the proven usefulness of computer modelling, it was only at the turn of the century, and with the development of systems biology as its own discipline, that the field of computational modelling in biology gained widespread traction [150]. There are three main reasons for this increased applicability of computational modelling: 1) The increased availability of computers and the resulting increase in computer-literacy of scientists and people in general. 2) The sharp increase in data availability based on new techniques such as genomic sequencing and proteomics. 3) The development of user-friendly software that allows non-mathematicians to create mathematically sound models [151]. On the back of these developments, the field of modelling has become a central part of biological research. There are many different approaches to modelling but in cell biology and biochemistry they can be classified in either stoichiometric models or kinetic models [150]. The former type is mostly used in flux analysis in metabolic networks and paints a static reaction scheme. The kinetic models on the other hand are more information dependant but also allow for a more detailed description of the system.

One of the limiting resources required for kinetic modelling is the availability of enzyme kinetic data. The most common approach in determining enzyme kinetics is based on enzyme kinetic assays [152]. For these assays, the enzyme is first purified and then mixed with a substrate in a reaction buffer. In most cases, it is the enzymatically changed substrate that is then measured to determine the rate of catalysis by the enzyme. This approach is the basis for most known enzyme kinetic values, but it comes with several limitations: the purification step is labour intensive and can be challenging when the enzyme is not stable, hydrophobic or prone to aggregation. A suitable substrate must be identified which should be well quantifiable and representative of the 'real' substrate. The assay conditions must be chosen carefully since temperature, pH and buffer can be highly impactful on the enzymatic activity [153]. Furthermore, the resulting kinetic values are only an estimate for how well the enzyme performs its function in-vivo. The environment in cells is much more complex than any assay buffer could simulate and can be heavily influenced by competing reactions, co-factor availability or local concentrations just to name a few.

In systems biology enzyme kinetics can play an important role in describing a systems behaviour but with the increased availability of high-quality omics data, the top-down approach to systems biology has gained increased applicability [154]. This approach circumvents the need for detailed kinetic data by determining the system based on its overall phenotypical behaviour. This is opposed to the bottom-up approach which aims to describe the pathways and interactions first, and then refer knowledge of the system based on these descriptions. In the following publication (Chapter 1) a novel approach that combines aspects of both bottom-up and top-down systems biology approaches to determine the kinetics and performance of the oxidative folding machinery in *E. coli* has been developed.

Project aims

The aim of my thesis project was formulated as part of the Marie Curie ITN grant ‘SECRETERS’ which consists of 15 PhD positions collectively aimed towards improving expression and secretion of recombinant proteins in microbial cells, formally in *E. coli*, *B. subtilis* and *P. pastoris*. Within this grant the aim of this thesis project was described as “Redox related modelling for *E. coli* and *P. pastoris* cell factories”.

In practice, this project was started by adapting the *S. cerevisiae* based oxidative folding model previously established at the University of Kent by Beal et al. under the supervision of my supervisor Dr. Tobias von der Haar [155]. Chapter 1 is the result of the continuation of this research in the model organism *E. coli*. While the project was initially envisioned to include in-vivo and in-vitro generated data from other project in the grant, it was adapted early in the project to be based on other available data sources instead. The development of the novel approach to enzyme kinetic estimation was developed in part out of necessity from by the disruption caused by the SARS-CoV2 pandemic.

Chapter 2 covers the second organism mentioned in the proposal, *P. pastoris*, and supplements the oxidative folding research of this project with in-vitro data. During this part of the project the key *P. pastoris* oxidative folding enzymes pPDI and ERp38 were investigated.

While chapters 1 and 2 cover the research outlined in the proposal, the research conducted in chapter 3 was envisioned based on experiences and knowledge gained during the thesis project itself. The large-scale calculation of disulfide bonds was only possible as a result of recent achievements in the scientific field of protein structure prediction and required the skills developed during the first part of the thesis to see through.

Furthermore, the review written by the consortia was also outlined in the proposal although the scale and quality achieved exceeded the outlined expectations [5].

Chapter 1

The following chapter is a published paper with the title “A quantitative interpretation of oxidative protein folding activity in *Escherichia coli*”. It was published in December 2022 in the journal Microbial Cell Factories [156].

This paper is the culmination of my work from the first two years of my PhD thesis. It was written within the broader context of the EU funded Marie Curie SECRETERS grant by which my thesis project is funded. The grants aim is to increase the capabilities and applicability of protein production in microbial productions systems, one of which is *E. coli*. Within this context, the research presented in this chapter aims to increase our understanding of the quantitative aspects of disulfide bond formation in *E. coli*. It does so by combining ordinary differential equation-based modelling of the disulfide bond forming machinery in *E. coli* with published quantitative proteomes of this organism. The modelling approach was built on previous work done by Beal et al. in their paper on modelling oxidative folding in *S. cerevisiae* [155]. However, my own paper utilises quantitative proteomics data which is becoming more readily available and is currently underutilized in its potential for modelling and protein production predictions.

The study was designed jointly by me and my supervisor, Tobias von der Haar. All analysis scripts, data analyses, and calculations were done by me. The first manuscript draft was written by me and finalised in discussion with Tobias von der Haar.

Rettenbacher and von der Haar
Microbial Cell Factories (2022) 21:268
https://doi.org/10.1186/s12934-022-01982-3

Microbial Cell Factories

RESEARCH

Open Access

A quantitative interpretation of oxidative protein folding activity in *Escherichia coli*

Lukas A. Rettenbacher and Tobias von der Haar*

Abstract

Background: *Escherichia coli* is of central interest to biotechnological research and a widely used organism for producing proteins at both lab and industrial scales. However, many proteins remain difficult to produce efficiently in *E. coli*. This is particularly true for proteins that require post translational modifications such as disulfide bonds.

Results: In this study we develop a novel approach for quantitatively investigating the ability of *E. coli* to produce disulfide bonds in its own proteome. We summarise the existing knowledge of the *E. coli* disulfide proteome and use this information to investigate the demand on this organism's quantitative oxidative folding apparatus under different growth conditions. Furthermore, we built an ordinary differential equation-based model describing the cells oxidative folding capabilities. We use the model to infer the kinetic parameters required by the cell to achieve the observed oxidative folding requirements. We find that the cellular requirement for disulfide bonded proteins changes significantly between growth conditions. Fast growing cells require most of their oxidative folding capabilities to keep up their proteome while cells growing in chemostats appear limited by their disulfide bond isomerisation capacities.

Conclusion: This study establishes a novel approach for investigating the oxidative folding capacities of an organism. We show the capabilities and limitations of *E. coli* for producing disulfide bonds under different growth conditions and predict under what conditions excess capability is available for recombinant protein production.

Keywords: *Escherichia coli*, Disulfide bond formation, Oxidative folding, Disulfide proteome, Kinetic modelling, Systems biology, Recombinant protein production

Background

Cystine disulfide bonds, covalent connections between the thiol groups of cysteine amino acids, are essential for the correct fold and catalytic activity of many proteins. They are formed by a dedicated cellular machinery, which in native *E. coli* is located in the periplasm [1]. This localisation of the disulfide bond forming machinery, and the high content of reductases and reducing agents such as glutathione in the cytoplasm [2] restrict formation of stable disulfide bonds to the periplasm.

E. coli is a commonly used host for recombinant protein expression, but the inability to form stable disulfide bonds in the cytoplasm can restrict its usefulness for expression of recombinant proteins that require such bonds to adopt the correct fold, including many proteins of strong industrial interest like antibody fragments, growth factors, blood clotting factors and enzymes. To enable production of these proteins in a functional form in *E. coli*, either direction to the periplasm is required, or engineering strategies need to be applied that enable disulfide bond formation in its normally strongly reducing cytoplasm. A number of engineering strategies have been proposed, including deletion of the main cytoplasmic thioredoxin reductases [3–5], or expression of recombinant sulfhydryl oxidases and disulfide bond isomerases [6].

Whether disulfide bonds in recombinant proteins are formed by the native *E. coli* machinery upon export to

*Correspondence: T.von-der-Haar@kent.ac.uk

Division of Natural Sciences, School of Biosciences, University of Kent, Canterbury, UK

the periplasm, or by engineered pathways in the cytoplasm, host cells must continue the formation of essential disulfide bonds in their native proteins at the same time as meeting the added requirements of recombinant protein expression. To our knowledge, this interaction between the requirements of the native and recombinant proteome has so far not been addressed. Quantifying the normal disulfide bond formation requirements in *E. coli*, relating them to the capacity for disulfide bond formation of the native oxidative folding pathways, and understanding how individual recombinant proteins change this balance of required and provided activity, would enable the further optimisation of engineering strategies for enhancing recombinant protein production in this organism.

The disulfide forming machinery in *E. coli* primarily consists of the 'Dsb' family of enzymes. The most abundant of these is the periplasmic DsbA, the cell's primary thiol disulfide oxidoreductase. DsbA contains a catalytic cysteine bond which can introduce new cystines into unfolded substrates in the periplasm, a process that leads to the reduction of the catalytic cystine in DsbA and the formation of two unpaired cysteines. To regenerate the catalytic activity of DsbA, the cysteine disulfide bond is reformed through an interaction with the periplasmic side of the transmembrane enzyme DsbB, which itself transfers the excess electrons to quinones and eventually to molecular oxygen as the terminal electron acceptor.

In addition to the de novo formation of disulfide bonds, cells require means of correcting proteins in which inappropriately formed disulfide bonds have been formed. DsbA does not possess strong chaperone or isomerising activities, which are instead associated with the dedicated isomerases DsbC and DsbG, which differ in terms of their substrate specificities. Whenever DsbA introduces an incorrect disulfide bond into a substrate, DsbC or DsbG are needed to either reduce the incorrect disulfide or isomerise it to form the correct version. While isomerisation is an electron-neutral reaction, the reduction of a misfolded disulfide bond without its ensuing re-oxidation results in an oxidised, inactive isomerase which can be re-reduced and thereby reactivated by DsbD. In similar fashion to DsbB, DsbD is located in the inner membrane and can facilitate electron transfer; however, in this case it transports electrons into the periplasm. The cytoplasmic side of DsbD can transfer a disulfide bond on to thioredoxins which in turn allows DsbD to accept an excess disulfide bond from the isomerases.

The different enzymes of the disulfide forming machinery all have different concentrations and enzyme kinetics. This machinery introduces disulfide bonds into a wide range of host substrates, including recombinant proteins where these contain disulfide bonds. Moreover, differing growth conditions can impact on the oxidative folding machinery as well. The dynamic interactions in this complex system have so far not been fully addressed experimentally. Here, we develop computational models of the oxidative folding process in *E. coli* and use proteomic data to estimate the relationship between provided and required activity under different growth conditions, and to describe and predict the impact of cell engineering and recombinant protein production on the native disulfide machinery.

Results

A quantitative estimation of the oxidative folding machinery in *E. coli*

An initial aim in this study was to estimate the required rates of de novo disulfide bond formation and disulfide bond isomerisation in *E. coli* on the one hand, and the abundance of components of the oxidative folding machinery and the enzymatic activity they provide on the other. We then set out to bring these two elements together by using a quantitative modelling approach to describe the oxidative folding system dynamically.

We initially collected a total of 73 quantitative *E. coli* proteomes from seven different publications [7–13]. Six of these publications provide absolute protein quantification values in the form of protein copy numbers per cell. A seventh study by Peebo et al. provides protein concentrations, which we converted to protein copy numbers per cell by estimating the cell size from data provided in the study as explained in the Materials and Methods section. The collected proteomes cover a variety of growth conditions, differing in media composition, carbon sources, growth rates, *E.coli* strains and stresses imposed on the cells. To select the most suitable datasets for modelling, we analysed the proteomes in terms of their quantitative protein coverage as well as completeness of the additional information provided with the proteomic data. In a first step we excluded proteomes without reported growth rates since this information was essential for estimating protein synthesis rates (see below). In order to estimate proteome coverage in these studies, reported values for total cellular protein count were compared to the corresponding theoretical total cellular protein count based on published calculations [14], which exploit the fact that cellular protein count correlates with both cell size and cellular growth rates. Growth rates were used to estimate the corresponding cell sizes using Eq. (1). The resulting estimates of fractional proteome coverage are displayed in Additional file 2: Fig S1. Proteomes with a quantitative coverage below 50% were not considered for the quantitative modelling part.

Good overall proteome coverage is important for good representation of oxidative folding substrates in the datasets. In addition, we intended to use the datasets also as a source for evaluating the abundance of oxidative folding enzymes, and we therefore specifically investigated how well the Dsb enzymes were represented in them.

The primary oxidase DsbA is an abundant, soluble protein detected in all datasets with a mean abundance of 696 ppm (proteins per million proteins) or around 4000 proteins per cell (Fig. 1). The two isomerases DsbC and DsbG are also soluble proteins. The more abundant DsbC was again represented in all datasets with a mean concentration of 144 ppm, or 800–900 proteins per cell. The less abundant DsbG was not represented in two of the datasets, but in those datasets where it was represented the concentration was reported with a mean abundance of 27 ppm or around 160 proteins per cell. In contrast to the good representation of the soluble enzymes, the membrane-bound DsbB and DsbD had no associated abundance data in the majority of studies, only being covered in 25 and 23 of the 73 datasets, respectively. This was expected, since membrane-associated proteins are frequently under-reported in proteomics datasets if not specifically accounted for during sample preparation [15]. The large whiskers of the boxplots shown in Fig. 1 demonstrate how heterogeneous the observed levels of these enzymes can be. This relatively large variance is in part derived from variations in measurement of the different proteomes, but also from different protein expression levels under different growth conditions.

Because of the poor representation of the membrane-associated oxidative folding enzymes, we estimated their abundance from synthesis rate data reported by Li et al. [16]. This study used a ribosome footprinting approach to characterise protein synthesis activity in the *E. coli* translatome. By determining apparent synthesis rates for all Dsb proteins, we were able to establish a relationship between synthesis rates and steady state levels for DsbA, C and G, which we then used to predict steady state levels for DsbB and D from their synthesis rates. These analyses yielded mean concentrations of 139 ppm and 29 ppm for DsbB and DsbD, respectively (corresponding to around 480 and 100 proteins per cell).

Based on the overall quantitative coverage analysis, the availability of additional information regarding growth rates, cell size and stress conditions as well as the coverage of the key 'Dsb' enzymes, the proteomes reported by Schmidt et al. [13] were selected for the rest of this study (unless noted otherwise). Additionally, only disulfide bonds and the enzyme functions associated with their formation were considered in this study. Other cysteine modifications such as sulfenic acids and their reduction via DsbC or DsbG were not included in this analysis [17].

The *E. coli* disulfide proteome

Following initial quality controls and selection of suitable datasets, we used the proteomics data do estimate the volume of disulfide bonds processed by the Dsb enzymes in native *E. coli* cells. Proteins are substrates if they are located in the periplasm and contain disulfide bonds in their folded state, and we used ancillary data sources to

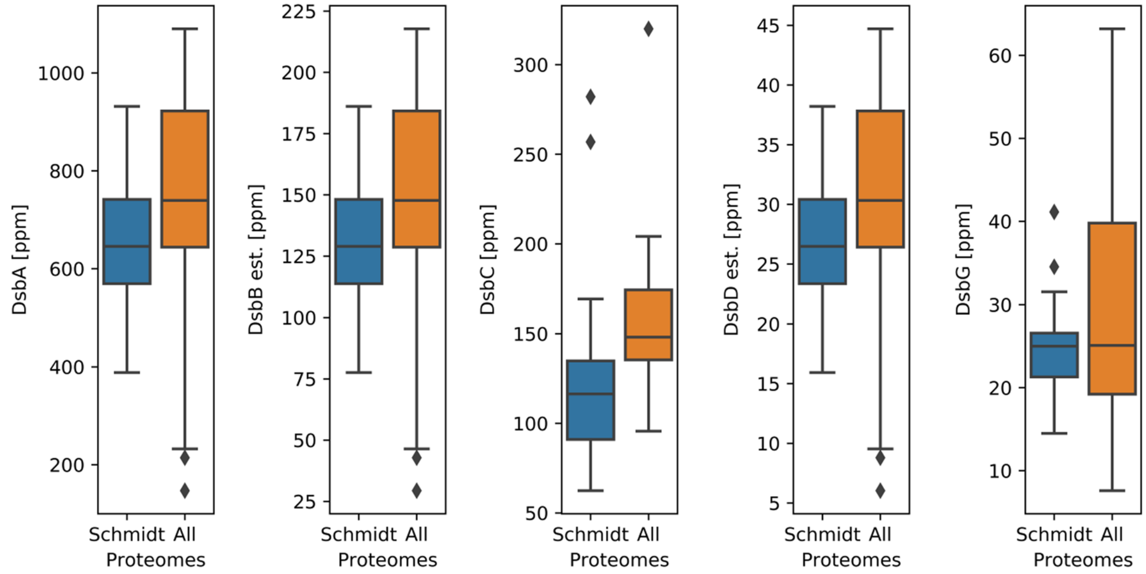


Fig. 1 Boxplots displaying the concentration ranges for the oxidative folding enzymes. For the membrane proteins DsbB and DsbD the estimated concentrations are displayed (DsbBe and DsbDe). The concentrations are given in parts per million (ppm), i.e. enzyme count per million host cell proteins. On the left (blue) the concentration ranges based on the quantitative proteomes by Schmidt et al. [13] are displayed. On the right (orange) the whole range of concentrations from all collected quantitative proteomes are displayed

identify the subset of cellular proteins to which these criteria apply.

Three different sources of known disulfides in the *E. coli* proteome were considered to identify disulfide-bond containing proteins: The Uniprot database [18] which contains curated annotations from the research literature and two proteomics studies [19, 20] which used labelling techniques to identify native *E. coli* disulfide bonds by mass spectrometry. The three data sets collectively identified 360 distinct disulfide bonds that can form in *E. coli* proteins (Fig. 2A); however, only 45 of these were identified in all three (39 of which are periplasmic, Fig. 2B). These initial observations suggest considerable residual uncertainty when it comes to the *types* of disulfide bond that can form in *E. coli* proteins. However, agreement between the studies is better for highly expressed proteins, and in consequence the uncertainty regarding the total *number* of disulfide bonds that are formed in bacterial cells is much lower (Fig. 2B).

To identify proteins located in the periplasm, we used data generated by Loos et al. [21] who trained a machine learning algorithm to annotate protein locations based on a protein's primary amino acid sequence, which suggests locations for 98% of all known *E. coli* proteins with high confidence. We considered all proteins annotated in this dataset as 'secreted' and 'secreted outer membrane' as potential substrates for the Dsb machinery. An overview of numbers of periplasmic disulfide bonds identified in this way is presented in Fig. 2.

A closer look at the reported disulfide bonded proteins confirms the common assumption that *E. coli* has a relatively simple disulfide proteome (Fig. 2A). The total set of 360 reported disulfide bonds are located on 285 different proteins, with only 18 proteins having more than one potential disulfide bond. 174 of the 360 identified disulfide bonds can be allocated to either the periplasm or the outer cell membrane. Of those, 140 are formed by consecutive cysteines, 12 by non-consecutive cysteines and 22 are intramolecular. Out of the 174 disulfides in *bona fide* periplasmic proteins, 97 are described by 2 + sources, and we used this "higher confidence" subset for the kinetic modelling studies described in the following (Fig. 2B). This set of 97 disulfide bonds is located on 82 individual proteins, with only 6 proteins having more than 2 potential disulfide bonds. The list of all 360 identified disulfide bonds and their protein IDs is provided in the Additional file 1: (sheet 2—'DSB data').

The disulfide bond datasets provide a static picture of disulfide bonds in the *E. coli* proteome. However, it does not yet incorporate information on the folding pathways by which individual disulfide bonds are formed, and in particular whether correct bonds are formed immediately through the action of DsbA or following initial incorrect formation and subsequent isomerisation by DsbC or DsbG. To our knowledge there is no quantitative,

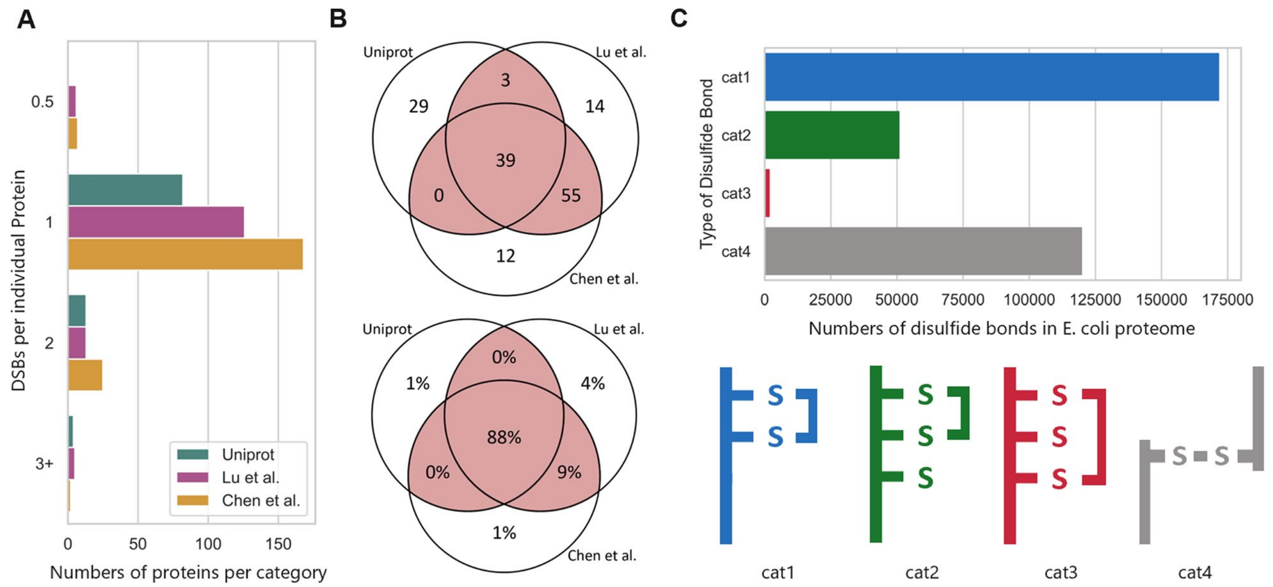


Fig. 2 Summary of the disulfide bond composition of the E. coli proteome. **A** Number of disulfides per protein per literature source. "0.5" refers to intermolecular disulfide bonds. Not shown is one protein with 2.5 disulfide bonds identified in Chen et al. **B** Numbers of disulfide bond containing periplasmic proteins (top) and estimated total numbers of periplasmic disulfide bonds (calculated by multiplying the numbers of disulfide bonds in each protein with protein abundance during growth on glucose, bottom) in three data sources. Disulfide bonds identified by 2 + sources are highlighted in red. Areas in the Venn diagram are not proportional to numbers. **C** Number of disulfide bonds per "folding difficulty" category

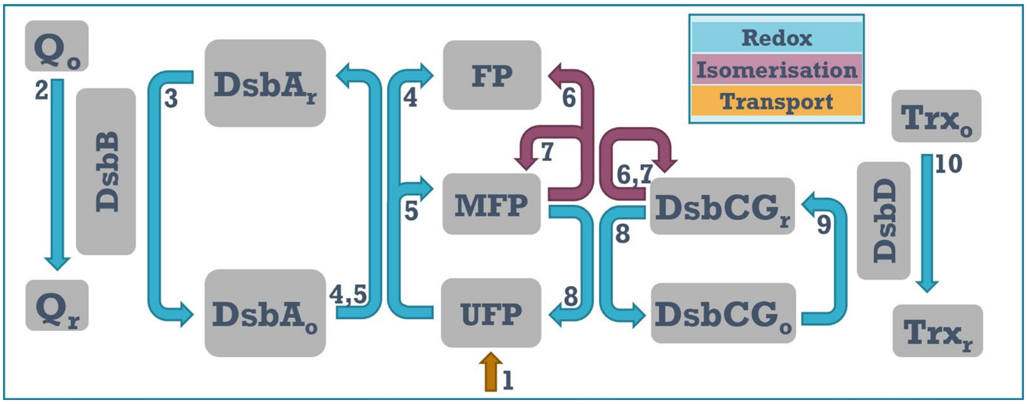


Fig. 3 The oxidative folding model. (1) Synthesis of proteins with a reduced disulfide bond. In the model there are three synthesis rates, one for each difficulty category, and synthesis rates are estimated from the known steady-state abundance of proteins in each category and the growth rate. (2) Oxidation of DsbB via quinone. (3) Oxidation of DsbAr by DsbB. (4) Correct substrate oxidation by DsbAo. (5) Incorrect substrate oxidation by DsbAo. (6) Correct substrate isomerisation by DsbCGr. (7) Incorrect substrate isomerisation by DsbCGr. (8) Reduction of wrongly oxidised substrates by DsbCGr. (9) Reduction of DsbCGo by DsbD. (10) Reduction of DsbD by Thioredoxin

proteome-wide information available to address this question. We therefore introduced a number of semi-quantitative assumptions that we then used to formulate boundary conditions for required levels of isomerase activity in the cell. We categorised disulfide bonds into four categories which we assume are of increasing risk of misfolding, based on relative cysteine locations on each protein's amino acid chain. The first category contains proteins that only have two cysteines, where no mispairing is possible. The second category contains proteins where disulfide bonds are formed between consecutive cysteines, but where additional cysteines exist that could mis-pair with either of the cysteines involved in the native bond. Thirdly we consider disulfide bonds between non-consecutive cysteines, where we assume a more substantial risk of incorrect disulfide bond formation with the intervening cysteine. A fourth category contains intermolecular disulfide bonds and was not further considered in this analysis.

A quantitative evaluation of proteins and associated disulfide bonds in each "folding difficulty" category is shown in Fig. 2C. This graph displays total numbers of disulfide bonds in each category, calculated as the number of proteins multiplied with each protein's abundance. Overall the risk of disulfide-bond related misfolding in *E. coli* appears relatively low, since two thirds of disulfide bonded proteins contain exactly the two cysteines required for the disulfide bond to form, without any scope for misfolding. This finding is consistent in principle with the observation reported above that the enzymes involved in disulfide bond isomerisation (DsbCDG) are expressed at much lower levels compared to DsbAB.

Modelling oxidative folding and isomerisation

To investigate the dynamics of periplasmic oxidative folding processes in *E. coli,* we used an ODE-based computational model with a reaction scheme as depicted in Fig. 3. The model assumes a steady influx of folding substrates into the periplasm, where the substrates are grouped into the different folding categories shown in Fig. 2. The rate of substrate influx into the periplasm is estimated from the quantitative disulfide proteome and the cells' growth rate, which gives the minimum rate of protein synthesis required to maintain a stable proteome in the steady state. In reality additional protein synthesis is required to counteract protein turnover, but under rapid growth conditions this proportion is small compared to that required due to growth [22].

In the model, proteins enter the periplasm in a reduced and unfolded state (*UF*) but can be oxidised by interaction with $DsbA_O$. The outcome of this oxidation can be either the adoption of a misfolded (*MFP*) or a correctly folded (*FP*) state. We assume that the reaction kinetics leading from *UF* to *FP* and *MFP* are identical, but that the probabilities of immediately adopting the *FP* state differ for the three substrate classes, being 100% for 'cat1', 50% for 'cat2' and 0% for 'cat3'. Although actual proteins are more likely to form a continuum of folding probabilities, we assume that these discrete protein categories in the model capture the different types of behaviour observed in biological proteomes both in terms of the types of reactions that occur and (in a first approximation) in terms of their quantitative requirements of de novo folding and isomerase activities provided by the *E. coli* oxidative folding pathways.

Following oxidation of *UF* proteins, DsbA is reduced and can be regenerated by DsbB, which in turn is regenerated in a redox reaction with quinone. If *UF* oxidation leads to formation of *MFPs*, these can further interact with the isomerase DsbCG where one of three things can happen: (1) the disulfide bond is successfully isomerised to form *FP*, (2) the isomerisation is unsuccessful and the protein remains in the *MFP* state or (3) the disulfide bond on the substrate is reduced by the isomerase, returning the protein to the *UF* state. In the latter case, the isomerase itself becomes oxidised and has to be regenerated via DsbD, which in turn transfers the excess electrons onto thioredoxin. In the model representation in Fig. 3, quinones and thioredoxin are depicted in the periplasm for simplicity, even though in reality these reactions take place on the cytoplasmic side of the transmembrane proteins and the inner cell membrane respectively. However, since the model simplifies the reoxidation of DsbB and the reduction of DsbD into pseudo-first order reactions, the actual location of these terminal components is irrelevant in this context.

Each model reaction is represented by an ordinary differential equation and has an associated kinetic parameter. These kinetic parameters dictate the speed of each reaction, and in cases where there is more than one possible reaction outcome, also the ratio between the possible outcomes.

Kinetics

We initially parameterised the model using enzyme concentrations derived from the quantitative proteomics data, and substrate concentrations derived from proteomics data covering disulfide-bonded proteins in the different "folding difficulty" categories outlined above. We assume that clients of the Dsb proteins are predominantly newly translated proteins which are not yet correctly folded. The rates with which such Dsb clients are generated can be estimated from their cellular abundance in the steady state and from the growth rate, since the rate of growth dilution dominates rates of protein turnover in fast growing microbial cells [23].

Based on these known substrate production rates, we then characterised minimal enzyme rate constants that were compatible with the essential requirement of doubling the *E. coli* proteome once per generation. This strategy allows estimating minimal required enzyme activities for the core reactions in the model but may underestimate the actually required activity if futile cycles occur frequently. For example, if a disulfide bond formed by DsbA is resolved again by DsbCG, or if a protein in a mis-folded state is simply transferred to another misfolded state rather than the correctly folded one, enzyme activity is engaged without a net change in substrate or product concentrations. Because we have no information allowing us to estimate the frequency of such cycles, we assumed here that such cycles are rare compared to productive folding events. In our model parameters, we assumed that futile cycles make up one third of all isomerase-catalysed reactions which we considered to be a conservative if not over-estimation of the futile reactions taking place in the cell.

We applied this strategy to all datasets generated by Schmidt et al. [13], thus generating specific minimally required enzyme rate constants for each of the growth conditions investigated in this study. Due to the specific reaction structure employed in the model, the results are returned in the form of apparent association rate constants for the formation of enzymes–substrate complexes. It is worth noting the relationship between these reported apparent rate constants and the actual enzyme rate constants: because we characterise *minimal* rate constants required for the system to cope with the observed substrate influx, these are expected to be slower than actual biochemical enzyme rate constants if an enzyme is not engaged at its maximum capacity. On the other hand, the modelled apparent rate constants cannot be faster than the actual rate constant as this would be biochemically impossible and would indicate that either enzyme or substrate concentrations have been reported incorrectly, or that the model structure has been chosen inappropriately.

To facilitate interpretation, we multiplied the enzyme concentrations for each condition with the modelled apparent association rate constants, thereby creating a pseudo first-order rate constant expressing how rapidly substrates are likely to be processed in each of the different growth conditions (Table 1). Lower first-order rate constants indicate that enzymes engage less readily with their substrate, because in the respective condition the ratio of provided to required activity is lower. In terms of the question we initially asked, high rate constants thus imply a degree of oversupply in the system, which could be exploited for example for more efficient recombinant protein processing.

We observed significant variation in oxidative folding capacity between the different growth conditions (Table 1). The de novo folding reactions (R4,5 in Fig. 3) vary over a two- to four-fold range between conditions, and the isomerisation reactions vary over a six- to eight-fold range. As the demand for oxidation and isomerization changes, so does the demand on the enzymes that catalyse these reactions. What we observe here is that the range of demand for the oxidative system is lower compared to the range of demand on the isomerization system. Each of the investigated proteome datasets reflects specific combinations of growth rates, oxidative folding

Table 1 Reaction rates for oxidative folding in *E. coli*

	Reaction(s)						
	R2	**R3**	**R4 (Cat1)**	**R4,5 (Cat2)**	**R5 (Cat3)**	**R6,7,8 (Cat2)**	**R6,7,8 (Cat3)**
Chemostat							
Chemostat μ = 0.5	0.090	5.388	0.946	0.433	1.297	0.227	0.527
Chemostat μ = 0.35	0.091	4.526	1.093	0.703	2.106	0.673	1.957
Chemostat μ = 0.20	0.040	3.943	0.791	0.535	1.603	0.295	0.886
Chemostat μ = 0.12	0.016	3.181	1.367	0.924	2.770	0.185	0.557
High Growth							
LB	0.228	6.727	1.894	0.337	1.131	0.246	0.584
Glycerol + AA	0.199	6.708	2.519	0.471	1.151	0.364	0.371
Stress							
pH6 glucose	0.152	7.571	1.599	0.733	2.084	0.227	0.486
42 °C glucose	0.048	6.801	1.023	0.370	1.108	0.112	0.240
Osmotic-stress glucose	0.111	6.637	1.107	0.482	1.518	0.192	0.412
Non-stress, sub-optimal							
Acetate	0.046	2.758	0.613	0.394	1.181	0.100	0.299
Fructose	0.079	3.230	1.061	0.425	1.149	0.187	0.301
Fumarate	0.065	3.888	0.988	0.556	1.664	0.158	0.406
Galactose	0.044	3.549	0.857	0.562	1.737	0.165	0.505
Glucosamine	0.083	5.008	1.005	0.537	1.608	0.210	0.515
Glucose	0.080	4.309	0.796	0.404	1.092	0.168	0.292
Glycerol	0.048	5.183	0.892	0.486	1.426	0.122	0.321
Mannose	0.072	3.599	0.722	0.320	0.960	0.119	0.321
Pyruvate	0.042	3.351	0.896	0.576	1.726	0.104	0.312
Succinate	0.065	3.925	0.873	0.481	1.440	0.176	0.453
Xylose	0.077	4.400	0.772	0.413	1.059	0.145	0.344
Median	**0.07**	**4.35**	**0.97**	**0.48**	**1.43**	**0.18**	**0.41**
Range (fold)	14.3	2.7	4.1	2.9	2.9	6.7	8.2
Inter quartile range (fold)	1.9	1.6	1.3	1.4	1.5	1.6	1.6

enzymes and their substrates, and the observed folding capacity most likely changes as the result of relative changes in these parameters.

We observed a clear dependence of the capacity for DsbA reoxidation by DsbB with growth conditions. The highest capacities for this reaction, with substrate processing rates of 6 min^{-1} and higher, were observed during stress conditions which in this dataset included low pH, high temperature and osmotic stress; as well as growth in LB and amino-acid supplemented glycerol, two non-stress conditions with high growth rates. Moreover, the chemostat series of experiments, in which growth rates are directly controlled by the dilution rate with otherwise identical parameters, revealed a strong correlation between the capacity to regenerate DsbA and growth rates (Fig. 4A. Pearson's Product-Moment Correlation Coefficient for the correlation between growth rate and R3 rate parameters for this reaction is 0.98). None of the other reaction rates show similar patterns, and in particular the de novo folding reactions (R4,5) show no clear correlation with the same conditions. Indeed, the majority of the apparent R4/5 rates appears remarkably constant with the lowest inter-quartile range of all reactions. One interpretation of these findings is that *E. coli* cells adjust the expression and subsequent availability of DsbA, including the cells' ability to reactivate this enzyme, in line with demand arising from increasing growth speed, thus enabling the timely processing of inflowing substrates.

Interestingly, the maintenance of DsbA capacity under high growth and stress conditions appears to be the result of distinct set-ups of the oxidative folding machinery (Fig. 4). During chemostat growth, the abundance of both DsbA and DsbB increases, resulting in an overall increase in the capacity to regenerate DsbA (Fig. 4A). This scenario is consistent with an increasing need for de novo folding in response to the increased dilution rates during fast growth, when the influx of DsbA substrates increases proportionally with growth and dilution rates. Under such conditions, both DsbA and the capacity to

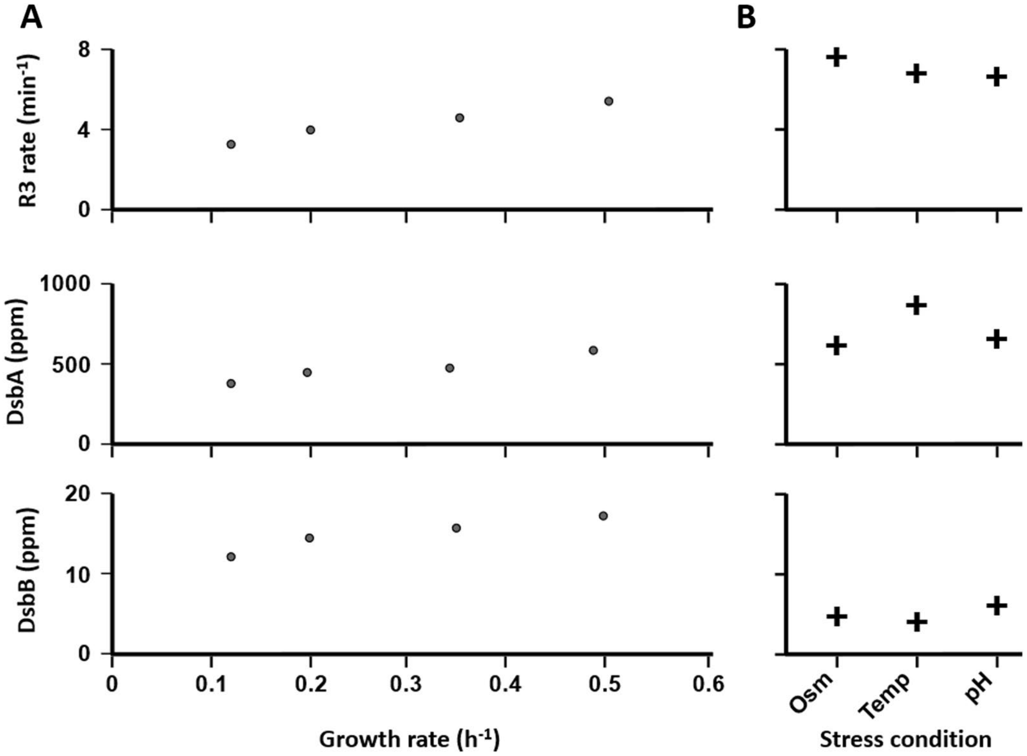


Fig. 4 The rate of DsbB- catalysed reoxidation of DsbA and DsbA/B enzyme abundance under different growth conditions. **A**, during chemostat growth the abundance of DsbA, DsbB and the capacity for DsbA-reoxidation by DsbB (represented here by the modelled R3 rate) all correlate positively with growth rates. **B**, under stress conditions, DsbA abundance and DsbB capacity are high despite low DsbB enzyme concentration

regenerate this enzyme must be adjusted concomitantly. In contrast, the high levels of DsbA under stress conditions appear to be efficiently reoxidised (they have high R3 rates) despite low DsbB concentrations. As the overall capacity of DsbA to catalyse de novo folding is still high under stress conditions compared to normal growth (see apparent R4 rates in Table 1), the most parsimonious explanation is that during stress high DsbA levels are maintained despite relatively low de novo folding demand.

Oxidative folding demand and recombinant protein production capacity

In addition to analysing the native capacity for oxidative folding in *E. coli*, the model allows estimating the capacity of this organism for dealing with additional substrates such as recombinant proteins. We investigated this by increasing the production rate for oxidative folding substrates of varying "folding difficulty", i.e., where the probability of immediately adopting the correct fold decreases and the probability of adopting an incorrect fold which needs to be further corrected by an isomerase increase. We assume that there is no regulatory adaptation to the new substrate. In order to monitor the capacity to process recombinant substrates under particular growth conditions, we increase the rate with which additional recombinant proteins are produced until the cells' native substrates begin to accumulate (we use an accumulation-threshold of the host cell disulfides of 0.5%- as a cut-off for determining the point at which the recombinant protein production starts to impact the upkeep of the host proteome). The amount of recombinant protein that can be introduced before this threshold is reached is displayed in Fig. 5.

The results suggest that the capacity to produce recombinant protein varies strongly with growth conditions, as well as with the requirement for isomerisation. Growth in glycerol, glucose and pyruvate are predicted to allow the highest yields in principle, with an estimated capacity of processing up to 150% of recombinant protein over and above the normal cellular protein complement. When isomerisation steps are required, the yield is predicted to drop strongly, with very-difficult-to-fold proteins that rely highly on isomerase activity showing only a fraction of the predicted yield of an otherwise equivalent easy-to-fold protein. Interestingly, the capacity to process isomerase-requiring substrates differs more strongly between conditions than non-isomerase-requiring ones, and the

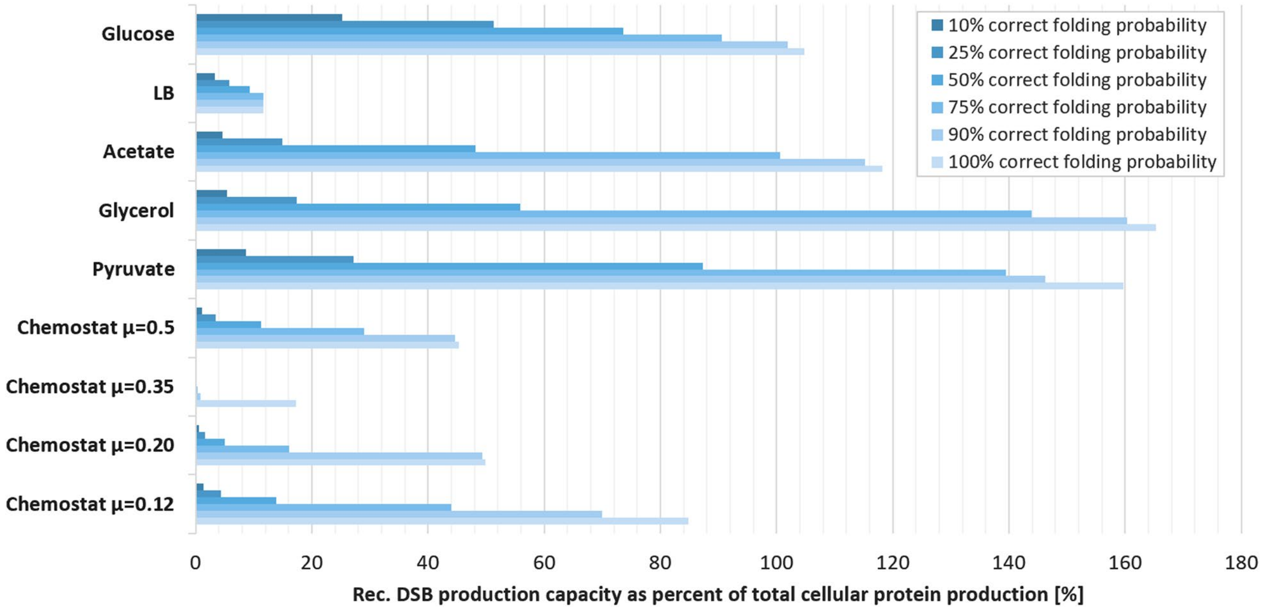


Fig. 5 Predicted recombinant protein production capacity for different growth conditions. Estimated by introducing a theoretical recombinant disulfide bonded protein into the model and simulating its impact on the host proteome in terms of oxidative folding. The calculated synthesis rates are shown as a percent of the overall protein production rate of the cell at each given growth condition. Results shown for different disulfide bond folding complexities of the recombinant protein

relationship is not proportional: for example, during growth in glucose the capacity to process an easy-to-fold substrate is predicted to be less than during growth in glycerol, but the capacity to process a difficult-to-fold substrate is 4–5 times higher during growth in glucose. We assume that these differences reflect different proportions of DsbAB vs DsbCDG concentrations, and indeed examinations of the proteome datasets reveals that whereas concentrations of most Dsb enzymes are comparable under both growth conditions, the concentration of DsbC is increased during growth in glucose.

Discussion

Our model-based investigation of the disulfide-bonded proteome in *E. coli* suggests that cells use different strategies for providing the required oxidative folding capacity under different growth conditions (Table 1 and Fig. 4). Under non-stress conditions, concentrations of DsbA and DsbB appear to be adjusted strictly in parallel, both increasing with faster growth and the resulting requirement for faster processing of de novo folding substrates (Fig. 4). This results in the provided DsbA activity remaining well matched with requirements, as indicated by the relatively low inter quartile range for modelled de novo folding rate constants during non-stress growth (Table 1). During stress conditions, the cellular strategy appears to differ substantially from the non-stress one in that here atypically high DsbA concentrations coincide with atypically low DsbB ones. The elevated DsbA:DsbB ratio should result in an increased cellular concentration of reduced DsbA. This could benefit the cell by facilitating a less ‘generous’ substrate oxidation strategy, only providing disulfide bonds to certain, high affinity substrates. The elevated levels of reduced DsbA could also help alleviate the mis-oxidation stress of the cell by acting as a disulfide bond acceptor for mis-oxidized substrates. An alternative explanation for this elevated DsbA:DsbB ratio could be that cells are preparing for a “growth ready” strategy: this could provide sufficient DsbA to rapidly cater for folding demand when growth rates increase again following stress recovery, then only requiring adjustment of DsbB levels which are much lower than for DsbA and can therefore be increased relatively more quickly.

Although the modelling outputs presented in Table 1 are meant to indicate comparative activity between different growth conditions only, it is useful to ask how they relate to the biochemical rate constant reported in the literature. Darby and Creighton reported biochemical assays in which they used DsbA to fold the three disulfide bond-containing Bovine Pancreatic Tryspin Inhibitor (BPTI) [24], where they observed initial association rate constants above 10^5 M^{-1} s^{-1} whereas the release into the MFP or FP product occurred at estimated rates of 2.7 s^{-1}. DsbA concentrations estimated from the proteome datasets are between 5 and 30 µM under all growth conditions, so that formation of the catalytic complex would typically be rate limiting. Interestingly, the minimally required de novo folding rates we report in Table 1 (R4 and R5) are typically around 1 s^{-1}, whereas the equivalent actual rates revealed by the biochemical experiments would be between 0.5 s^{-1} at 5 μM and 3 s^{-1} at 30 μM DsbA concentration. These analyses assume that DsbA is maintained in a mostly or wholly oxidised state. If, under steady state conditions, a substantial proportion of DsbA was in a reduced state awaiting reoxidation, the ability to catalyse *do novo* folding would be reduced proportionally. Overall, the comparison to available biochemical data suggests that the minimally required and actually provided DsbA capacity is within one order of magnitude and support the notion that DsbB concentrations need to be adjusted in concert with DsbA levels in order to maintain sufficient continuous DsbA activity.

In previous work, Karyoleimos et al. investigated how different recombinant protein production rates affected steady-state expression levels of secreted single chain antibody (scFv) and human growth hormone (hGH), two recombinant proteins with differing disulfide bond patterns [25]. This study reported that gene expression pathways became quickly overwhelmed as expression levels increased, and that efficient processing of the recombinant proteins required adjusting expression levels to the lower range of the rhamnose-inducible system used in that study. While this study focused on the limiting capacity of the Sec translocon as the main bottleneck for the production of periplasmic recombinant proteins, some of the presented data indicate that there are concurrent issues with protein folding, including the apparent inability of the Dsb machinery to produce fully disulfide bonded hGH when expression levels were adjusted to allow for most efficient secretion (cf. Figure 5 B in Karyoleimos et al. 2019). One of the testable predictions resulting from our analyses is that adjusting growth media could be a viable strategy for adjusting the available oxidative folding capacity to the needs of individual recombinant proteins, by allowing to adjust both the overall folding capacity and the ratio of *do novo* folding to isomerase activity (Fig. 5). Our results suggest that the 'ideal' growth conditions for recombinant disulfide bond formation depends strongly on the type of disulfide bond(s) required by the target protein. For 'simple' disulfides, which rely solely on the oxidative folding machinery, growth on glycerol with a moderate growth rate (~ 0.5 μ) appears favourable based on our modelling results. For recombinant proteins with complex disulfide pattern, glucose-based growth with a moderate growth rate (~ 0.6 μ) results in the best yield prediction. While fast growing cells on LB media (1.9 μ) exhibit a low excess capacity for oxidizing recombinant proteins, the highly elevated biomass formation can compensate this disadvantage. In cases where volumetric yields are more important than effective C-source usage, this growth strategy is also predicted to yield good results for both complex and simple disulfide patterns. Chemostat growth seems to be unsuited for the efficient production of disulfide bonded proteins compared to the other observed growth conditions. However, given the relatively close match between required and provided Dsb activity, the success of such media-based strategies will likely remain limited and substantial increases in oxidative folding capacity would require the introduction of engineered systems such as *CyDisCo* [26], which provides oxidative folding capacity in the cytoplasm thereby circumventing both Sec and Dsb bottlenecks.

Conclusion

In summary, our study shows that the combination of genome-wide datasets and modelling approaches can be used to explore feasible rate constants even when information on the actual biochemical rates in a system is limited. This approach is particularly useful for estimating the capacity of cell-wide pathways to cope with both endogenous demand and any additional demands arising from bioprocessing, bioengineering or synthetic biology needs, and the resulting information can be used to inform strain and process engineering strategies to optimise relevant cellular pathways.

Materials and methods

Datasets

Data manipulations including cleaning and merging of different data sources were performed using the Python numpy [27] and pandas [28] libraries.

The quantitative proteomes used in this study were extracted from Additional files available with the relevant literature [7–13]. Information regarding strain types, growth conditions and protein quantification methods were also extracted from these publications. The quantitative data sets from these publications were merged into a single data table based on individual proteins' Uniprot IDs.

To assess the number of *potential* disulfide bonds per protein in the *E. coli* proteome, protein sequences were downloaded from Uniprot [18] and the maximum number of cysteine pairs was calculated as the number of cysteines divided by two, rounded down to the nearest integer.

Actual numbers of disulfide bonds per protein were collected by merging three experimental data sources. One was derived from annotated disulfide bonds in the uniport database. The other two were based on disulfide-labelled proteomes. All three data sources identify the specific cysteins in the protein sequence which form disulfide bonds and the data sets were merged based on these cysteins and the corresponding protein IDs. In cases were a single cystein can form disulfide bonds with different partners, only a single disulfide was considered for the evaluation of the cellular oxidative folding requirement. In cases were a cystein forms an oxidative bond with itself (on another copy of the same protein), the disulfide bond was counted as 0.5 for the quantitative evaluation. Disulfide bonds identified between different proteins were not considered in this analysis.

The classification into "folding difficulties" (see the Results section for details) was then performed by programmatically examining whether a protein had exactly two cysteines and therefore no possibility of misfolding (category 1), more than two cysteines where the disulfide bond was formed between consecutive cysteines (low misfolding probability, category 2), or more than two cysteines where the disulfide bond was form between non-consecutive cysteines (high misfolding probability, category 3).

A master table containing quantity information from the different datasets, numbers of cysteines and disulfide bonds and intracellular protein locations [19–21] was generated by merging the individual data tables listed above (Additional file 1: Table S1).

Evaluation of proteome coverage

Most of the protein datasets provide information in absolute protein numbers, which need to be converted to concentrations for modelling biochemical reactions. Most of the proteome studies used here do not report cell volume but do report cell growth rates, and the *E. coli* cell volume is known to vary linearly with growth rates [29]. We used two publications that report both cell size and growth rate data [13, 30] to create a conversion factor for estimating cell sizes from reported growth rates. Equation 1 is based on growth rates between 0.1 and 1.9 h^{-1} and was used for estimating quantitative protein coverage. Equation 2 is based only on growth rate values between 0.1 and 1 h^{-1} and was used for converting concentration values to protein count per cell values.

$$\textit{cell size}\left[\mu\mathrm{m}^3\right] = 1.44 \cdot \textit{growth rate}\left[\mathrm{h}^{-1}\right] + 1.90 \tag{1}$$

$$\textit{cell size}\left[\mu\mathrm{m}^3\right] = 1.83 \cdot \textit{growth rate}\left[\mathrm{h}^{-1}\right] + 1.74 \tag{2}$$

Estimation of Dsb enzyme abundance

Abundance data for DsbA, C and G were directly extracted from the proteomics datasets via their Uniprot IDs (DsbA, P0AEG4; DsbC, P0AEG6; DsbG, P77202). DsbB and DsbD are membrane-anchored proteins and the membrane association likely leads to depletion of these proteins during sample preparation. Abundance of these proteins was therefore estimated by comparing their synthesis rates inferred from ribosome-profiling based data set by Li and colleges [16]. We assumed that the ratio between synthesis rate and steady-state protein abundance is similar for all Dsb proteins, and there for calculated apparent DsbB/DsbD abundance from their synthesis rates, based on the observed synthesis rate/abundance ratio for DsbA and DsbG.

Kinetic modeling and parameters estimation

Ordinary differential equation (ODE) models were created using the complex pathway simulation software Copasi [31]. These models were imported into python using the tellurium library [32] and converted to the human-readable antimony model script using pycotools3 [33]. The same package was used to simulate model behaviour over time. The resulting data was processed, analysed and displayed using the Python libraries pandas, matplotlib [34] and seaborn [35]. The iterative loops for identifying minimal kinetic parameter sets were also created using the same python libraries.

Estimating minimal enzyme activity required for proteome maintenance

Each proteome has a set of kinetic values that need to be achieved in order to satisfy the cells reported doubling time. The kinetic values are gradually reduced until either substrate accumulation exceeds a certain threshold (0.5% per substrate species) or the theoretical proteome doubling time reaches an 5% increase compared to the reported doubling time of the proteome.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12934-022-01982-3.

Additional file 1: Data tables. Details of disulfide bonded proteins in the *E. coli* proteome. Background colours distinguish data from the seven individual data sources used.

Additional file 2: **Figure S1**. Analysis of quantitative proteome coverage. Calculations based on cell size estimate based on growth rate (eq. 1, main text). Reported total protein counts are compared to the theoretical total protein count derived from the calculated cell sizes and total protein estimate for *E. coli* by Milo Ren (2013).

Acknowledgements

We thank Prof Lloyd Ruddock (University of Oulu) and the members of the Secreters Network for discussion.

Author contributions

LAR and TvdH designed the research plan. LAR conducted the research. LAR wrote the manuscript. TvdH revised the manuscript. Both authors contributed to the article and approved the submitted version. Both authors read and approved the final manuscript.

Funding
This work was funded by the People Programme (Marie Skłodowska-Curie Actions) of the European Union's Horizon 2020 Programme under REA Grant Agreement No. 813979 (SECRETERS).

Availability of data and materials
The original source data used for this research was collected from the referenced publications or the uniport database and can be requested from the authors upon reasonable request. The research results and data used in this article have been included in the manuscript and the Additional file.

Declarations

Ethics approval and consent to participate
This work is purely based on the bacteria ***Escherichia coli*** and does not include any data or research on humans or animals.

Consent for publication
Not applicable.

Competing interests
No competing interests.

Received: 24 October 2022 Accepted: 5 December 2022
Published online: 22 December 2022

References

1. Bardwell JCA. Disulfide bond formation enzymes. Enzymes. 2007;25:111–28.
2. Berkmen M. Production of disulfide-bonded proteins in *Escherichia coli*. Protein Expr Purif. 2012;82(1):240–51.
3. Derman AI, Prinz WA, Belin D, Beckwith J. Mutations that allow disulfide bond formation in the cytoplasm of *Escherichia coli*. Science. 1993;262(5140):1744–7.
4. Prinz WA, Åslund F, Holmgren A, Beckwith J. The role of the thioredoxin and glutaredoxin pathways in reducing protein disulfide bonds in the *Escherichia coli* cytoplasm. J Biol Chem. 1997;272(25):15661–7.
5. Bessette PH, Åslund F, Beckwith J, Georgiou G. Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. Proc Natl Acad Sci USA. 1999;96(24):13703–8.
6. Gaciarz A, Khatri NK, Velez-Suberbie ML, Saaranen MJ, Uchida Y, Keshavarz-Moore E, et al. Efficient soluble expression of disulfide bonded proteins in the cytoplasm of *Escherichia coli* in fed-batch fermentations on chemically defined minimal media. Microb Cell Fact. 2017;16(1):108.
7. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science. 2010;329(5991):533–8.
8. Valgepea K, Adamberg K, Nahku R, Lahtvee PJ, Arike L, Vilu R. Systems biology approach reveals that overflow metabolism of acetate in *Escherichia coli* is triggered by carbon catabolite repression of acetyl-CoA synthetase. BMC Syst Biol. 2010;4(1):166.
9. Arike L, Valgepea K, Peil L, Nahku R, Adamberg K, Vilu R. Comparison and applications of label-free absolute proteome quantification methods on *Escherichia coli*. J Proteomics. 2012;75(17):5437–48.
10. Wiśniewski JR, Rakus D. Multi-enzyme digestion FASP and the 'total protein approach'-based absolute quantification of the *Escherichia coli* proteome. J Proteomics. 2014;109:322–31.
11. Peebo K, Valgepea K, Maser A, Nahku R, Adamberg K, Vilu R. Proteome reallocation in *Escherichia coli* with increasing specific growth rate. Mol BioSyst. 2015;11(4):1184–93.
12. Soufi B, Krug K, Harst A, Macek B. Characterization of the *E. coli* proteome and its modifications during growth and ethanol stress. Front Microbiol. 2015;6:103.
13. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent *Escherichia coli* proteome. Nat Biotechnol. 2016;34(1):104–10.
14. Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values. BioEssays. 2013;35(12):1050–5.
15. Kongpracha P, Wiriyasermkul P, Isozumi N, Moriyama S, Kanai Y, Nagamori S. Simple but efficacious enrichment of integral membrane proteins and their interactions for in-depth membrane proteomics. Mol Cell Proteom. 2022;21(5):100206.
16. Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 2014;157(3):624–35.
17. Roos G, Messens J. Protein sulfenic acid formation: from cellular damage to redox regulation. Free Radical Biol Med. 2011;51(2):314–26.
18. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–9.
19. Lu S, Fan SB, Yang B, Li YX, Meng JM, Wu L, et al. Mapping native disulfide bonds at a proteome scale. Nat Methods. 2015;12(4):329–31.
20. Chen ZL, Meng JM, Cao Y, Yin JL, Fang RQ, Fan SB, et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. Nat Commun. 2019;10(1):1–12.
21. Loos MS, Ramakrishnan R, Vranken W, Tsirigotaki A, Tsare E-P, Zorzini V, et al. Structural basis of the subcellular topology landscape of *Escherichia coli*. Front Microbiol. 2019. https://doi.org/10.3389/fmicb.2019.01670.
22. Nath K, Koch AL. Protein degradation in *Escherichia coli*: i measurement of rapidly and slowly decaying components. J Biol Chem. 1970;245(11):2889–900.
23. von der Haar T. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst Biol. 2008;2(1):87.
24. Darby NJ, Creighton TE. Catalytic mechanism of DsbA and its comparison with that of protein disulfide isomerase. Biochemistry. 1995;34(11):3576–87.
25. Karyolaimos A, Ampah-Korsah H, Hillenaar T, Mestre Borras A, Dolata KM, Sievers S, et al. Enhancing recombinant protein yields in the *E. coli* periplasm by combining signal peptide and production rate screening. Front Microbiol. 2019. https://doi.org/10.3389/fmicb.2019.01511.
26. Hatahet F, Nguyen VD, Salo KEH, Ruddock LW. Disruption of reducing pathways is not essential for efficient disulfide bond formation in the cytoplasm of *E. coli*. Microbial Cell Factories. 2010;9(1):67.
27. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357–62.
28. McKinney W. Data structures for statistical computing in python. Proceedings of the 9th python in science conference. 2010 56–61.
29. Schaechter M, Maaloe O, Kjeldgaard NO. Dependency on medium and temperature of cell size and chemical composition during balanced growth of salmonella typhimurium. J Gen Microbiol. 1958;19(3):592–606.
30. Volkmer B, Heinemann M. Condition-dependent cell volume and concentration of *Escherichia coli* to facilitate data conversion for systems biology modeling. PLoS ONE. 2011;6(7): e23126.
31. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, et al. COPASI–a complex pathway simulator. Bioinformatics. 2006;22(24):3067–74.
32. Choi K, Medley JK, König M, Stocking K, Smith L, Gu S, et al. Tellurium: an extensible python-based modeling environment for systems and synthetic biology. Biosystems. 2018;171:74–9.
33. Welsh CM, Fullard N, Proctor CJ, Martinez-Guimera A, Isfort RJ, Bascom CC, et al. PyCoTools: a python toolbox for COPASI. Bioinformatics. 2018;34(21):3702–10.
34. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9(03):90–5.
35. Waskom ML. Seaborn: statistical data visualization. J Open Source Software. 2021;6(60):3021.

Chapter 2

Characterisation of PDI and ERp38, two protein disulfide bond isomerases from *Pichia pastoris*

Introduction

The work presented in this chapter has derived from a collaboration with my colleague Arianna Palma, who is part of the same Marie Curie Grant (SECRETERS) that funded my own PhD project. Her PhD project focuses on the investigation of the oxidative folding in the methylotrophic yeast *P. Pastoris*. During my secondment at the University of Oulu both Arianna and I were under the supervision of Prof. Lloyd Ruddock. The protein expression constructs for PDI and ERp38 used in the following research were constructed by Arianna Palma while the construct for BTPI expression was already available at the hosting laboratory in Oulu. The ERp38 protein was purified by her while all other proteins were purified by me. All other work as well as all the results shown in this Chapter were done by me. Some of the data presented in this Chapter has also been recently published [129].

The PDI family

As previously mentioned in the thesis introduction, the PDI enzyme family is central to disulfide bond formation and isomerisation in eukaryotes. The defining features of this family is the thioredoxin fold together with a CXXC motif at its core [94]. For *Homo sapiens*, *S. cerevisiae* and *P. pastoris* the most studied PDI family member (often referred to as PDI1) has the central motif CGHC. However, in all three members PDI1 is not the only PDI family representative present in their genome. In fact, in *H. sapiens* roughly 20 different family members have been identified, 5 in *S. cerevisiae* and 5 in *P. pastoris* [157]. All of these enzymes have at least one thioredoxin domain with a CXXC motif. How many (thioredoxin-) domains each enzyme has as well as the two amino acids found in between the two cysteines of the motif (-XX-) can vary. Both these factors have a major impact on the function of the enzyme. The two amino acids in between the two central cysteine amino acids have a particularly important role in modulating the redox potential of the two neighbouring cysteines [158]. This in turn determines the oxidative strength of a disulfide bond formed between the two cysteines or between one of the cysteines and another interacting proteins' cysteine. Also, the variations in domain structure of each family member will have large implications on the specific function each protein has. Steric limitations can determine if thioredoxin folds can interact with each other and which substrates can be oxidised, reduced or isomerised by a given PDI [159]. This is essential in regulating the redox state of compartments as well as avoiding interactions that would result in futile cycles of repeated reduction and oxidation.

The primary function of PDI is oxidative folding in the ER, however, the enzymes and their functions have been associated with a wider role in cell biology and medicine [160]. ER-Associated Degradation (ERAD), an essential part of the ER quality control system, is partially modulated by PDI enzymes [161]. In ER Ca^{2+} homeostasis, a PDI family member (ERp44) plays an important role by inhibiting the channel protein Inositol 1,4,5-trisphosphate receptors (IP3Rs) via oxidation of its free cysteines [162]. IP3Rs are involved in several biological functions and disfunction has been linked to diseases such as Alzheimer's disease [163]. As a final example, the PDI family member ERp57 is deeply involved in the presentation of intracellular derived antigen peptides on the cell surface. The heavy chain of both the Major Histocompatibility Complex (MHC) as well as members of the CD1 (cluster of differentiation 1) membrane glycoprotein family (both of which present antigens on the cell surface) interact with ERp57 for DSB formation and in the case of CD1d the corresponding cysteine has been shown to be highly conserved [164], [165].

Yeasts (and fungi as a whole) are of great scientific interest. They are used as model organisms for fundamental evolutionary and medical research, as well as in protein production due to their incredible ability to secrete high amounts of proteins [166]. Since the turn of the century the methylotrophic yeast *P. pastoris* has been established as a commonly used protein production host due to its ability to metabolise methanol as well as its powerful *AOX1* promotor [167]. All yeasts are eukaryotic organisms and therefore contain many of the same cellular compartments and similar enzymes as humans do. One example of this is the ER and the PDI enzyme family. Mammalian PDI was first described in the 60s and has been intensively researched since [168] while *P. pastoris* PDI was described much later in the early 2000s [169]. Therefore, research into the yeasts' oxidative folding apparatus and the functionality of its PDI family members is more relevant than ever in order to facilitate the current emergence of *P. pastoris* as a recombinant protein production host. Furthermore, comparing the oxidative folding capabilities of *P. pastoris* to that of humans will help improve our scientific understanding of human-like protein production in *P. pastoris*.

BPTI

Oxidative folding and protein folding in general have been intensely researched for decades. One particular protein, the bovine pancreatic trypsin inhibitor (BPTI), has been a key model protein in studying both [170]. BPTI is a relatively short protein consisting of roughly 58 amino acids and functions (as the name suggests) as an inhibitor for the protease trypsin or trypsin-like proteases. Most importantly though, BPTI has 3 disulfide bonds which are essential for it to fold correctly. Reduced cysteines can be oxidised and blocked from forming disulfide bonds via trapping agents such as N-Ethylmaleimide (NEM). And since the formation of disulfide bonds is heavily linked to the folding process of BPTI, experiments based on cysteine trapping and isolation of folding intermediates has been of key interest in protein folding studies [171], [172]. And as such it was the first protein to have its folding pathway described in detail.

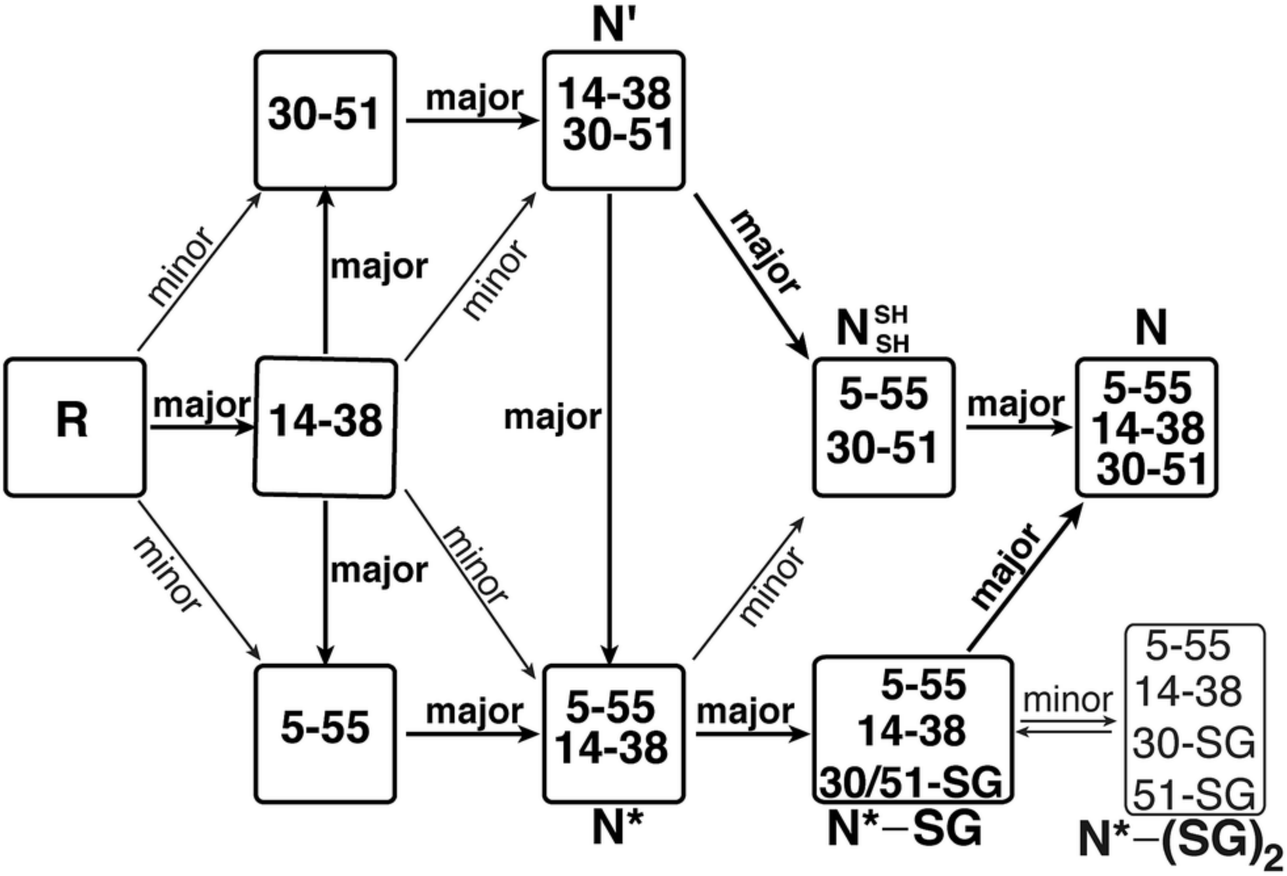


Figure 14 Known folding pathways of the three DSB if BPTI. The pathways are categorised into more common (major) and less common (minor) pathways. Figure from Mousa et al. [173]

The three native disulfide bonds of BTPI are between cysteines 5 and 55, 14 and 38, 30 and 51 of the protein sequence (commonly written as [5-55], [14-38] and [30-51] respectively). As depicted in Figure 14, disulfide bonds are mostly not introduced in their final position, but rather subject to inner-protein reshuffling. The [14-38] disulfide bond is most readily oxidised, and the more common folding pathways are reliant on a transfer of this disulfide bond to one of the other two positions. In most folding conditions, the first two disulfide bonds are introduced relatively fast while the last reshuffling step (resulting in [5-55] and [30-51]) is the time limiting step. The final oxidation of the third disulfide bond is then again relatively fast.

Glutathione

Glutathione (GSH) is a tripeptide functioning as a cellular antioxidant in many organisms, including humans and yeasts [174], [175]. It is a highly abundant thiol which is formed by the three amino acids glutamate, cysteine and glycine with a non-typical peptide bond between the two former amino acids [176]. The monomeric form of GSH is readily oxidised to the dimeric form of GSSH via a DSB in-between the cysteine side chains. Together the two forms function as the central redox couple in mammals controlling the oxidative state of the cell [177].

Project aim

Oxidative folding is of central interest for both understanding and utilizing cells. The yeast *P. pastoris* is an important protein production host and it is therefore of great scientific interest to investigate the organisms native oxidative folding machinery together with its enzymes. As in most eukaryotic cells, PDI1 is the most abundant protein of the PDI family members identified in *P. pastoris*. The enzyme Erp38, on the other hand, is found in relatively few other organisms. As such, in the following chapter both enzymes will be investigated regarding their oxidative folding capabilities. This will be facilitated via observing the enzymes' ability to oxidise and isomerise the well-studied substrate protein BPTI.

Methods and Materials

Proteins

Table 2 List of used proteins together with their length, weight and UniProt accession number. For the length and weight parameters the values corresponding to the his-tagged versions of the proteins are displayed in brackets.

Protein	**Abbrev.**	**Length [amino acids]**	**MW [kDa]**	**UniProt ID**
***P. pastoris* PDI**	pPDI	517 (524)	57.8 (58.7)	B3VSN1
***P. pastoris* ERp38**	ERp38	369 (376)	42.0 (42.9)	C4QWA2
***H. sapiens* PDI**	hPDI	508 (515)	57.1 (58.1)	P07237
BPTI	BPTI	100 (107)	10.9 (11.9)	P00974

Constructs

The cloning was conducted by my colleague Arianna Palma as part of her own PhD thesis and their exact construction can be found in her thesis (not yet completed at the time of writing this) as well as the corresponding paper [129].

pET23 was used as the backbone for the construction of the plasmids with the lactose (and IPTG) inducible Tac promoter. The constructs were propagated in the *E. coli* XL1-blue (Agilent) strain and subsequently transferred into the *E. co*li BL21(DE3) strain for expression. For BPTI production an additional plasmid with the CyDisCo construct was also included in order to avoid inclusion body formation. All proteins where his-tagged for IMAC purification.

Protein Production and Purification

Growth media: AIM LB Broth without trace elements (Formedium), 0.8 % glycerol

IMAC loading buffer: 20 mM phosphate (9.636 mM monosodium phosphate; monohydrate and 10.360 mM disodium phosphate, heptahydrate), 150 mM NaCl, pH 7.0

IMAC washing buffer: 20 mM phosphate (as above), 500 mM NaCl, 50 mM imidazole, pH 7.0

IMAC elution buffer: 20 mM phosphate (as above), 150 mM NaCl, 500 mM imidazole, pH 7.4

RP loading buffer: 2 % acetonitrile, 0.1 % Trifluoroacetic acid (TFA)

RP elution buffer: 90% acetonitrile, 0.1 % TFA

Growing cells for pPDI & hPDI

E. coli BL21(DE3) cells transformed with the respective expression plasmids were thawed and plated out on agar plates from which a single colony was picked and transferred to 10 mL autoinduction growth media and grown at 30 degrees Celsius under constant shaking (incubator) for 4 hours. Both preculture and inoculum were grown with Ampicillin (Amp) present at 100 µg/mL. 5 times 100 mL of culture media (autoinduction) aliquoted into high-yield growth flasks were inoculated with 1% inoculum at OD_{600} of roughly 1 and grown for 24 hours. Lastly the growth was stopped by centrifugation at 4000 rpm for 30 minutes at 4 degrees and the pellets stored at -80 degrees until further use.

Growing cells for BPTI

The *E. coli* strain with the BPTI construct was unfrozen from the strain collection and plated on agar plates containing both Amp and chloramphenicol (Kan) present at 100 µg/mL and 35µg/mL respectively. The inoculum was grown in autoinduction media overnight and the 5 100 mL growth flasks inoculated at an OD_{600} of roughly 1 with 1 mL each. The inoculated high-yield growth flasks were grown at 30 degrees and 250 rpm for 24 hours. Afterwards the cells were separated from the growth media via centrifugation for 30 minutes at 4000 rpm and 4 degrees. The cell pellets were stored at -80 degrees.

Protein Purification

The frozen cell pellets were resuspended in IMAC loading buffer to an OD_{600} of roughly 30. The cells were disrupted via sonication. Total sonication time was 3 minutes with 40% intensity / amplitude and a ½ inch tapped sonication horn. Each sonication interval lasted for 5 seconds interspaced by 25 seconds of no sonication to avoid extensive head development. The sonication was performed under constant cooling in an ice bath. The ruptured cells were then separated from the soluble contents via filtration with a 0.45 μm syringe filter.

His-tagged proteins were separated from the rest of the host cell proteins via Immobilized Metal Affinity Chromatography (IMAC) using a HiTrap Chelating HP column (GE Healthcare) with a column volume of 5 ml. The IMAC column was primed with Ni-Chloride solution which turns the column visibly green and ready for use. Afterwards the column was washed with 2 column volumes (CV) water and 2 CV loading buffer. The soluble cell lysate was then loaded onto the column at 2 mL / minute overnight. After loading the column was washed with 2 CV loading buffer followed by 4 CVs of washing buffer and another 2 CVs of loading buffer. The elution was done over the duration of 20 CVs with a flow rate of 2 mL / minute and from 0 to 100 % elution buffer and the individual elution fractions were collected. The collected fractions were then analysed for target protein content via sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS PAGE). Based on the SDS PAGE results the fractions with high target protein content were selected and pooled. The pooled fractions were then buffer exchanged back into the loading buffer via centrifugation filtration with a molecular weight cut-off of 10k for pPDI and hPDI and a cut-off of 3k for BPTI. The buffer exchange was repeated 5 times and the final step was also used to concentrate the target protein down to a high protein concentration. For pPDI and hPDI, the final protein concentration was then measured, and the enzymes were aliquoted, shock frozen in liquid nitrogen and stored at -80 degrees.

After IMAC chromatography, BPTI was buffer exchanged back into the IMAC loading buffer and diluted in 6 M guanidinium chloride solution (pH 7) for a final concentration of 4 M. Then Dithiothreitol (DTT) was added at 5 times molar excess for each mole of BPTI in order to fully reduce the BPTI. After 30 minutes of incubation the reduced BPTI was loaded onto a reverse phase column (Source 5 RPC St 4.6/150 by Amersham Biosciences) which had been equilibrated with 2% ACN (plus 0.1%

Trifluoroacetic acid (TFA)) loading buffer. BPTI was loaded onto the column with 0.2 mL per minute and eluted with 90% ACN elution buffer (plus 0.1% TFA). The collected elution fractions were tested for target protein content via SDS PAGE and pooled accordingly. The selected fractions were then lyophilised overnight and resuspended in 10 mM HCl the next day. As a final step the protein concentration of BTPI was measured before it was aliquoted, shock frozen in liquid nitrogen and stored at -80 degrees until further use.

BPTI refolding assay

The refolding assays for BPTI were performed in 0.1 M sodium phosphate buffer at pH 7. Additionally, EDTA was present at a final concentration of 1 mM. The redox environment was controlled with the addition of glutathione in both its reduced form (GSH) and its oxidised form (GSSG). Unless stated otherwise the final concentration for GSH and GSSG was 2 mM and 0.5 mM respectively. Both glutathione stock solutions were made in 0.1 M phosphate buffer, pH 7 and stored frozen as aliquots to avoid refreezing the stocks. Depending on the experiment, pPDI, hPDI, ERp38 or an equivalent volume of buffer (blank) was added to this solution and phosphate buffer was added to the mixture for a total volume of 100 µL. The second assay mixture only contained the reduced BPTI at a final assay concentration of 50 µM together with phosphate buffer for a final volume of 100 µL. The refolding reaction was started by adding the BTPI mixture to the other assay mixture containing the glutathione and enzyme. After adding the mixtures together and quickly mixing them together via repeated up and down pipetting, the assay mixture was aliquoted into 10 times 20 µL samples and placed on a 25 degrees heat block. The first sample was immediately stopped (time point 0) by adding the aliquot into a reaction tube with 20 µL stopping solution already present. The stopping solution contains 100 mM N-Ethylmaleimide (NEM) dissolved in 0.1 M phosphate buffer at pH 7 and was made fresh each day. The remaining 9 assay samples were stopped via addition of 20 µL stopping solution at 2.5, 5, 7.5, 10, 15, 20, 40, 60 and 120 minutes. Each of the 10 aliquots were left to incubate with the stopping solution for 1 minute before flash freezing them in liquid nitrogen and storing them at -80 degrees.

LC-MS

The BTPI assay samples were thawed and diluted 1:5 in 1% TFA solution. Each sample was measured on a LC-MS system using a ACQUITY UPLC Protein BEH C4 Column with 2.1 mm inner diameter and 100 mm column length together with water as loading buffer and acetonitrile as elution buffer both with 0.1 % formic acid. 2 µL sample volume was used for injection. The elution profile was set from 3% to 40% elution buffer in 15 minutes. Following the separation via LC the samples were analysed in the MS using electron spray ionisation followed by an orbitrap system for ion trapping and mass analysing. The resulting mass fragment information was then analysed using the

Xcalibur Data Acquisition and Interpretation Software from Thermo-Fisher Scientific. The m/z range was set from 800-2000 and the spectrum was integrated between 2-10 minutes of retention time. The folding state of the BTPI was then calculated from the measured molecular weights of the detected BPTI fragments.

Kinetic Calculations

The enzyme kinetic calculations are based on the triplicate timeseries assay data and are fitted to the respective functions by sum of squared error loss function using the python package scipy [178]. The 3 equations describing the distinct BPTI folding steps are listed here (Eq. 1-3) and again further down together with the corresponding results.

$$[BPTI_{1S}] = [BPTI_{0S}] * \ e^{-k_1 * t}$$

(Eq. 1)

$$[BPTI_{2S}] = \frac{[BPTI_{0S}] * k_1}{k_2 - k_1} * \left(e^{-k_1 * t} - \ e^{-k_2 * t}\right)$$

(Eq. 2)

$$[BPTI_{3S}] = [BPTI_{0S}] - \left([N^*] * e^{-k_1 * t} + [N'] * e^{-k_2 * t}\right)$$

$$with\ [N'] + \ [N^*] = \ [BPTI_{0S}]$$

(Eq. 3)

Peptide Oxidation

The 10 amino acid long peptide used for detecting disulfide bond formation was already readily available at the hosting university [179]. The assays were conducted in freshly made pH 7.0 McIlvaine buffer based on a mixture of 0.2 M Na_2HPO_4 and 0.1 M citric acid stock solutions (roughly 200 and 40 mL respectively). The fresh buffer was place into a fluorescence cuvette together with GSSG, GSH and the peptide substrate for final assay concentrations of 0.5 mM, 2 mM and 5 µM respectively. After 1 minute equilibration time the assay was started by adding the enzyme to a final concentration of 0.05 µM for pPDI and 0.03 µM for ERp38.

The fluorescence measurements were conducted in a FluoroMax-4 Spectrophotometer with a 310 µL glass cuvette at 25 °C. excitation wavelength was set at 280 nm and emission wavelength at 350 nm. The size of the excitation slit was 1 nm and the emission slit 5 nm. The total measurement time was set at 900 seconds with measurements taken every 6 seconds with an integration time of 1 second. To avoid unnecessary bleaching effects the shutter was set to close in-between the individual measurements.

Results

Protein Production

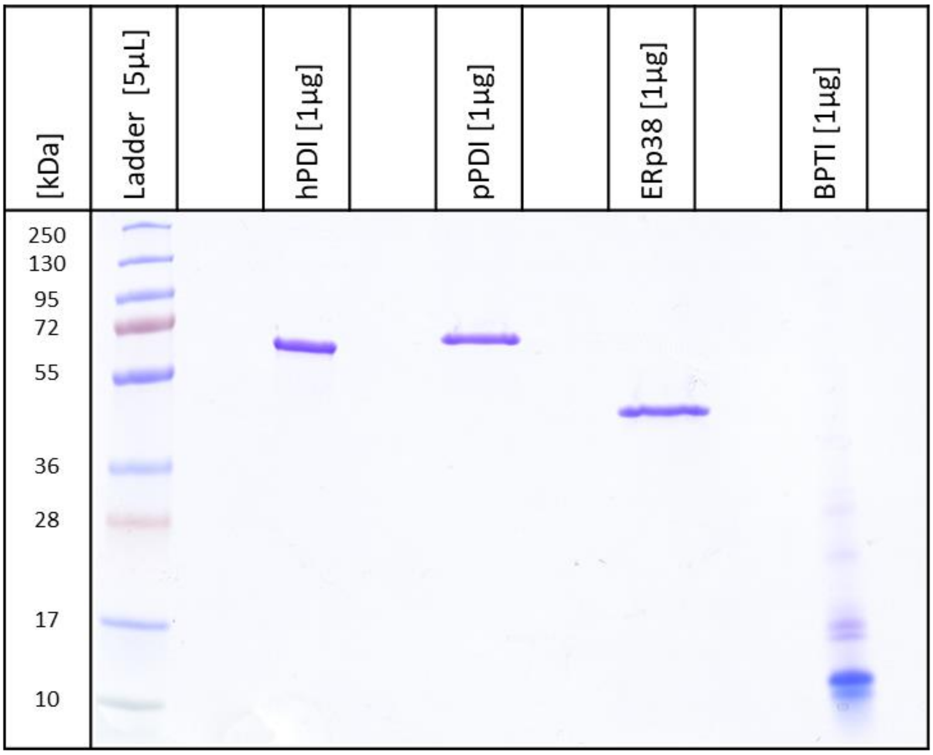


Figure 15 SDS PAGE gel of the purified proteins used in this Chapter. The samples were prepared in reducing loading SDS PAGE buffer and 1 µg of each protein was loaded each. The expected weights are 58.1, 58.7, 42.9 and 11.9 kDa for hPDI, pPDI, Erp38 and BPTI respectively.

Production and purification of hPDI

Following the expression of His-tagged protein in *E. coli*, the final OD_{600} for the 500 mL of cells producing hPDI was measured at an average of 19.5. After purification the final concentration was 13.83 mg/mL in a total of 1.5 mL as determined by microvolume spectrophotometer (NanoDrop). The final concentration in molarity was 235.96 µM and the final product quality was assessed via SDS PAGE (Figure 15).

Production and purification of pPDI

Following the expression of His-tagged protein in *E. coli*, the final average OD_{600} at harvest was 12 for the 400 mL of growth media. After purification the final yield was 14.5 mg/mL in roughly 3 mL as determined by microvolume spectrophotometer (NanoDrop). The final concentration in molarity was 250 µM and the final product quality was assessed via SDS PAGE (Figure 15).

Production and purification of BPTI

The final average OD_{600} at the time of harvested was 23.4 for the 500 mL of BPTI producing cells. After all purification steps and resuspension in 10 mM HCl after lyophilisation the final reduced BPTI yield was 2.57 mg/mL in a volume of 6mL. In molarity the final yield was 337.7 µM and the final product quality was assessed via SDS PAGE (Figure 15). The oxidation state of the final BPTI was tested via MS showing more than 99% of the identified BPTI species corresponding to the fully reduced (0S) form (as can be observed in Figures 17, time-point “0”). The redox heterogeneities visible for BPTI in the SDS PAGE gel in Figure 15 are most likely due to incomplete reduction in the gels loading buffer and the double bands can be associated with the His-tag cleaved BPTI species identified in the MS results. A complete discussion of the BPTI species identification can be found in the following section.

Identifying the BPTI fragments and corresponding folding states

The fully reduced BPTI used in the refolding assays has 6 free cysteines which are free to react with the NEM in the stopping solution. This NEM addition to the BPTI adds 125 MW per added molecule. For each folded disulfide bond on the BPTI two cysteines become unavailable for NEM trapping and no additional weight is added to the protein. However, the final mass of the BPTI is reduced by 2 MW for each DSB present due to the loss of two Hydrogen atoms compared to the reduced state. This difference in weight from the NEM and the loss of hydrogen can be detected in the masses measured in the MS. The possible weights observed in the MS are further expanded and complicated by the possible addition of oxygen onto none, one, two or all of the three methionine (MET) present in the BPTI sequence. Each oxidation adds a molecular weight of 16. Furthermore, the NEM trapping does not always reach 100 % coverage of all available reduced cysteines before the reaction is stopped via flash freezing. Or the opposite can happen where an additional NEM is added to the weight of the BPTI based on an unspecific addition onto the protein [180]. In contrast to the ‘correct’ NEM addition which always adds NEMs in pairs, these two undesired but unavoidable reactions add or miss only a single NEM. These additions of single NEM weights further complicate the mixture of detected masses by the MS analysis. Lastly, cleavage of the His-tag can further generate aberrant peptide. How exactly this cleavage occurs is unclear, however, the detected masses fit the theoretical His-tag-less BPTI masses very precisely.

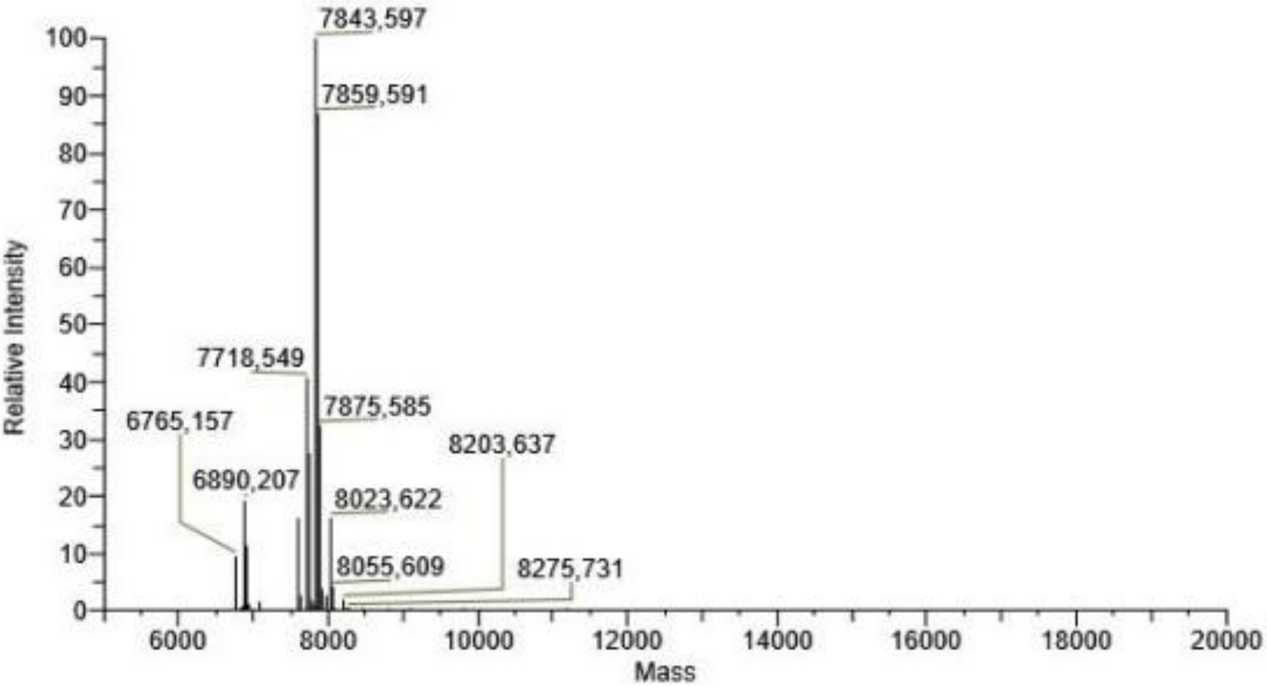


Figure 16 Mass spectrum from BPTI refolding assay. The spectrum shows the common identified masses which are linked to specific folding and oxidation states via Table 3.

The fully reduced version of BPTI used in this research (without any NEM added yet) has a molecular weight of 7591.55 (oxidised) or 7597.55 (fully reduced). The possible fragment weights identified and discussed above are listed in Table 3. Each of the possible combinations between methionine oxidation, added or missed NEM-addition or His-tag cleavage were considered for the analysis but are not listed in Table 3. Any other masses detected but not listed below were not considered for the BPTI oxidative state analysis.

Table 3: Masses of BPTI possible variations occurring in the assay samples. Different masses are linked to their respective oxidation states. The rows list BPTI and the different oxidation states including extra or missed NEM additions as well as the weights of cleaved his-tag BPTI. The columns account for the 3 methionines and their commonly occurring oxidation states. Cleaving the his-tag removes one of the methionines from the sequence which makes the 3 oxidised variation non applicable.

	NEMs added	0 MET oxidised	1 MET oxidised	2 MET oxidised	3 MET oxidised
BPTI 0S:	6	8347.84	8363.84	8379.84	8395.84
BPTI 0S: +1 extra NEM	7	8472.89	8488.89	8504.89	8520.89
BPTI 0S: -1 NEM missed	5	8222.79	8238.79	8254.79	8270.79
BPTI 0S: cleaved His-tag	6	7394.44	7410.44	7426.44	N/A
BPTI 1S:	4	8095.74	8111.74	8127.74	8143.74
BPTI 1S: +1 extra NEM	5	8220.79	7611.55	7627.55	7643.55
BPTI 1S: -1 NEM missed	3	7970.69	7986.69	8002.69	8018.69

BPTI 1S: cleaved His-tag	4	7142.34	7158.34	7174.34	N/A
BPTI 2S:	2	7843.65	7859.65	7875.65	7891.65
BPTI 2S: +1 extra NEM	3	7968.69	7609.55	7625.55	7641.55
BPTI 2S: -1 NEM missed	1	7718.60	7734.60	7750.60	7766.60
BPTI 2S: cleaved His-tag	2	6890.25	6906.25	6922.25	N/A
BPTI 3S:	0	7591.55	7607.55	7623.55	7639.55
BPTI 3S: +1 extra NEM	1	7716.60	7732.60	7748.60	7764.60
BPTI 3S: cleaved His-tag	0	6638.15	6654.15	6670.15	N/A

Enzyme-free oxidation of BPTI by GSSG (Blank)

When no enzyme is added to the BPTI refolding assays the GSSG / GSH mix in the assay buffer is still capable of slowly oxidising BPTI towards the 3S state. However, only the oxidation steps are performed with any relevant efficiency which results in the accumulation of the 2S species [181]. As discussed in the BTPI section of the Introduction the main routes of oxidation for BPTI are a combination of the two steps; initial oxidation of 2 accessible cysteines (mainly CYS14-CYS38) followed by an internal reshuffling of the new DSB onto one of the less accessible cysteine pairs (mainly CYS30-CYS51 or CYS5-CYS55). Without an isomerase present the 2S BPTI stays ‘stuck’ in either of the two 2S forms (CYS14-CYS38 + CYS30-CYS51 or CYS14-CYS38 + CYS5-CYS55). This can be seen in Figure 17 which displays the combined results of 3 runs of this uncatalyzed (blank) refolding assay. The displayed y-axis values of normalized relative abundance are calculated based on all the identified BPTI species listed in Table 3. For all assays discussed in this chapter this accounts for almost all of the detected species in the relevant elution areas of the reverse phase chromatography.

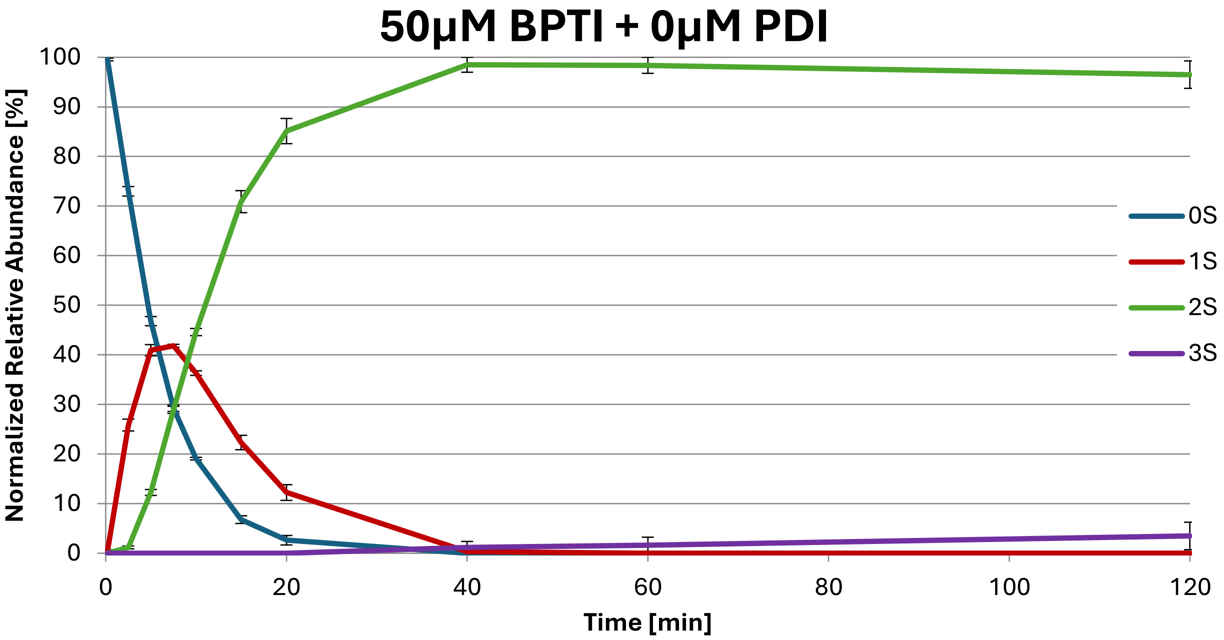


Figure 17: Timeseries of triplicate BPTI refolding assays with 50 µM BPTI in refolding assay buffer and no added enzyme. The oxidation is mediated by the glutathione in the buffer. After 2 hours of total assay time only a negligible amount of BPTI is able to reach the fully folded state.

Figure 17 shows how even in the uncatalyzed assay runs the initial fully reduced BPTI (0S) species quickly disappears, with only a few percent left after 20 minutes. A similar depletion of the 1S species can also be observed with less than 1% left after 40 minutes assay time. At this 40-minute time point the vast majority of BPTI is in the 2S oxidative state. From this point onwards the observed change in BPTI oxidative state is very slow with only a few percent finding the way towards the final 3S species without the help of an isomerase.

BPTI refolding activity of pPDI, hPDI and ERp38

Investigating isomerisation capabilities with 7 µM pPDI

The investigation of the pPDI isomerisation activity was started with a refolding assay concentration of 7µM (the 50 µM of BPTI was kept constant in every assay). The results can be observed in triplicates in Figure 18. We can observe that the initial oxidations of BPTI occurs very fast. In fact, at this concentration they occur too fast to be properly observed. For the following experiment we reduced the concentration of pPDI; however, at this 'high' enzyme concentration we can observe that the fully oxidised BTPI (3S) species reaches higher levels much faster, with it being the dominant species after just 10 minutes. Nevertheless, it still takes the full 2 hours of the assay duration to fully oxidise the BPTI into its native 3S state. The trajectory of the 3S species suggest that the reaction from the 2S to the 3S species isn't a standard exponential enzymatic reaction. In the beginning the increase in 3S species is extremely fast (60% after only 10 minutes) while the later part of this curve is rather slow (110 minutes for the remaining 40%). This is of course because 2S BPTI has two main species which are known to exhibit different isomerisation speeds. This will be further discussed following the comparison with human PDI further down. For now, we can observe that at 7 µM pPDI it takes the full 2 hours to fully fold BPTI and we can use this concentration and these results to calculate the isomerisation activity of pPDI. The initial two oxidation steps to the 1S and 2S states happen within the first 2.5 minutes of the assay and will need to be investigated with a reduced enzyme concentration. This is further highlighted by the first '0' minutes time point starting with only 15 % 0S BPTI left. The time it takes to mix the two reaction mixtures together and pipetting the first 20 µM into the prepared stopping solution (and the time it takes for the stopping solution to stop the reaction completely) is roughly 10 seconds. As a consequence, the '0' time point roughly corresponds to 10 seconds after the start of the reaction (on average). Furthermore, with an isomerase concentration of 7 µM these 10 seconds are enough to oxidise 85% of the 0S BPTI.

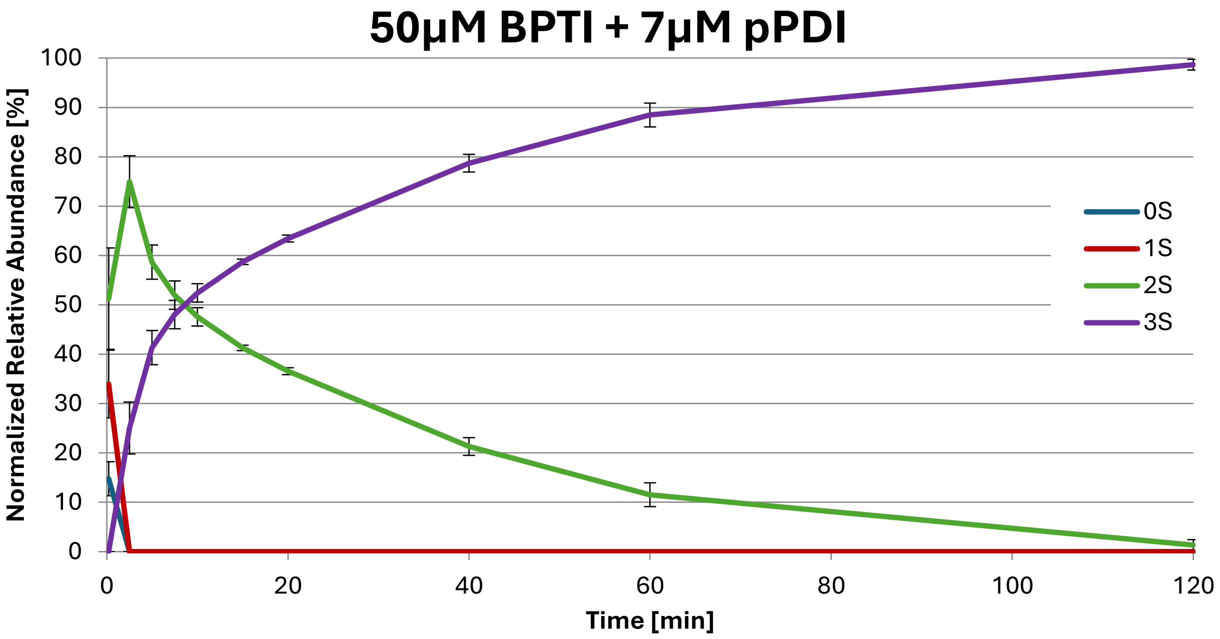


Figure 18: Timeseries of triplicate BPTI refolding assays with 50 µM BPTI in refolding assay buffer and 7 µM of pPDI. While the oxidation steps from 0S to 2S are completed within the first 2.5 minutes of the assay, the isomerisation towards the 3S state can be observed clearly.

Investigating oxidation steps with 0.1 µM pPDI

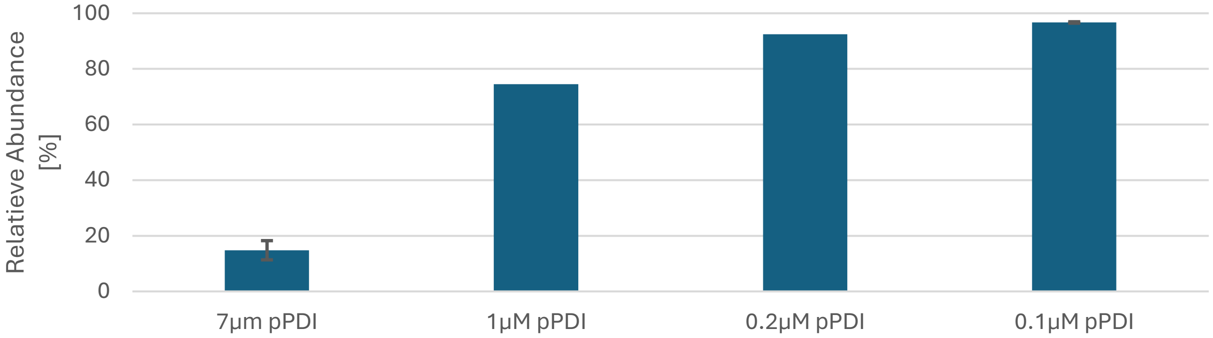


Figure 19 Impact of pPDI concentration on 0S BPTI concentration at time point 0 (i.e. after roughly 10 seconds). Measurements at 1 µM and 0.2 µM are based on single repeats only.

After testing various pPDI concentrations (Figure 19) the 0.1 µM concentration was selected as the most suitable for observing the first two oxidation steps of BPTI. The triplicate results are displayed in Figure 20. At this reduced isomerase concentration, the initial oxidation from 0S to 1S and 2S can now be observed. As opposed to the above-described concentrations, the initial 0S BPTI state at time point '0' is still observable at roughly 97%. Also, the transient 1S species is clearly observable between 0 and 15 minutes. After 15 minutes the first two oxidation steps are mostly completed and only the slow isomerisation reaction towards the 3S state remains to be completed. However, at this strongly reduced pPDI concentration, this final step takes too long to be completed within the 2-hour assay duration with only roughly 17% of the BPTI reaching the final oxidation state. Nevertheless, based on these results the kinetics of the first two steps could be calculated and will be discussed further down.

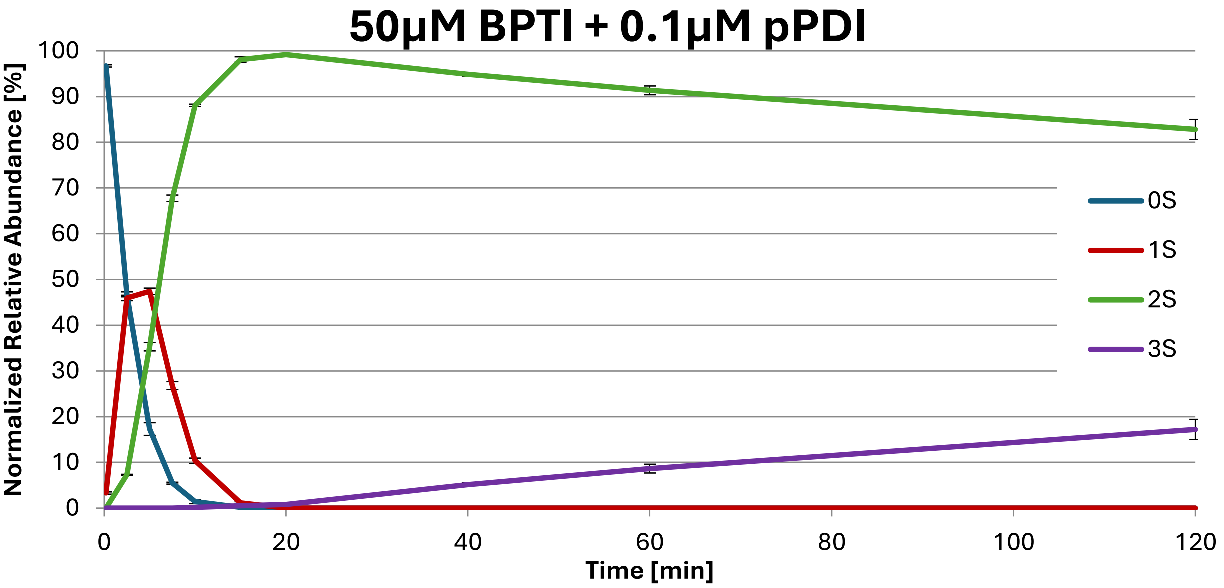


Figure 20: Timeseries of triplicate BPTI refolding assays with 50 µM BPTI in refolding assay buffer and 0.1 µM of pPDI. With this lower enzyme concentration, the oxidation steps from 0S to 1S and 2S are clearly visible. The isomerisation step, however, stays incomplete at the end of the two hours assay duration.

Investigating isomerisation capabilities with 7 µM hPDI

Human PDI has been extensively studied for decades ever since it was discovered as the primary enzyme linked to DSB formation in humans. As such the kinetic values of human PDI together with BPTI have been studied before [182]. However, in order to better compare the *P. pastoris* PDI results to its human equivalent, BPTI refolding was also investigated at the same concentrations for hPDI. The high 7 µM hPDI concentration results are displayed in Figure 21. When comparing this figure to the pPDI equivalent results in Figure 18, we can observe very similar trends. In similar fashion to pPDI, the first two oxidation steps are completed within the first 2.5 minutes of the assay and the remainder of the 2-hour assay time is required for the completion of the final isomerisation and oxidation step (2S to 3S). The two PDI versions will be more thoroughly compared later on.

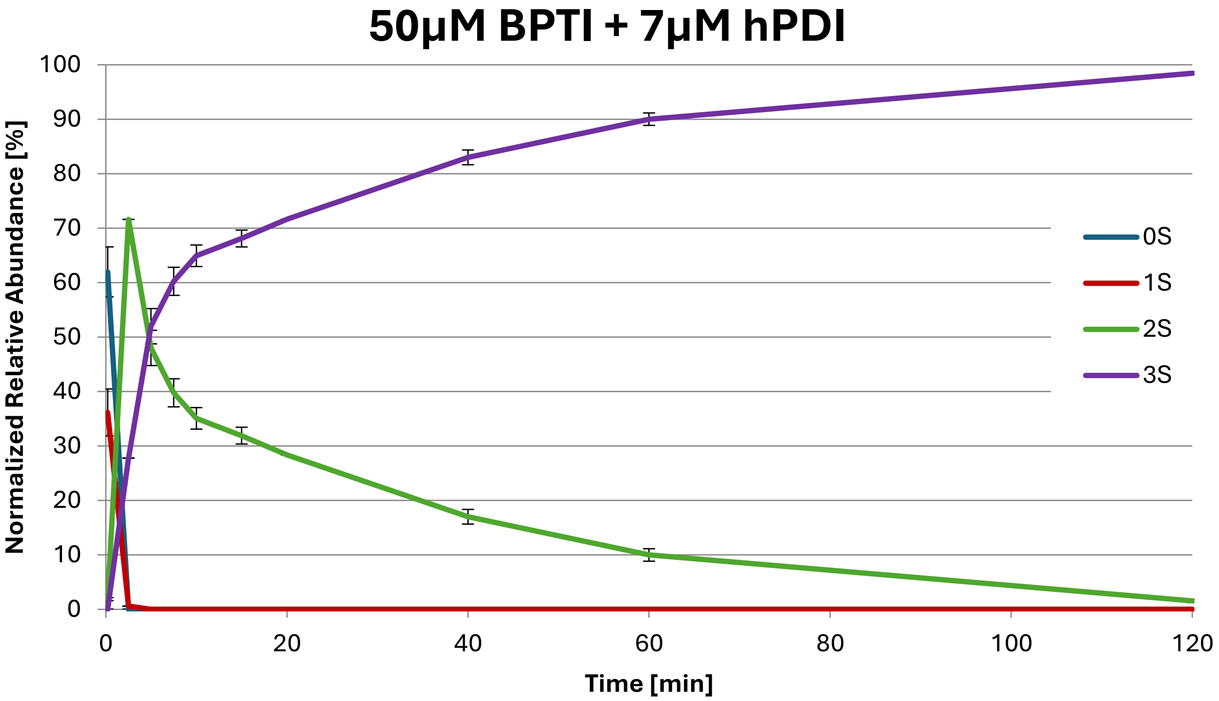


Figure 21: Timeseries of triplicate BPTI refolding assays with 50 µM BPTI in refolding assay buffer and 7 µM of hPDI. While the oxidation steps from 0S to 2S are completed within the first 2.5 minutes of the assay, the isomerisation towards the 3S state can be observed clearly.

Investigating oxidation steps with 0.1 µM hPDI

In order to properly observe the first two oxidation steps the reaction had to be slowed down which was achieved, like with pPDI before, by reducing the concentration of the isomerase down to 0.1 µM. The results of this heavily slowed down reaction can be seen in Figure 22. As observed with the high enzyme concentrations, the profile of the oxidative folding states of BPTI looks very similar to the previously tested pPDI. With the reduced PDI concentration the two first oxidation steps can now be properly distinguished from each other, and the corresponding kinetic parameters can be calculated. This calculation will be done and discussed further down together with a thorough comparison to the pPDI results.

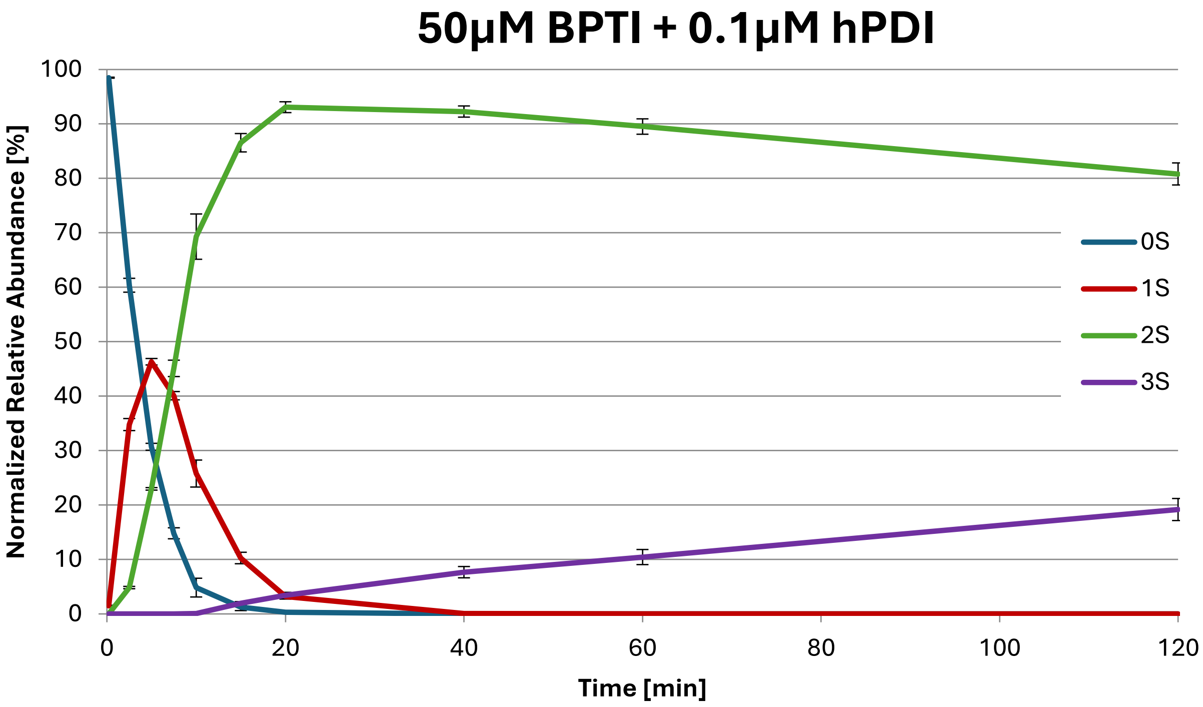


Figure 22: Timeseries of triplicate BPTI refolding assays with 50 µM BPTI in refolding assay buffer and 0.1 µM of pPDI. With this lower enzyme concentration, the oxidation steps from 0S to 1S and 2S are clearly visible. The isomerisation step, however, stays incomplete at the end of the two hours assay duration.

Investigating isomerisation capabilities with 7 µM ERp38

The results for 7 µM ERp38 are displayed in Figure 23.

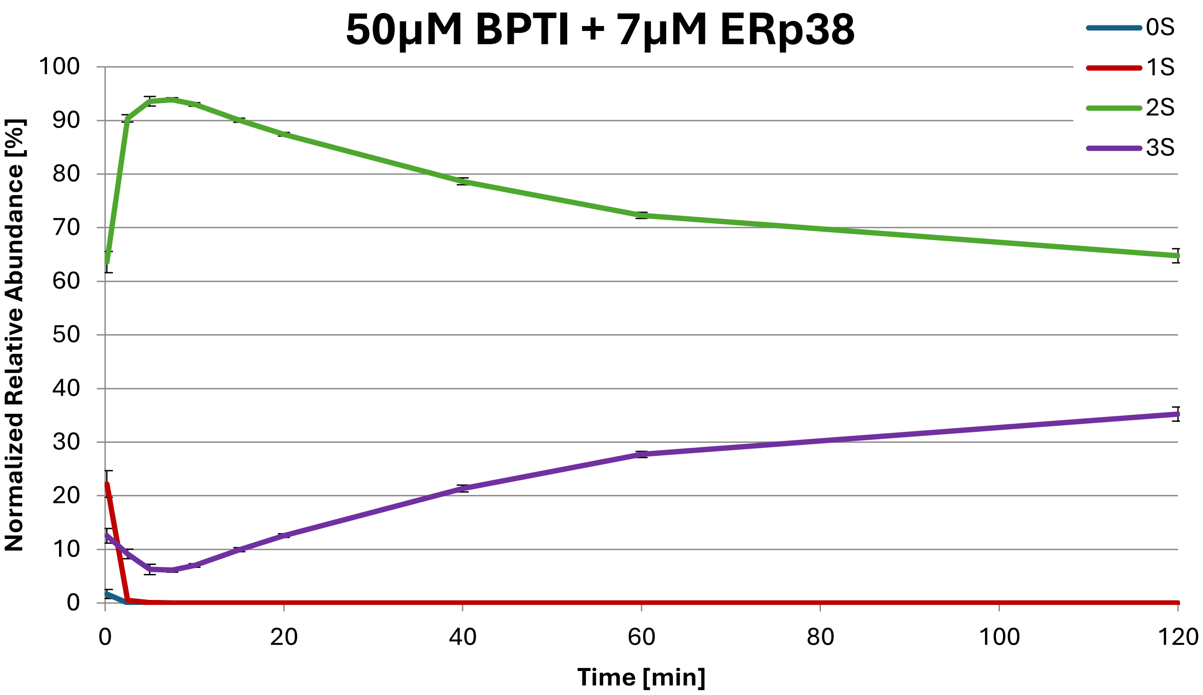


Figure 23: Timeseries of triplicate BPTI refolding assays with 50 µM BPTI in refolding assay buffer and 7 µM of ERp38. Despite the high concentration of added enzyme, the isomerisation of BPTI is not completed after the 2-hour timepoint. However, the two oxidation steps are completed within the first 2.5 minutes.

In similar fashion to the two previously characterised enzymes pPDI and hPDI, the initial oxidations of BPTI to the 2S state are completed within the first 2.5 minutes of the assay and will have to be investigated at a reduced enzyme concentration. However, unlike the PDIs, ERp38 is not able to complete the final isomerisation step towards the 3S state in the assays' timespan of 2 hours. Only roughly 35% of the BPTI has reached its final oxidation state at the end of the assay. We can assume that the isomerisation activity of this enzyme is reduced compared to the PDIs, although still present, since the final 3S concentration is still well above the blank run with only the glutathione present (Figure 17). The resulting kinetics will be discussed further below.

Investigating oxidation steps with 0.04 µM ERp38

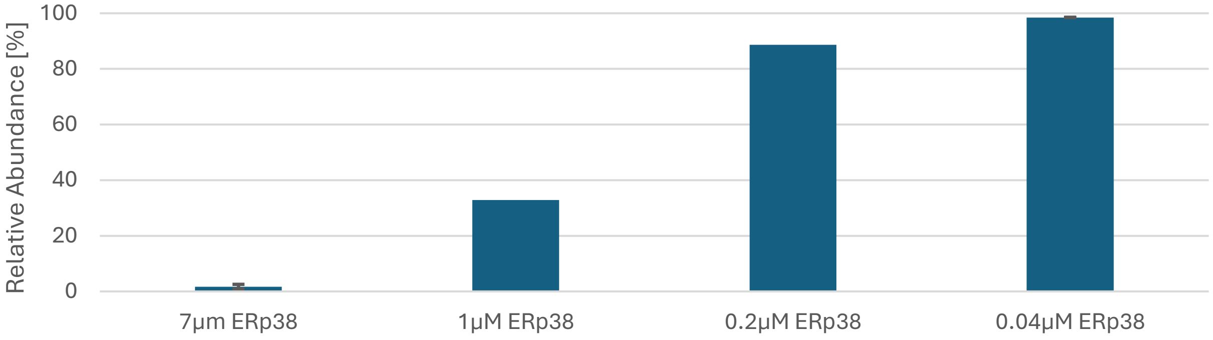


Figure 24 Impact of ERp38 concentration on 0S BPTI concentration at time point 0 (i.e. after roughly 10 seconds). Measurements at 1 µM and 0.2 µM are based on single repeats only.

The assay concentrations of ERp38 were gradually reduced until the reaction has slowed down enough in order to observe the initial oxidation steps of the BPTI refolding. The results of this process can be observed in Figure 24. For the PDIs this was achieved at a concentration of 0.1 µM, however, for ERp38 this concentration was still too high and only after reducing the concentration down to 0.04 µM were the initial oxidation stages properly observable. The results of the assay with this concentration are displayed in Figure 25. At this heavily reduced enzyme concentration the changes in the BPTI folding states look similar to the ones observed for the PDIs. For a concentration independent comparison, the kinetic value will need to be calculated which will be done below. For now, we can conclude that ERp38 seems to be a significantly stronger oxidase but a weaker isomerase compared to the two PDIs.

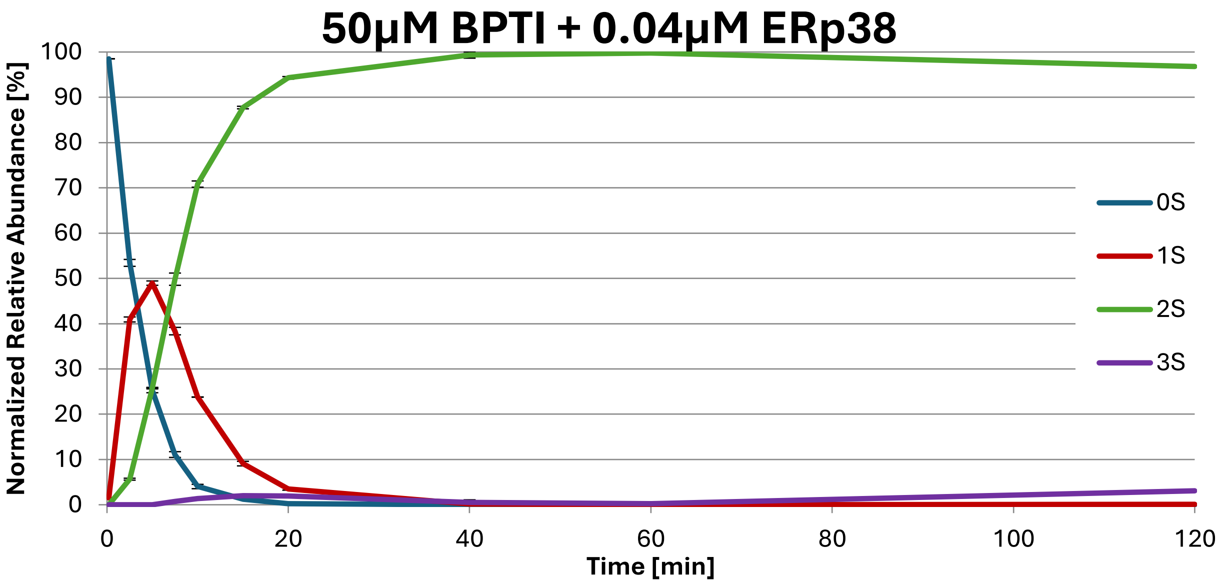


Figure 25: Timeseries of triplicate BPTI refolding assays with 50 µM BPTI in refolding assay buffer and 0.04 µM of ERp38. At this heavily reduced enzyme concentration the two initial oxidation steps become visible in the beginning on the assay.

7 µM pPDI with less glutathione

The glutathione in the assay buffer functions as the terminal electron acceptor for the oxidation of BPTI. More specifically it is the oxidised glutathione (GSSG) in the buffer that accepts the electrons and in doing so transfers its own disulfide bond onto BPTI. GSSG forms a redox pair with its reduced form, GSH. This reduced glutathione is also added to the buffer and can function as an electron donor and reduce oxidised PDI, which needs to be in this form to function as an isomerase. With glutathione being central to the function of the isomerase, it is important to also investigate the effect it has on its ability to refold BPTI. For this purpose, the glutathione concentrations were reduced to observe their impact on the 7 µM pPDI refolding results. In order to only observe the concentration effects of a changed glutathione concentration and not the effects of a change in redox environment, the two glutathione species were changed according to Equation 4.

$$\frac{[GSH]^2}{[GSSG]} = constant$$

(Eq. 4)

With the target value of 0.15 mM GSSG (compared to the 0.5 mM used in all other experiments) the corresponding concentration of GSH was 1.095 mM (down from 2 mM). The result of the assay runs with the reduced glutathione concentration are displayed in triplicates in Figure 26. In contrast to the results displayed in Figure 18, the reduced concentration of glutathione slowed down the speed of the initial BPTI oxidation occurring within the first few seconds of the assay. At time point ‘0’, roughly 59% of the initial 0S BPTI is remaining. Also, only ~7% of the BPTI has reached the 2S state at this time point compared to the ~51% in Figure 18. However, an impact of the reduced glutathione concentrations on the final isomerisation step from 2S to 3S could not be observed, with the completion of the BPTI refolding still occurring after roughly 2 hours.

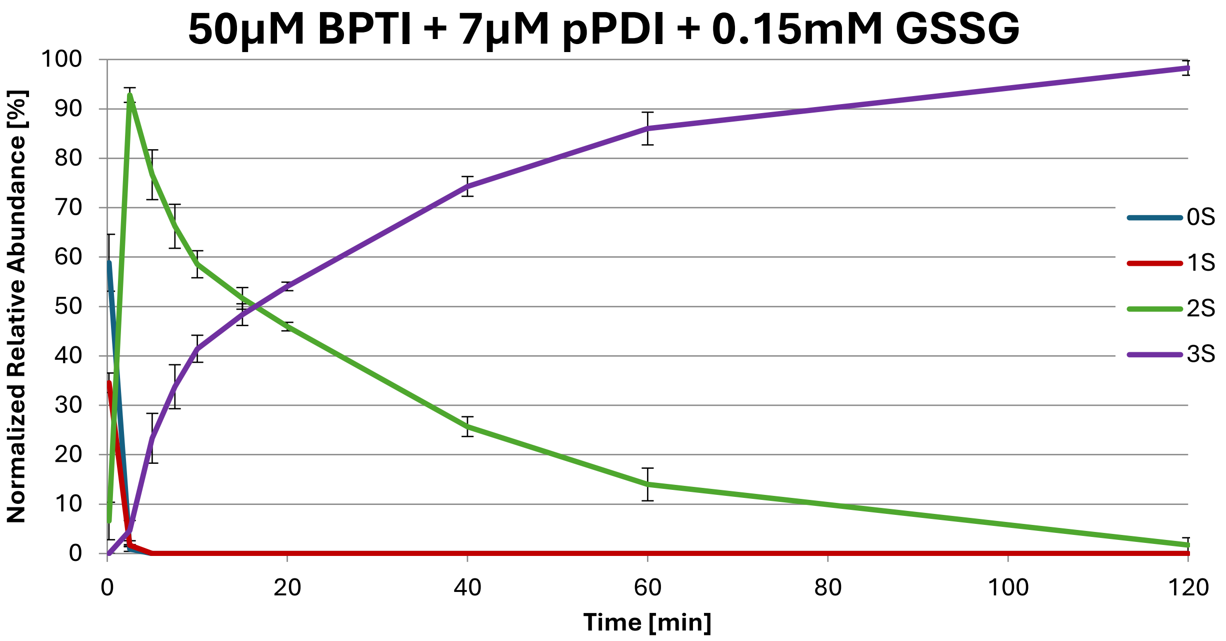


Figure 26: Timeseries of triplicate BPTI refolding assays with 50 µM BPTI with 7µM pPDI and a slightly changed assay buffer condition. The glutathione concentration in the buffer was lowered to 0.15 mM GSSG and 1.095 GSH (from 0.5 and 2 mM respectively).

The kinetics of oxidation and isomerisation of pPDI, hPDI and ERp38

Fully reduced BPTI has 6 available cysteines for disulfide bonding. However, during folding, not all possible cysteines combinations form a relevant amount of disulfide bonds between them with only the 3 native cysteine pairs being quantitatively relevant (i.e. cysteine pairs 5-55, 14-38 and 30-51). As described in the BPTI introduction, the major pathway for BPTI oxidative folding is via the oxidation of the 14-38 disulfide bond, followed by a subsequent reshuffling of this new disulfide bond on to one of the other two native disulfide bond positions. The reduced 14-38 cysteines are then oxidised again and after another reshuffling, they are oxidised a third and final time resulting in natively folded BPTI. The second reshuffling step is the slowest step in the folding process. This is the step that is heavily reliant on an isomerase which can be observed in the assays with the reduced enzyme concentrations (i.e. Figure 20) or no added enzyme at all (Figure 17). Here the final 3S fraction never exceeds a couple of percent.

Based on the timeseries data it is already possible to estimate that ERp38 is a stronger oxidase compared to the two PDIs. However, this comes at the cost of reduced isomerisation capabilities. pPDI and hPDI on the other hand, show very similar behaviour and will require the following kinetic analysis to determine how they might differ from each other.

First Oxidation step (0S to 1S)

The first oxidation step from the fully reduced BPTI to the BPTI with a single disulfide bond follows an exponential decay function as displayed in Equation 1. The kinetic parameter k_1 can be calculated by fitting this function to the refolding assay data, particularly the assays with the lower enzyme concentrations (0.1 µM pPDI and hPDI as well as 0.04 µM ERp38).

$$[BPTI_{1S}] = [BPTI_{0S}] * \ e^{-k_1 * t}$$

(Eq. 1)

The fitting results for this first BPTI oxidation step are displayed in Figure 27 together with the calculated kinetic parameters k_1.

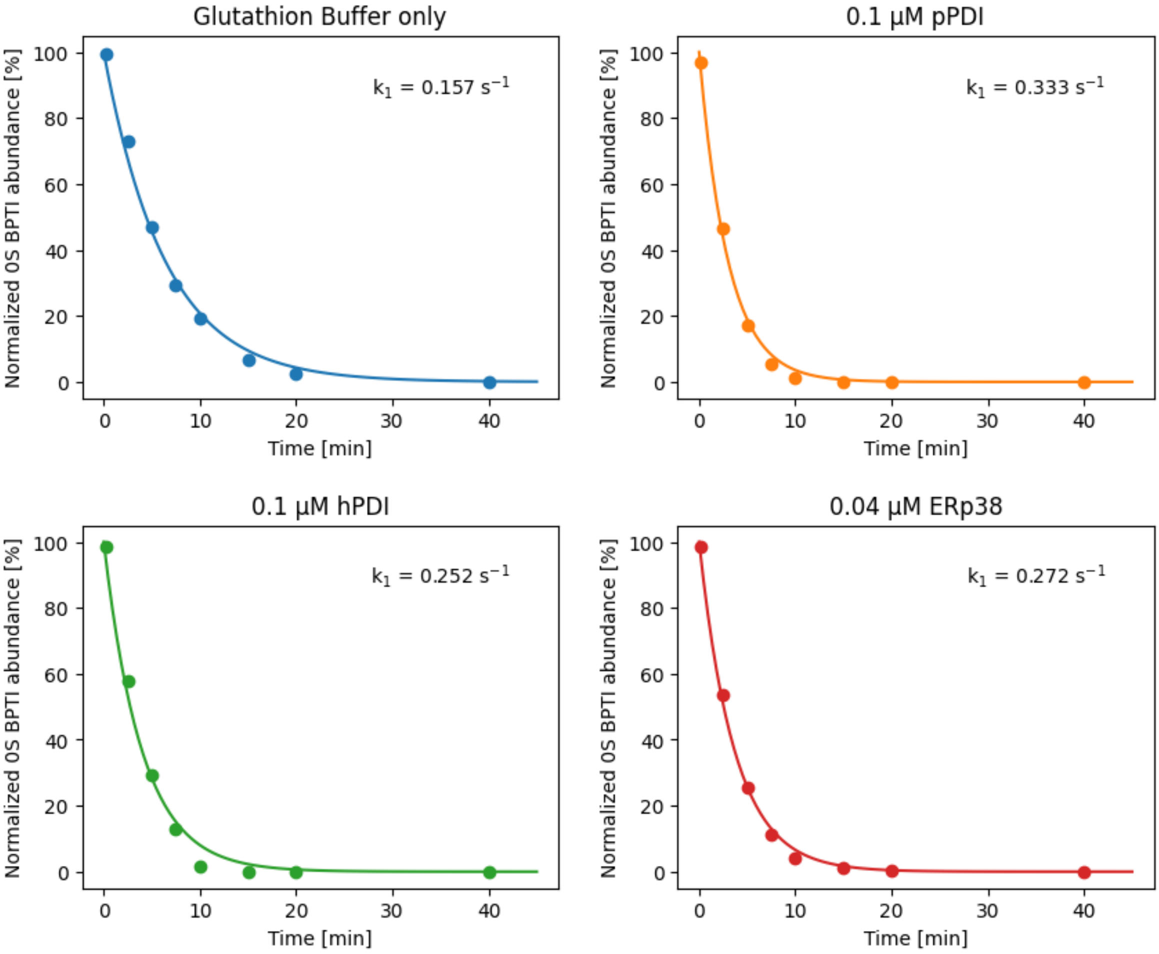


Figure 27 Curve fitting results of the first BPTI oxidation step using Equation 4. Kinetic parameters obtained from the fitting are displayed in the top-right corner of each graph.

The higher the k_1 values, the faster the exponential decay happens, which corresponds to a sharper decline in the 0S species. We can observe that the uncatalyzed reaction (Figure 27, top-left) has the lowest k_1 value with 0.157 s^{-1}. All three enzymes tested show a higher k_1 value which translates to all three enzymes having oxidase activity. In order to compare the kinetic values between different enzyme concentrations, the rate constant can be calculated. This is done via Equation 5. Before however, the rate constant for the glutathione oxidation still has to be subtracted from the rate constants of the enzymatically catalysed reactions since glutathione is also present and active during their assays.

$$k = \frac{k_1 - k_{1\,[GSSG]}}{[Enzyme]}$$

(Eq. 5)

With this we can now calculate the rate constant of the reactions and compare them to each other. For 0.1 µM pPDI, 0.1 µM hPDI and 0.04 µM ERp38 we get 1.76, 0.95 and 2.86 s^{-1} μM^{-1} respectively. With these values we can conclude that in regard to the initial oxidation step of BPTI, ERp38 is more than 1.5 times as fast as pPDIs and more than twice as fast as hPDI. Between the two PDIs it seems that pPDI is roughly 54% faster.

Second Oxidation step (1S to 2S)

The second oxidation step in the BPTI folding pathway is a combination between two steps: a reshuffling of the first DSB onto one of the two other DSB positions (5-55 or 31-51) followed by another oxidation of the now free 14-38 DSB. Just like in oxidation step 1, this isn't the only oxidation pathway that BPTI can take to get to the 2S state, but it is the dominant pathway.

Since this second oxidation is dependent on the product of the first oxidation (the 1S BPTI), it follows a more complex trajectory. This is because the starting species for this reaction (1S) is being created at the same time as this second oxidation step is occurring. The result of this can be seen in Figure 17 to Figure 26, with the 1S species appearing and disappearing within the duration of the assays. This second oxidation step follows a trajectory which can be described by Equation 2. The fitting results in conjunction with the two calculated kinetic parameters are displayed in Figure 28.

$$[BPTI_{2S}] = \frac{[BPTI_{0S}] * k_1}{k_2 - k_1} * \left(e^{-k_1 * t} - e^{-k_2 * t}\right)$$

(Eq. 2)

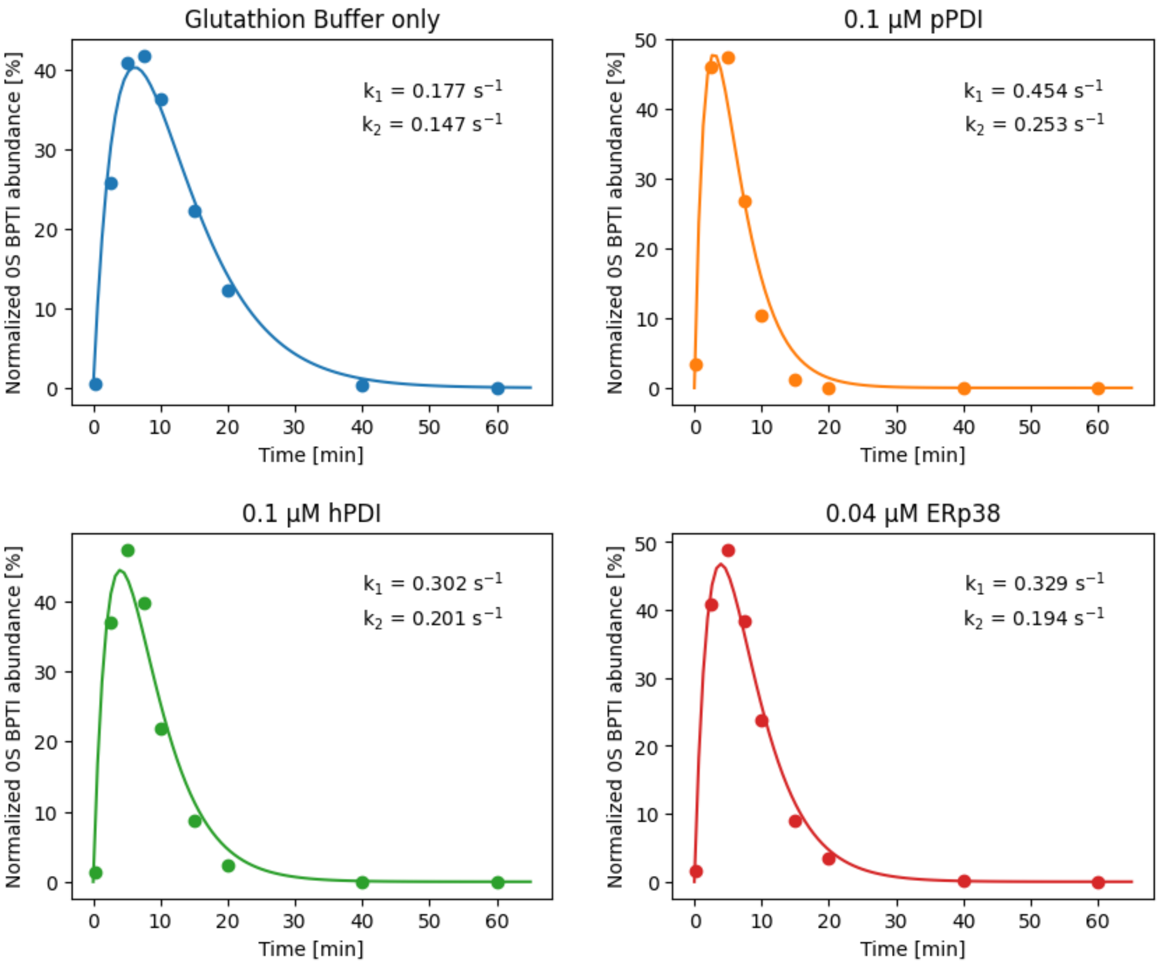


Figure 28 Curve fitting results of the second BPTI oxidation step using Equation 5. Kinetic parameters obtained from the fitting are displayed in the top-right corner of each graph.

The kinetic fitting results show a similar relation between the enzyme compared to the first oxidation step. If we use Equation 5 again to calculate the rate constants based on the 3 estimated k_1 values and the 3 estimated k_2 values we receive 2.77 & 1.06, 1.25 & 0.54 as well as 3.80 & 1.18 s^{-1} μM^{-1} for the three enzymes respectively. ERp38 is still the fastest of the three enzymes but the difference to pPDI is less distinct. The difference between the two PDIs is more pronounces with pPDI being almost twice as fast as hPDI. Also, we can observe that the second oxidation step is slower compared to the first step.

Third Oxidation step (2S to 3S)

The third oxidation step in the BPTI folding pathway happens in a similar way to the previous step from 1S to 2S. The newly introduced second DSB must be shifted onto the remaining unoxidized DSB position (either 5-55 or 31-51, which ever isn't oxidised yet). This is the slowest step in the BPTI folding pathway and relies heavily on the help from an isomerase.

Based on observations made in the plotted timeseries we know that all three enzymes have a higher isomerase activity than glutathione alone, since the final 3S concentration is significantly higher compared to the glutathione buffer only assays. However, ERp38 seems to show less isomerase activity compared to the two PDIs. This is not too surprising, after the above observed oxidative strength of ERp38. The ability to isomerase a DSB is linked to an enzymes ability to form a temporary mixed disulfide with a substrate [94]. In most cases this requires the isomerase to reduce the misfolded DSB temporarily and this step is energetically more difficult for a strong oxidase to perform efficiently. The isomerisation step is further complicated by the efficiency of the enzyme-substrate interactions, the reoxidation of the enzyme by glutathione or other electron acceptors, and other oxidation control mechanisms present. Nevertheless, while the two reactions - oxidation and isomerisation – share functionality, there is a trade-off between the two reaction which is why even comparatively simple systems such as *E. coli* have different enzymes for the two functions.

As mentioned in the Introduction, the refolding of BPTI has been extensively studied and the different states of the oxidative refolding states have been described. As can be seen in Figure 14, BPTI has 3 different 2S states, commonly labelled as N' (14-38 & 30-51), N* (5-55 & 14-38) and N_{SH}^{SH} (5-55 & 30-51). The first two states are considered 'energetically trapped' and require isomerisation to reach the N_{SH}^{SH} state. From this last state, which has the last remaining reduced cysteines 14 and 38 surface exposed, the final native 3S state can be reached very quickly. Since this last step is so fast and the previous two steps are so slow, the here observable reaction kinetics is almost exclusively dominated by the isomerisation step which is why this step is commonly referred to as 'Isomerisation' even though it also includes the subsequent oxidation of the final disulfide bond.

In the three high enzyme concentration assays (7 µM) the initial two oxidation steps are completed within the first couple of minutes and the isomerisation step can begin with roughly the full 100 % of 2S species. On a first glance, the trajectory of the 3S formation seems to follow an exponential trajectory (Equation 6) and the results of such a fitting can be seen in Figure 29. However, no proper fitting can be achieved with this single exponential function. As mentioned in the previous section, the 2S state has two 'energetically trapped' states N' and N* which have different speeds at which they can be typically isomerised. As such we can use the double exponential function displayed in Equation 3 to calculate the kinetic parameters of the isomerisation step. The results of this curve fitting can be seen in Figure 30.

$$[BPTI_{3S}] = [BPTI_{0S}] - \left([2S] * e^{-k_1 * t}\right)$$

(Eq. 6)

$$[BPTI_{3S}] = [BPTI_{0S}] - \left([N^*] * e^{-k_1 * t} + [N'] * e^{-k_2 * t}\right)$$

$$with\ [N'] + [N^*] = [BPTI_{0S}]$$

(Eq. 3)

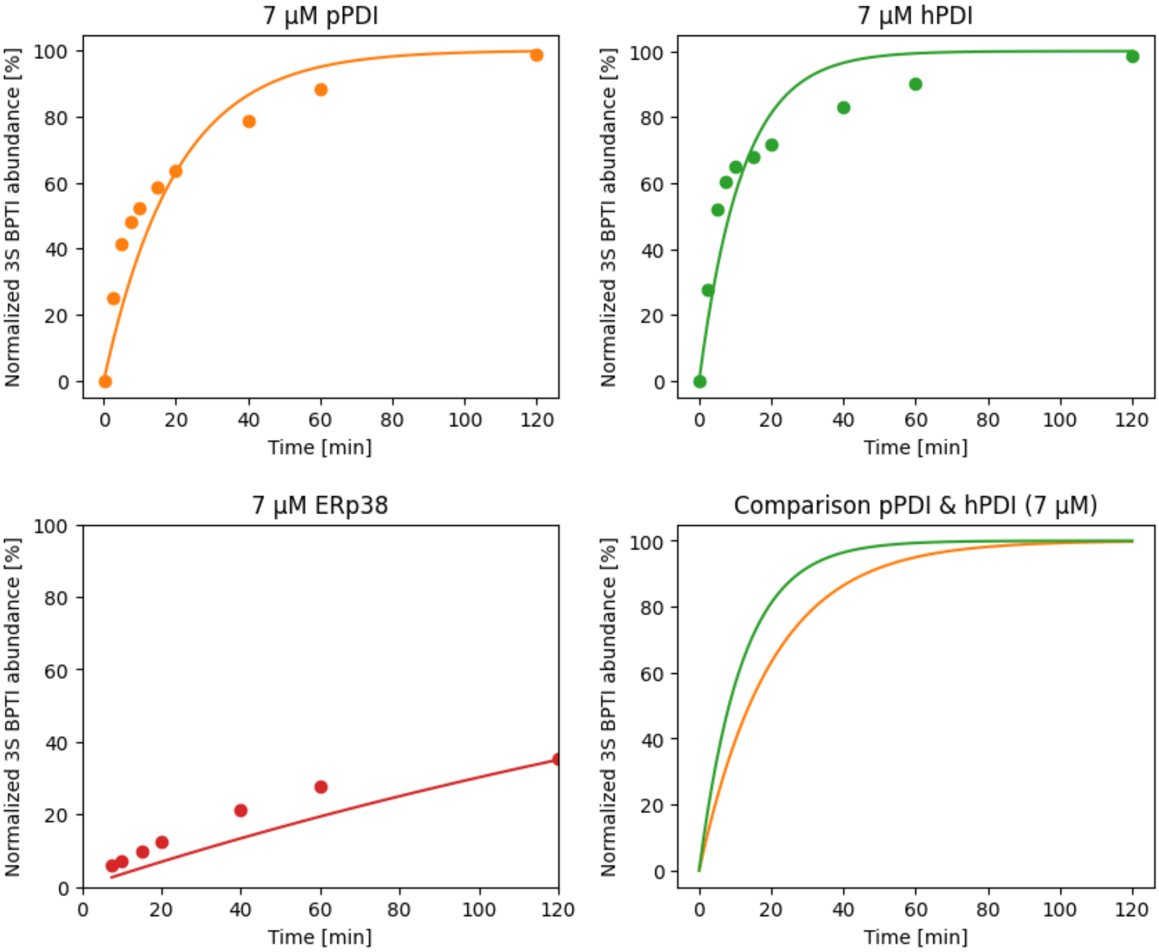


Figure 29: Curve fitting results of the BPTI isomerisation step using the single exponential equation (Equation 6) instead of the double exponential equation (Equation 3).

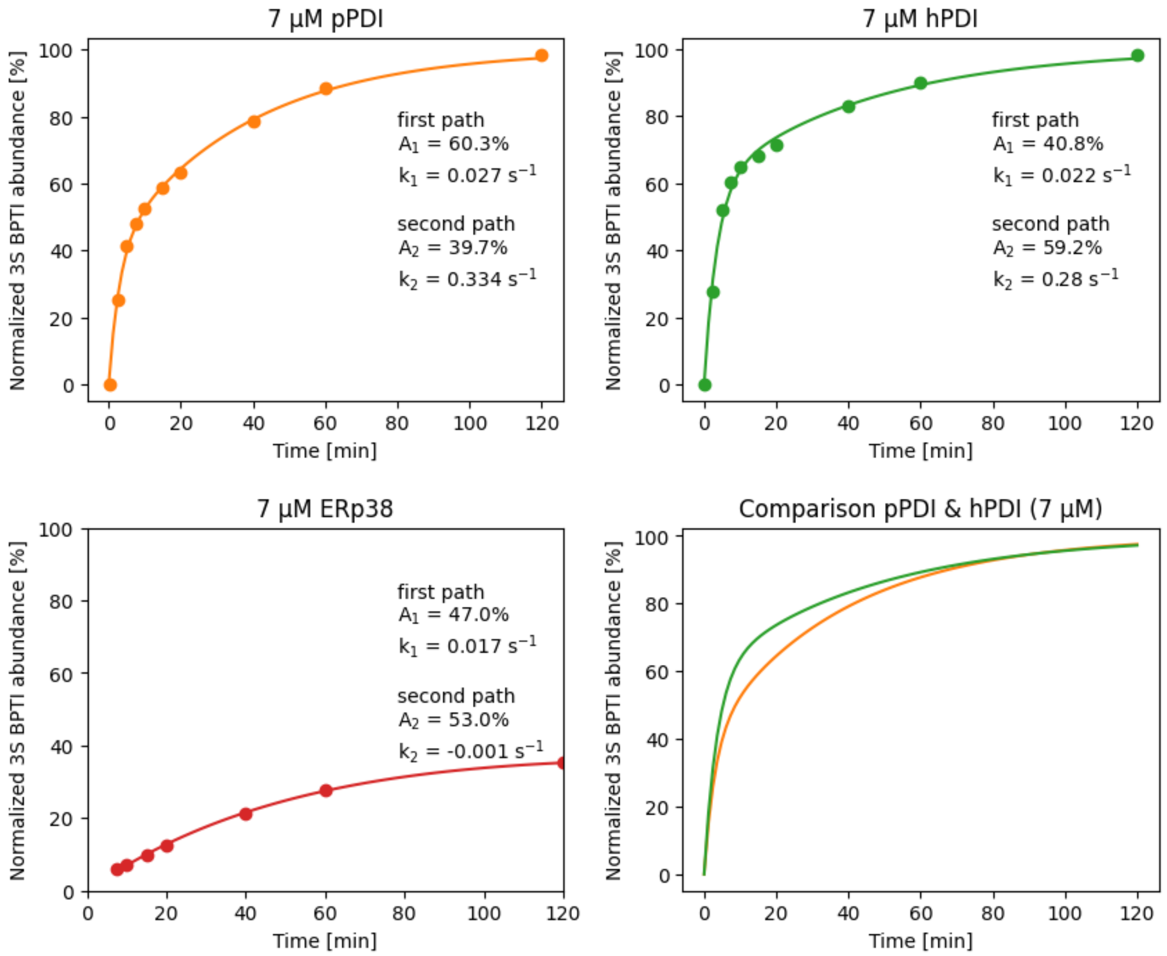
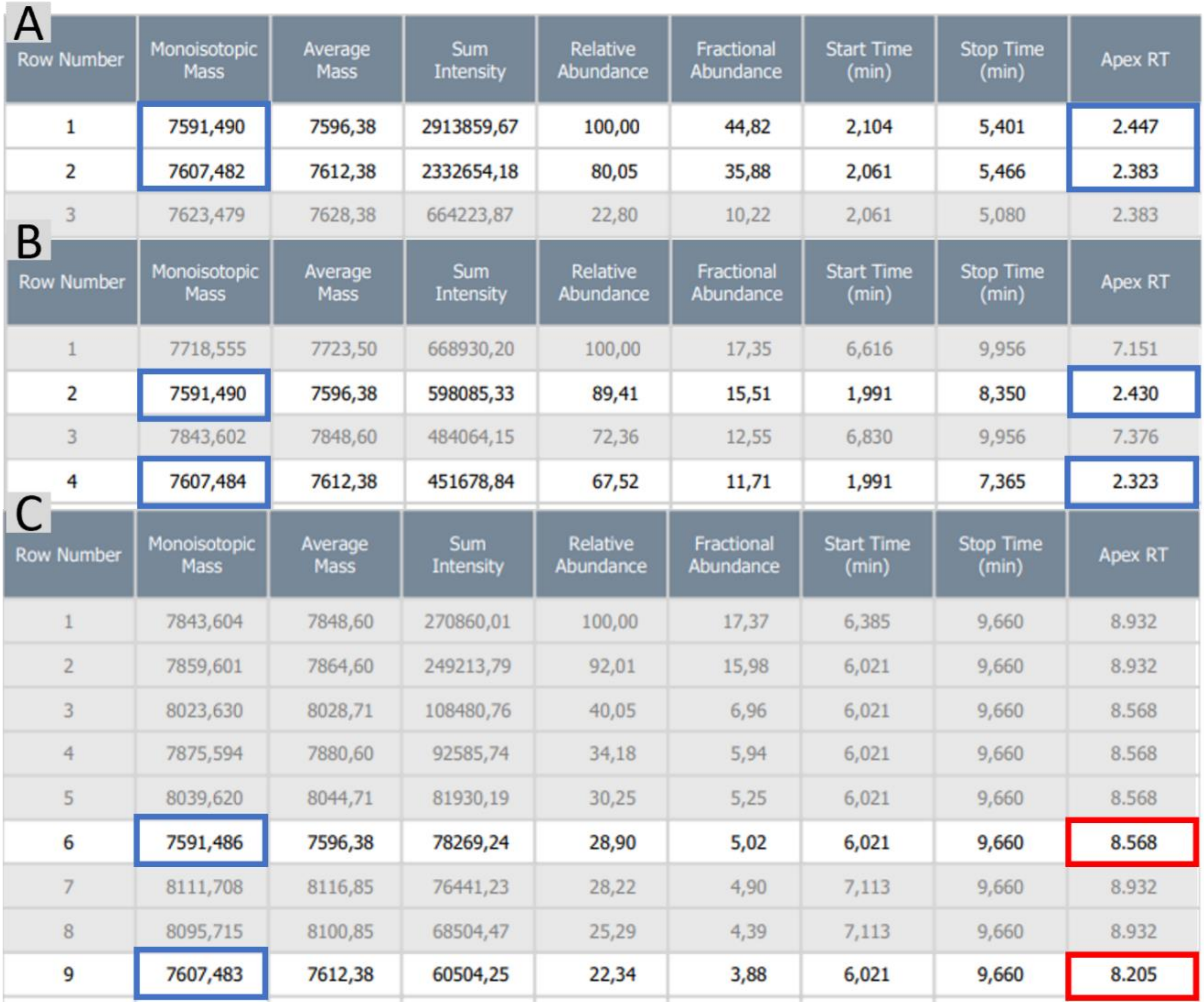


Figure 30 Curve fitting results of the BPTI isomerisation step using Equation 3. Kinetic parameters and fractions for each of the two isomerisation pathways are displayed in the top-right corner of each graph. The last panel compares the two fitted curves for pPDI and hPDI.

Via equation 5 we can now calculate the rate constants as 0.0039 & 0.0477 for pPDI, 0.0032 & 0.040 for hPDI as well as 0.0024 for ERp38 (all values in s^{-1} μM^{-1}). With the double exponential function the fitting works very well. With this function, two kinetic parameters k_1 and k_2 are estimated as well as the amount of substrate that passes through each (A_1 and A_2 respectively). While it is not possible to determine which pathway and kinetic parameters is linked to which of the two substrate species (N' and N*) it is known that the pathway involving N' is faster compared to the one involving N*. Since all three assays use the same enzyme concentrations, we do not need to calculate the rates to compare them to each other. For ERp38 we can observe that k_2 is close to 0 which would suggest that the enzyme cannot isomerise one of the two 2S species at all. The kinetic rate k_1 is slower compared to the two PDIs but is nonetheless able to facilitate the complete folding of BPTI. The comparison between the two PDIs is done more thoroughly below.

Non-native 3S species

The two ERp38 refolding assays (Figure 23 and Figure 25) display an unusual behaviour for the 3S species trajectories. In both assays and in each of the 3 repeats, an initial appearance of a BPTI species with 3 disulfide bonds can be observed, which disappears during the course of the assay only to reappear later on. Since this behaviour has not been observed for BPTI before, the corresponding MS results are investigated below in more detail.



A

Row Number	Monoisotopic Mass	Average Mass	Sum Intensity	Relative Abundance	Fractional Abundance	Start Time (min)	Stop Time (min)	Apex RT
1	7591,490	7596,38	2913859,67	100,00	44,82	2,104	5,401	2.447
2	7607,482	7612,38	2332654,18	80,05	35,88	2,061	5,466	2.383
3	7623,479	7628,38	664223,87	22,80	10,22	2,061	5,080	2.383

B

Row Number	Monoisotopic Mass	Average Mass	Sum Intensity	Relative Abundance	Fractional Abundance	Start Time (min)	Stop Time (min)	Apex RT
1	7718,555	7723,50	668930,20	100,00	17,35	6,616	9,956	7.151
2	7591,490	7596,38	598085,33	89,41	15,51	1,991	8,350	2.430
3	7843,602	7848,60	484064,15	72,36	12,55	6,830	9,956	7.376
4	7607,484	7612,38	451678,84	67,52	11,71	1,991	7,365	2.323

C

Row Number	Monoisotopic Mass	Average Mass	Sum Intensity	Relative Abundance	Fractional Abundance	Start Time (min)	Stop Time (min)	Apex RT
1	7843,604	7848,60	270860,01	100,00	17,37	6,385	9,660	8.932
2	7859,601	7864,60	249213,79	92,01	15,98	6,021	9,660	8.932
3	8023,630	8028,71	108480,76	40,05	6,96	6,021	9,660	8.568
4	7875,594	7880,60	92585,74	34,18	5,94	6,021	9,660	8.568
5	8039,620	8044,71	81930,19	30,25	5,25	6,021	9,660	8.568
6	7591,486	7596,38	78269,24	28,90	5,02	6,021	9,660	8.568
7	8111,708	8116,85	76441,23	28,22	4,90	7,113	9,660	8.932
8	8095,715	8100,85	68504,47	25,29	4,39	7,113	9,660	8.932
9	7607,483	7612,38	60504,25	22,34	3,88	6,021	9,660	8.205

Figure 31 Most abundant protein species detected from the BPTI refolding assays via LC-MS. The species are displayed with their monoisotopic masses detected by the MS together with their respective corresponding retention times in the RP chromatography. The apex time point for each species elution pattern is display on the far-right side. A: Top 3 most abundant species for assay time point 120 minutes with 7µM pPDI enzyme concentration. B: Top 4 most abundant species for assay time point 120 minutes with 7µM ERp38 enzyme concentration. C: Top 9 most abundant species for assay time point 0 with 7µM ERp38 enzyme concentration.

The top part of Figure 31 shows the MS results for a single measurement corresponding to the last time point in Figure 18 which predominantly contains BPTI fully oxidised and isomerised by 7µM pPDI. In the same way, Part 'B' shows the last assay time-point and part 'C' shows the first time point (Figure 23) for BTPI folded with 7µM ERp38. In all three parts of Figure 31 only the most common and relevant masses are listed. In all three cases the detected masses for the two most common 3S BPTI species (standard 7591.5 g/mol and single oxidised with 7607.5 g/mol) are within 0.1 g/mol between the three separate mass measurements. The observed retention times of the 3S species are however quite different between the ‘early’ ERp38 based 3S BPTI and the ‘standard’ 3S BPTI. The observed apex retention times for 3S BPTI was detected at roughly 2.3 – 2.5 minutes for all measurements based on assays with hPDI, pPDI and late ERp38 as well as non-refolded oxidised BPTI (collected after IMAC purification and before DTT reduction). For comparison, the retention time for 2S BPTI is always detected at roughly 7-8 minutes, i.e. much later compared to 3S BTPI. The retention of the partially unfolded 2S (and 1S or 0S) state is higher due a larger amount of hydrophobic amino acid residues present on the outside of the protein. This results in a stronger interaction of the 2S BPTI with the hydrophobic stationary phase of the RP column compared to the fully folded 3S BPTI. The shift in retention time and the accurately matching masses therefore suggest that the early 3S BPTI species folded via ERp38 corresponds to a fully oxidised but misfolded BPTI. To my knowledge such an enzymatically catalysed non-native 3S BPTI species has not been observed before.

Figure 32 and Figure 33 also display the chromatograms corresponding to part ‘A’ and ‘C’ of Figure 31. The first one shows the retention times observed for ‘standard’ 3S BPTI which forms a distinct ‘peak’ early in the chromatogram. The presumed non-native 3S BPTI species; however, elutes much later and overlaps with the elution times of 0S, 1S and 2S BPTI. Furthermore, the chromatogram in Figure 31 and the measured retention times suggest that this presumed non-native 3S BPTI might not correspond to a uniquely misfolded 3S BPTI which could explain why no distinct ‘peak’ can be observed. However, the concentrations of this non-native 3S species is relatively low with 15% relative abundance. Attempts to separate the non-native 3S species from the other BTPI species were made but where unfortunately beyond the scope of my research stay in Oulu University and my PhD project.

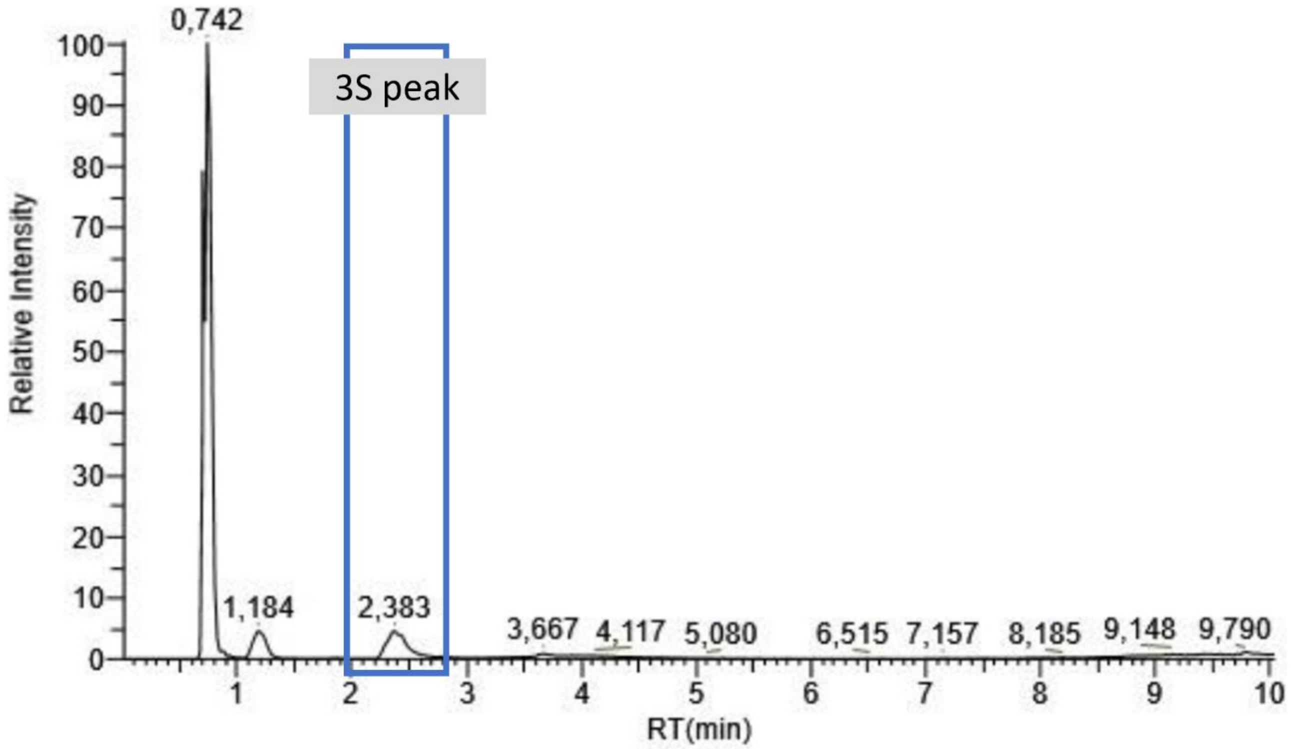


Figure 32 Typical chromatogram for assay results based on assay conditions with predominantly fully folded 3S BPTI. The typical elution time for 3S BPTI is highlighted with an elution peak that is visible and distinct.

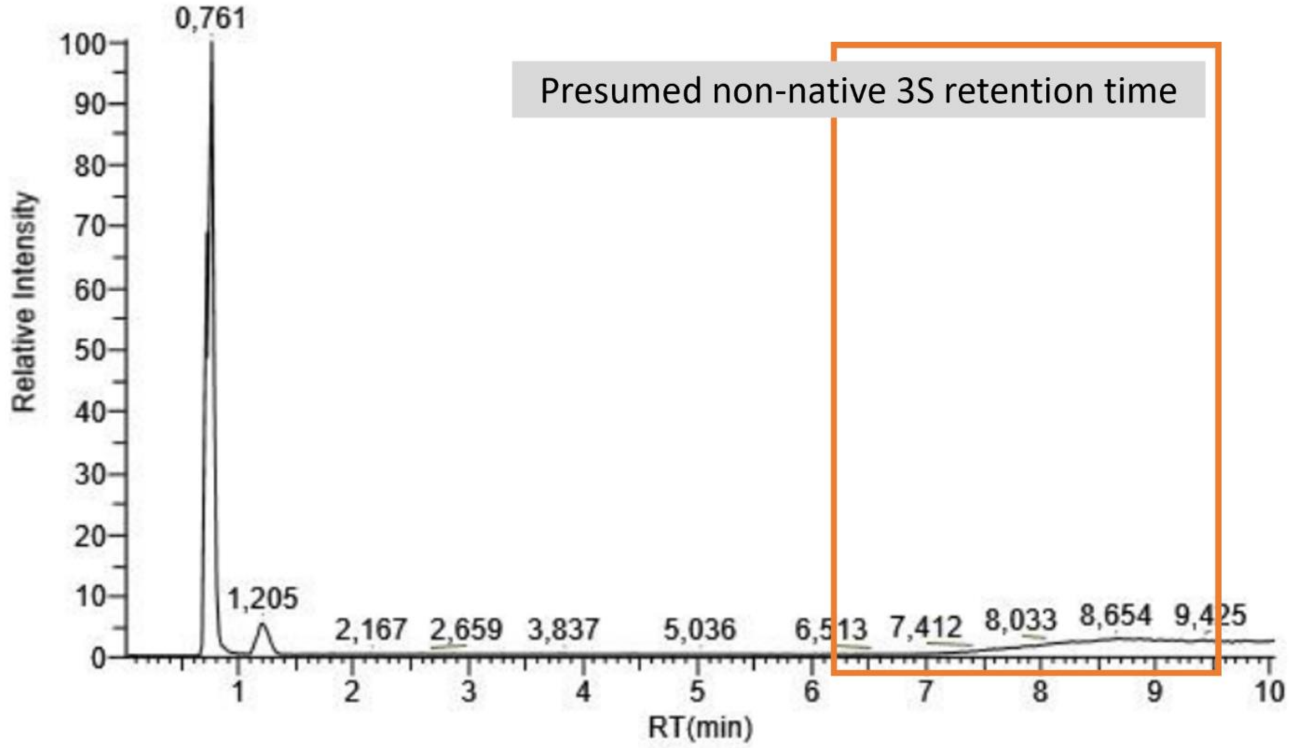


Figure 33 Chromatogram result for BPTI refolding assays with high amounts of ERp38 at time point 0. No distinct elution peak visible for either the predominant 2S species or the presumed non-native 3S species. The elution time frame for the non-native 3S as detected by the MS is highlighted.

High oxidation kinetics together with the non-native 3S BPTI species observed for ERp38 could hint at its role in the oxidative folding machinery of *P. pastoris*. The enzyme might employ a ‘fast and loose' approach to oxidative folding which results in extremely fast oxidation but might also result in more misfolded proteins, particularly in substrates with high DSB pattern complexity.

Peptide Oxidation

In the above experiments, the oxidative folding activities of pPDI, ERp38 (and hPDI) were investigated by observing the NEM mediated mass shift in oxidative folding intermediates as a result of a change in the availability of reduced cysteines. While this is a common approach to studying an enzymes oxidative folding activity it is not the only approach. Another strategy is to measure the change in fluorescence based on DSB formation. The formation of DSB changes the secondary and/or tertiary structure of a protein which can change the relative proximity of two amino acids which in turn can impact their fluorescence.

This approach does not require a full-length protein and a method for this approach had already been established at the host university; utilizing the 10 amino acid long peptide N-R-C-S-Q-G-S-C-W-N [179]. The amino acid tryptophan (W) is the strongest fluorophore of the three aromatic amino acids which leads to a strong fluorescent signal of the reduced peptide. However, when a DSB is formed between the two cysteines, the ends of the peptide are brought together. This brings the arginine (R) on the other side of the peptide into closer proximity with the Tryptophan. The charged arginine is a fluorescence quencher which results in a reduced fluorescence activity of the tryptophan and thus the peptide. This reduction in fluorescence is then measured and can be used to infer the rate of disulfide bond formation.

The results of this peptide refolding assay can be seen in Figure 34. The fluorospectrometer takes a measurement every 6 seconds with an integration time of 0.5 seconds. The resulting values are relatively noisy, and therefore the values displayed in Figure 34 are based on the averaging of 3 runs (with those averages being displayed as the individual points in the plot). The resulting trendlines are based on a 5-point floating average. Also, the absolute fluorescent starting point values for each series are noisy too and therefore each measurement series is normalized and converted to more comparable percentage-change values.

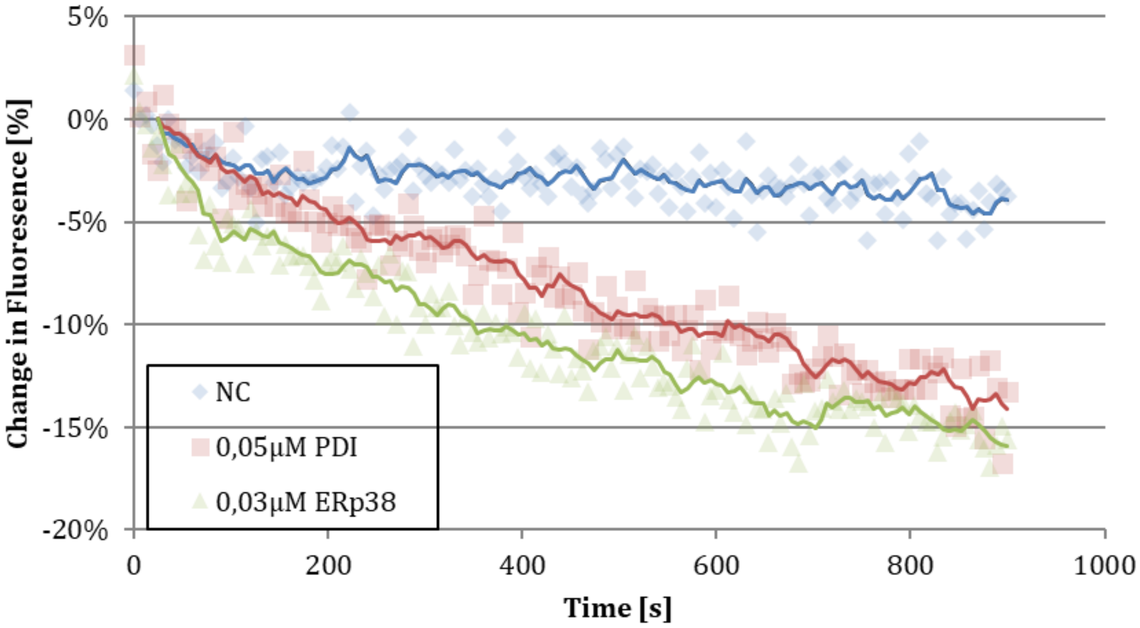


Figure 34: Results of the peptide oxidation assay for pPDI and ERp38 (as well as the uncatalyzed runs - NC). Each dot represents the average value at the respective timepoint based on three independent series measurements. The corresponding trendlines are calculated based on a 5-point floating average calculation.

The time series follows an exponential decay function which is displayed in Equation 7 and the fitting results together with their respective calculated kinetic values are displayed in Figure 35. Despite the strong noise signal in the fluorescence measurements a relatively good R^2 values were achieved for the fit with 0.924 and 0.929 for pPDI and ERp38 respectively. The values are calculated with a different substrate, buffer and measurement compared to the previous BPTI based analysis and are therefore not comparable. However, the values can be used to compare the two enzymes tested. ERp38, even though it's used concentration was only 3/5 of the of pPDI has been fitted to a more than twice as high kinetic value. When also considering the enzyme concentrations used, the rate constants can be calculated by dividing the k_1 values with their respective enzyme concentrations (Equation 5). In the present case this results in the values 0.022 and 0.077 s^{-1} μM^{-1} for pPDI and ERp38 respectively (disregarding the negligible impact of the oxidation from the glutathione buffer). This suggests that ERp38 is roughly 7 times as efficient at oxidatively folding the 10 amino acid long peptide compared to pPDI. For comparison the rate constants for the two enzymes and the first BPTI oxidation step only show a 2-fold higher rate constant for ERp38 compared to pPDI. This difference could be explained by pPDI having a higher binding affinity for BPTI compared to ERp38 or by ERp38 having a higher binding affinity for the short peptide (or both). Either might be linked to both enzymes having different tasks in the yeasts oxidative folding machinery. For a better understanding of their roles in their native environment more substrates would need to be tested, including native substrates.

$$peptide_{ox} = peptide_{red} * e^{(-k_1 * t)} + C$$

(Eq. 7)

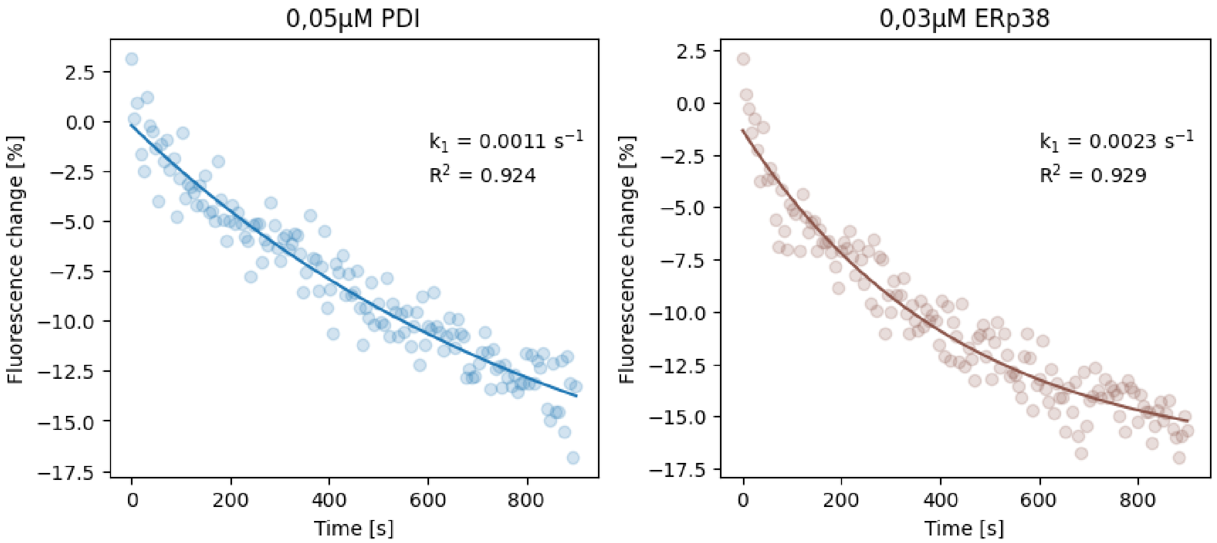


Figure 35 Curve fitting results of the peptide oxidation experiments using Equation 6. Kinetic parameters and coefficient of determination (R^2) are displayed in the top-right corner of each respective graph.

Discussion

Figure 30 shows how both pPDI and hPDI (with the same concentration) achieve 100% natively folded BPTI at roughly the same time (2 hours). However, the trajectories of their isomerisation step look different. Human PDI reaches a higher level of 3S species faster but then slows down. Pichia PDI takes longer in the beginning but does not slow down as much later on, as hPDI does. This behaviour can be described by the fitting with the double exponential equation and knowledge of the different 2S states. During the two oxidation steps, hPDI seems to have achieved a mix of roughly 60/40 between the two 2S species N' and N* respectively. And since the refolding of N' is faster compared to N*, the refolding assay with hPDI achieves a higher 3S yield faster (until roughly 60% is reached). On the other hand, pPDI achieves a ratio of roughly 40/60 between N' and N* which means that more BPTI must be isomerised through the slower pathway which explains the slower start. However, this alone does not explain why both reactions finish at roughly the same time. This can then only be explained by pPDI having faster kinetics compared to hPDI. And this can be seen in the fitted kinetic values for each respective enzyme with 0.027 s^{-1} / 0.334 s^{-1} and 0.022 s^{-1} / 0.28 s^{-1} for k_1 and k_2 as well as pPDI and hPDI respectively.

Looking at each step individually, pPDI is both the faster oxidase as well as the faster isomerase compared to hPDI. However, during the complete refolding process this kinetic advantage is cancelled out by hPDIs proclivity for forming more of the faster isomerising N' species. Since neither organism has BPTI as a native substrate, explaining this difference in behaviour through evolution is difficult although the cow is certainly closer related to humans than it is to yeasts. Another way to explain the different ratios of N' and N* between the two enzymes, could be a 'more careful' oxidation approach by hPDI compared to pPDI which results in the yeast enzyme producing more 2S faster but at the cost of forming more of the unfavourable N* species. If BPTI could be considered a representative example of the native substrates for both pPDI and hPDI, then this difference in oxidation pattern could hint at two different oxidative folding approaches between the two species. *P. pastoris* could be employing a 'fast and loose' approach while *H. sapiens* would be described by a more 'slow and careful' approach. This theory would link well with the results from the ERp38 assays which have shown it to be an extremely fast oxidase. So fast in fact, that it seems to catalyse the formation of a non-native 3S BPTI species which has not been described before for BPTI. However, these experiments are not sufficient to determine the overall oxidative folding relation between these two organisms, since this process is much more diverse and complicated and both organisms have a wide range of other oxidative folding enzymes and substrates. Nevertheless, the results presented in this Chapter provide important insight into the oxidative folding processes of *P. pastoris* and their characteristics compared to that of humans.

Chapter 3

Evaluating model organisms for their disulfide bond forming capabilities

Introduction

Disulfide bonds in POIs

Oxidative folding has been widely used by evolution to introduce function and stability into enzymes [183]. Disulfide bonded enzymes can be found in all domains of life and their (often complicated) folding process has been extensively studied for decades [184]. In biotechnology, DSB are of interest in two different areas. On the one hand, proteins of interest can have DSB which are essential to their folding or function. Unfolded or misfolded DSBs (or a mixture thereof) can introduce complications into their production, both in lab-scale as well as production-scale [185]. On the other hand, DSBs can complicate the production of POIs and put an increased strain on the selected production organism. Particularly microbial based protein production can struggle with complex DSB patterns and product titres can be infeasibly low as a result [5].

The production of disulfide bonded proteins at large scale has become an immense industry [186]. Many high demand and high revenue biopharmaceutical POIs include essential DSBs with the prime example being human growth hormone or antibodies such as Adalimumab (TNF-α blocker used against rheumatoid arthritis and related conditions) or Pembrolizumab (PD-1-inhibitor used in cancer immunotherapy) [19]. But DSBs are not only relevant in antibodies, many other biotechnologically relevant POIs, such as enzymes, also have DSBs [187]. Together these POIs form a spectrum of DSB complexity ranging from very simple to very complex DSB patterns. Another aspect to this DSB complexity is the organisms which are used for their production. These also range from very simple production hosts such as *B. subtilis* or *E. coli* to more complex ones such as mammalian cell lines. In principle, the complexity of the target POI dictates the required complexity of the producing organism [188]. However, appropriately matching the complexity of the DSB forming machinery in potential hosts to the requirements of the POI is currently still a process based on trial and error. A preferable approach would be to move towards a product development process that is based on biological understanding instead. The complex interaction between POIs and production host and the resulting production titres are in part linked to the POIs DSB pattern [156]. However, it is important to note at this point that this complexity is also heavily influenced by other factors such as other PTMs, protein secretion and growth conditions [5].

An organism's ability to produce DSBs is dictated by its respective oxidative folding machinery. For *E. coli* this machinery was thoroughly investigated in Chapter 1 and for the yeast *P. pastoris* novel insights into its oxidative folding machinery were presented in Chapter 2. For most other relevant organisms (excluding humans) this machinery is less well studied, with only the central enzymes such as PDI and ERO being well characterised. Human PDI is probably the most well studied oxidative folding enzyme and even though this is an immensely important enzyme in the oxidative folding machinery of humans, there are several other PDI family members in humans which are far less well studied [189]. This can be extended to other model organisms where the central oxidative folding enzymes might be characterised but the overall machinery is not well known. In most cases even the qualitative knowledge of which enzyme catalyses which part of the oxidative folding cascade of a given POI is very limited, and this knowledge becomes even more scarce when it comes to the quantitative aspect, which becomes highly relevant when trying to predict target POI titres.

What dictates an organisms oxidative folding capability is its own requirement for disulfide bonded proteins. In order to better understand and predict recombinant heterologous protein titres, it helps to first investigate an organisms own oxidative folding requirement and machinery. During the last decade the amount of available large scale biological data on both proteomes as well as protein structures has made it feasible to investigate an organisms oxidative folding machinery on a holistic level. Knowledge of each proteins disulfide pattern as well as its abundance can be used to reverse engineer the organisms oxidative folding machinery as demonstrated in Chapter 1.

In this chapter new genome-wide information is utilised to investigate the oxidative folding requirements (i.e. the prevalence of different types of DSBs) of these 12 model organisms: *E. coli, B. subtilis, S. cerevisiae, P. pastoris, C. albicans, A. niger, A thaliana, C. elegans, D. melanogaster, D. rerio, M. musculus* and *H. sapiens*. Not all of these are currently biotechnologically relevant production organisms or cell lines, however, they are among the most well studied organisms and as such have the highest level of available information and annotation currently available. Also, they cover a broad evolutionary spectrum of organisms from bacteria to fungi to plants to animals.

Methods and Materials

Python environment and packages

The analysis was done in python 3.6 as part of the anaconda distribution and utilising the jupyter notebook environment [190]. The packages numpy and pandas were used for basic data manipulation and basic mathematical operations [191], [192]. Matplotlib and Seaborn were used for data visualisation [193], [194]. The biopython package was used for handling the protein structure data as well as calculating the cysteine distances within the structures. The package os (Miscellaneous operating system interfaces) was used for interactions with Microsoft data structures (i.e. folders).

Handling the structure files

AlphaFold provides its structure files in the Crystallographic Information Framework file format (.cif). These files were read into Python using the biopython parser function MMCIFParser() [148]. The files downloaded from the protein structure database PDB were in the PDB file format (.ent). These were loaded into Python using the PDBParser() instead. For the model organisms (*E. coli, S. cerevisiae, H. sapiens* and *M. musculus*) the AlphaFold structure files were downloaded directly from the AlphaFold website. For all other organisms (*B. subtilis, P. pastoris* and *A. niger*) the structures were downloaded from the Google Cloud Public Datasets provided by AlphaFold under the CC-BY-4.0 licence.

In the human proteome, AF provides additional segmented structures for the proteins which are otherwise too long to model. For proteins which are more than 2700 residues long, AF splits the sequence into 1400 residue segments which are modelled as individual structures. These segments are overlapping by 1200 amino acids which results in the first structure covering amino acids 0 to 1400 and the second structure covering amino acids 201 to 1600 and so on. For appropriate DSB count estimation, DSBs where only counted in sections of fragmented structures which were not modelled in prior fragments to ensure that each DSB was counted exactly once.

Calculating cysteine distances

Distances between the cysteines in each protein were calculated using the coordinates of the cysteine sulfur atom. The distances were calculated using the distance formula for a three-dimensional space shown in Equation 8. While all possible distances between each pair of cysteines in a given protein were calculated only the distances between each cysteine and its closest neighbouring cysteine were considered for the subsequent analysis.

$$Distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

(Eq. 8)

Closest Cysteine Search

Protein structures annotate the spatial location of the atoms in a given protein. The amino acid cysteine has a single sulfur atom which can bond with another sulfur atom to form a disulfide bond. This sulfur atom is annotated as ‘SG’ in protein structure files. In the first step the ‘Closest Cysteine Search’ script extracts amino acid residue list from a structure file. Then, the coordinates of the sulfur atom in every cysteine amino acid are queried and stored in a temporary data frame. Next, the distance formula (Equation. 8) was used to calculate the spatial distance between every cysteine sulfur atom in a given protein sequence. From this matrix of spatial distances, the closest neighbouring sulfur for every sulfur in the protein is selected and stored in a data frame alongside the sequence position of the respective cysteine and the calculated distance to its nearest neighbour.

Disulfide bonds and their complexity

Anytime the distance between two sulfur atoms in their respective cysteine in a given protein structure is below 3 Å they are considered disulfide bonded. In cases with more than two cysteines in a protein sequence, a disulfide bond can be either consecutive or non-consecutive [195]. Consecutive disulfide bonds are defined by the absence of cysteines that occur in the primary sequence between the two disulfide bonded cysteines. Non-consecutive cysteines on the other hand are always interrupted by cysteine occurring in-between the sequence positions of the disulfide bonded cysteines. Based on this definition the identified disulfide bonds were classified into consecutive and non-consecutive disulfide bonds.

Quantitative estimation

Quantitative proteomes were collected from the protein abundance database PaxDB for all selected model organisms except for *P. pastoris* for which no entry was available [139]. This database creates an integrated quantitative proteome based on individual quantitative proteomes for the respective organisms, thus creating a baseline quantitative proteome. The abundance values are given in parts per million (ppm) which are used for both intra-organism comparisons as well as inter-organism comparisons.

Results & Discussion

The AlphaFold protein structure predictions for *E. coli, S. cerevisiae, H. sapiens, M. musculus, A. thaliana, C. albicans, C. elegans, D. melanogaster, D. rerio, A. niger, P. pastoris* and *B. subtilis* were used in this analysis. The Gram-negative bacterium *E. coli* is central to biotechnology and the probably most commonly used organism in laboratories worldwide. *B. subtilis* is not only a common model organism for Gram-positive bacteria, but also widely employed as a recombinant protein production host. The three yeasts *S. cerevisiae, P. pastoris* and *C. albicans* are also included in this analysis. *S. cerevisiae* has been employed in biotechnology for centuries and is among the most intensely studied organisms in biology. *P. pastoris* is a methylotrophic yeast and is increasingly relevant in yeast based recombinant protein production. *C. albicans* is commonly found in the human gut microbiome. It is also an opportunistic pathogen that can cause death in immunocompromised patients. As a fourth representative of the fungi kingdom *A. niger* is also included in this analysis. *A. niger* is a filamentous fungus which is heavily employed in the production of citric acid and enzymes such as glucoamylases. The plant *A. thaliana* is a commonly used model organism for studying plant biology with a relatively small genome size compared to many other plants. *C. elegans* is the most studied representative of the Nematoda phylum and a commonly used organisms for studying gut microbiota and aging in animals. The zebrafish *D. rerio* is the most studied representative of the aquatic vertebrates and commonly employed in drug development. The invertebrate fruit fly *D. melanogaster* has been a model organism in biology for decades and is still heavily used in modern genetics and physiology research. The first of the two mammals represented in this study is the mouse *M. musculus*. No other animal has been as instrumental in medical research for human health and development as the mouse. And lastly, we've included *H. sapiens* into our analysis to complete the list of organisms relevant for biology and biotechnology.

The key data discussed and displayed in the following paragraphs is listed in the summary table at the end of the Results and Discussion section (Table 5).

Qualitative insight in the disulfide bonds of organisms

Figure 36 shows the full-length histograms based on all calculated closest inner-protein sulfur distances for *E. coli* and *H. sapiens*. The figure shows the raw data output of the closest-cysteine-search script applied to the AF proteomes of both organisms (Table 5). Both organisms have a few outlying cysteine distances which are much further apart from each other compared to the rest of the proteome. These outliers come mostly from very elongated proteins with isolated cysteines on either side. These most extreme outliers are P76237 and Q6ZMV7 for *E. coli* and for *H. sapiens* respectively. Particularly the later example is predicted to have an extremely long tail and the accuracy of the AF prediction for such proteins must be questioned. However, in this work we are more interested in the closely distanced cysteines which is why Figure 37 shows the more suitable histograms focused only on the short inner-protein cysteine distances for all 12 organisms.

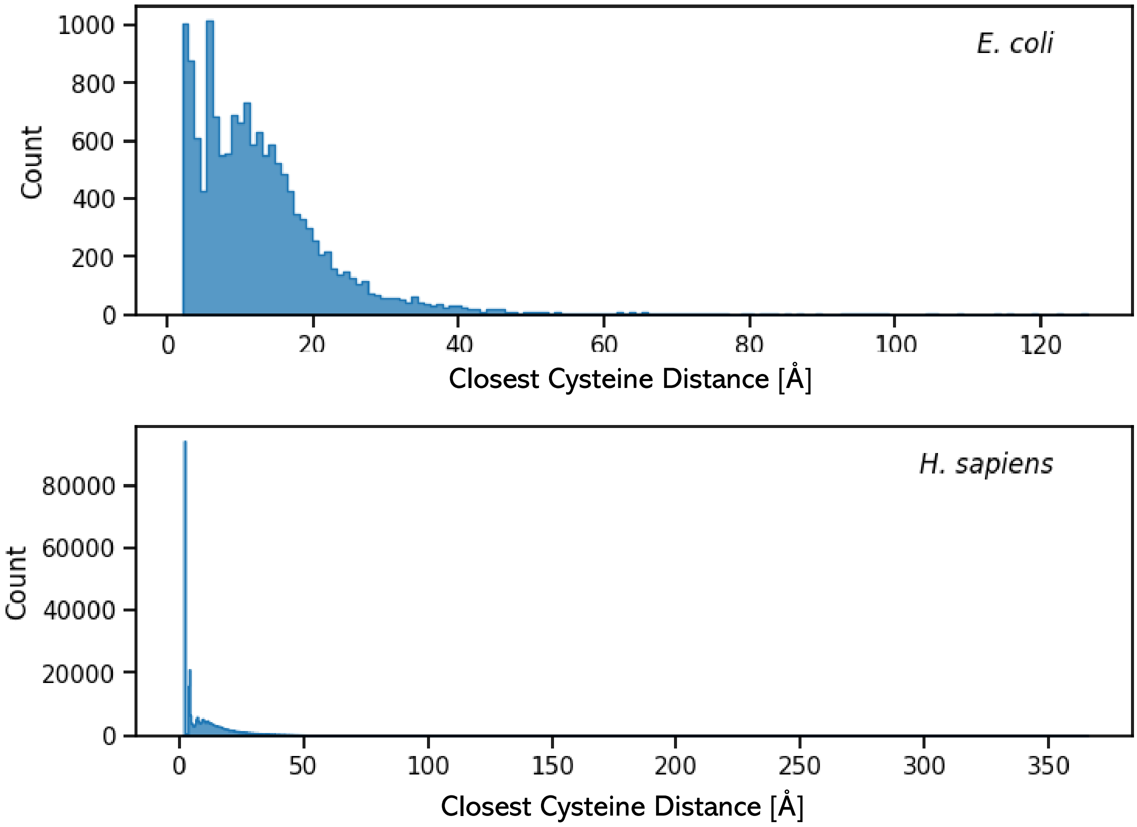


Figure 36 Histograms of the inner-protein cysteine distances in the AF predictions for the E. coli and H. sapiens proteomes.

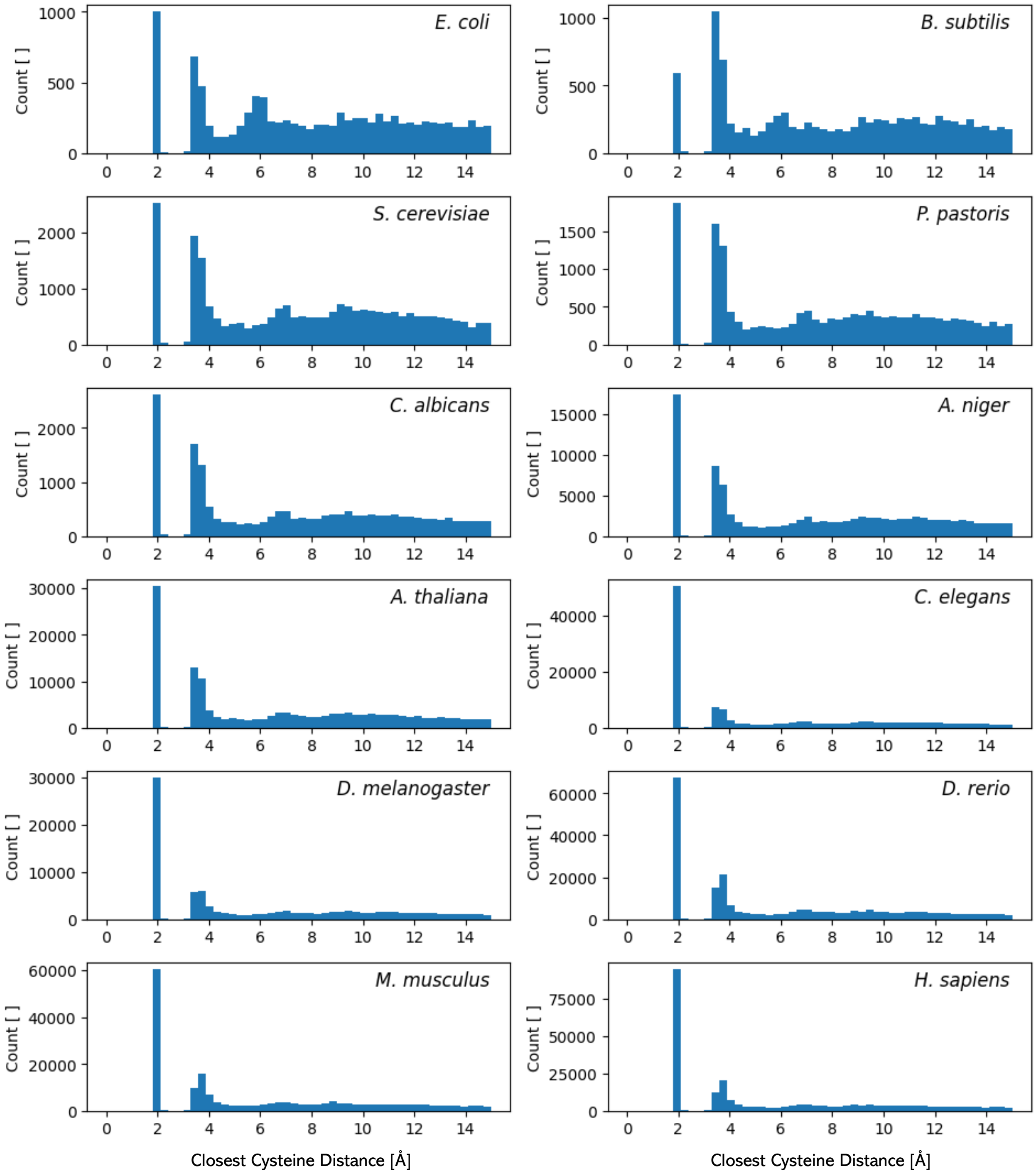


Figure 37 Histograms for the closest inner-protein cysteine distances for each protein predicted by AF for the 12 selected model organisms. The histograms show only the closest cysteine distances between 0 and 15 Å.

Clearly visible is the ‘peak’ of distances at around 2 Å followed by a clear gap between it and the next lowest distances starting at above 3 Å. A disulfide bond has a rough length of 2.05 Å and the PDB database uses 3 Å as the cut-off for disulfide bond classification [195]. As such the observed lack of distances between 2.5 and 3 Å is a good indication that AF does differentiate between disulfide bonded cysteines and cysteines that are merely in close proximity to each other without *necessarily* forming a disulfide bond.

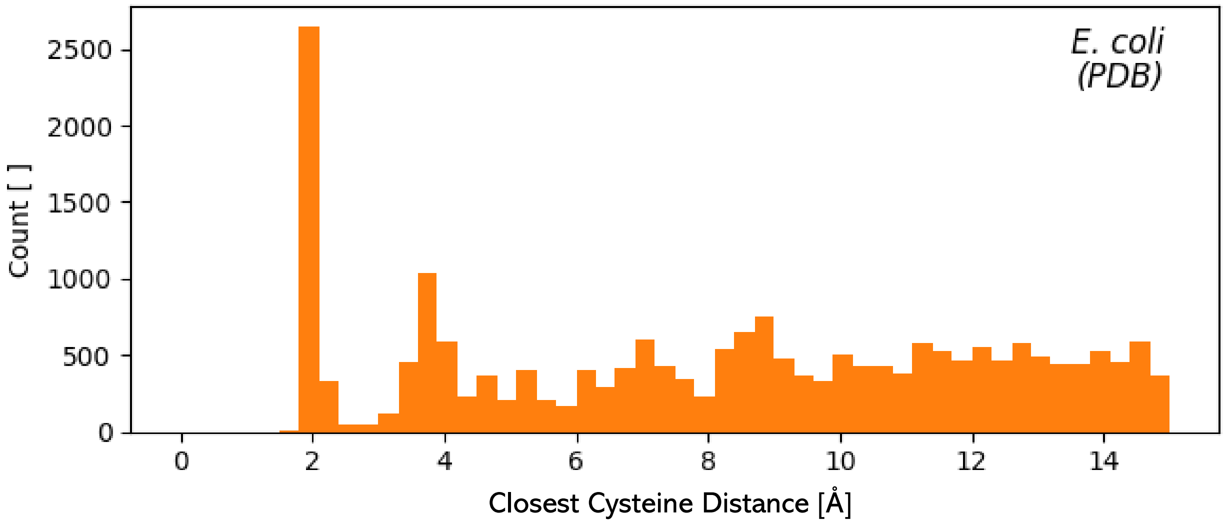


Figure 38 Cysteine Distances calculated from 4257 E. coli PDB protein structures.

This gap can also be observed in measured protein structures which is presumably where AF learned the behaviour from. An example of this can be seen in Figure 38 which displays as histogram based on 4257 PDB structures of *E. coli* proteins which have a resolution below 3 Å. And while the resulting histogram does not display the gap as clearly as the histograms based on modelled structures, the gap is nonetheless clearly visible in Figure 38. Chemically speaking this gap is based on the steric repulsion two SH-groups exhibit towards each other. This results in a minimum distance between the sulfur atoms of apparently roughly 3 Å whenever they are not connected by a covalent disulfide bond.

Most importantly however, in all 12 histograms in Figure 38 a large number of AF predicted DSBs can be observed. The qualitative and quantitative implications of these DSBs will be discussed thoroughly in this chapter, however, when looking at the Y-axis scale of the 12 histograms in Figure 37 we can already see that the number of predicted DSBs varies over more than 2 orders of magnitude between the different model organisms.

It is important to note here that cysteines with predicted sulfur atom distances above 3 Å can still form disulfide bonds but that these disulfide bonds are not included in the AF structure prediction and can therefore not be detected with the here employed method. These unpredicted DSBs do however impact the PDB based DSB quality control matching discussed further below.

Other closest-cysteine distances of interest

Next to the DSB ‘peak’ and the 3 Å gap two areas stand out from the rest of the histograms with much higher ‘peaks’ compared to the rest of the histograms. The first one at around 3-4 Å which is clearly visible in all 12 organisms and the other one at around 6 Å which is only distinct in the two prokaryotic organism histograms (Figure 37).

Zinc fingers

As mentioned above, the 3-4 Å closest-cysteine-distance spike can be partially explained with non-disulfide bonded cysteine pairs which could potentially form a disulfide bond between them. However, this 3-4 Å distance is also displayed in cysteine residues in one of the most commonly found protein structures, the zinc finger [196]. The most common variation of this feature contains a zinc ion (Zn^{2+}) coordinated by two cysteine residues on one side and 2 histidine residues on the other side. The centrally bound zinc atom created a pointed (finger-link) protein structure which is particularly suited for interacting with DNA, but also RNA and some proteins. As such the zinc finger is most commonly found in proteins linked to regulatory functions.

The distance between the cysteine sulfur atoms in this structure is mostly determined by the zinc-sulfur bond length and the angle of the cysteine residue towards the central zinc ion. Variations in these two parameters creates a range of potential sulfur atom distances which can be observed by measuring inner-protein cysteine-sulfur distances. This distance range can be described using protein structures from both AF and PDB and are displayed in Figure 39.

The data displayed in the top histogram in Figure 39 is based on AF structures of the 12 model organisms for which zinc fingers have been annotated in the UniProt database (5270 entries). The lower histogram is based on 1980 PDB structures which have Zn^{2+} as a ligand. Since zinc is not only found in zinc fingers the resulting zinc finger ‘peak’ is not as pronounced as the one above based on AF and proteins which have specifically annotated zinc fingers. Nevertheless, both histograms clearly display the distinct ‘peak’ just below the 4 Å distance.

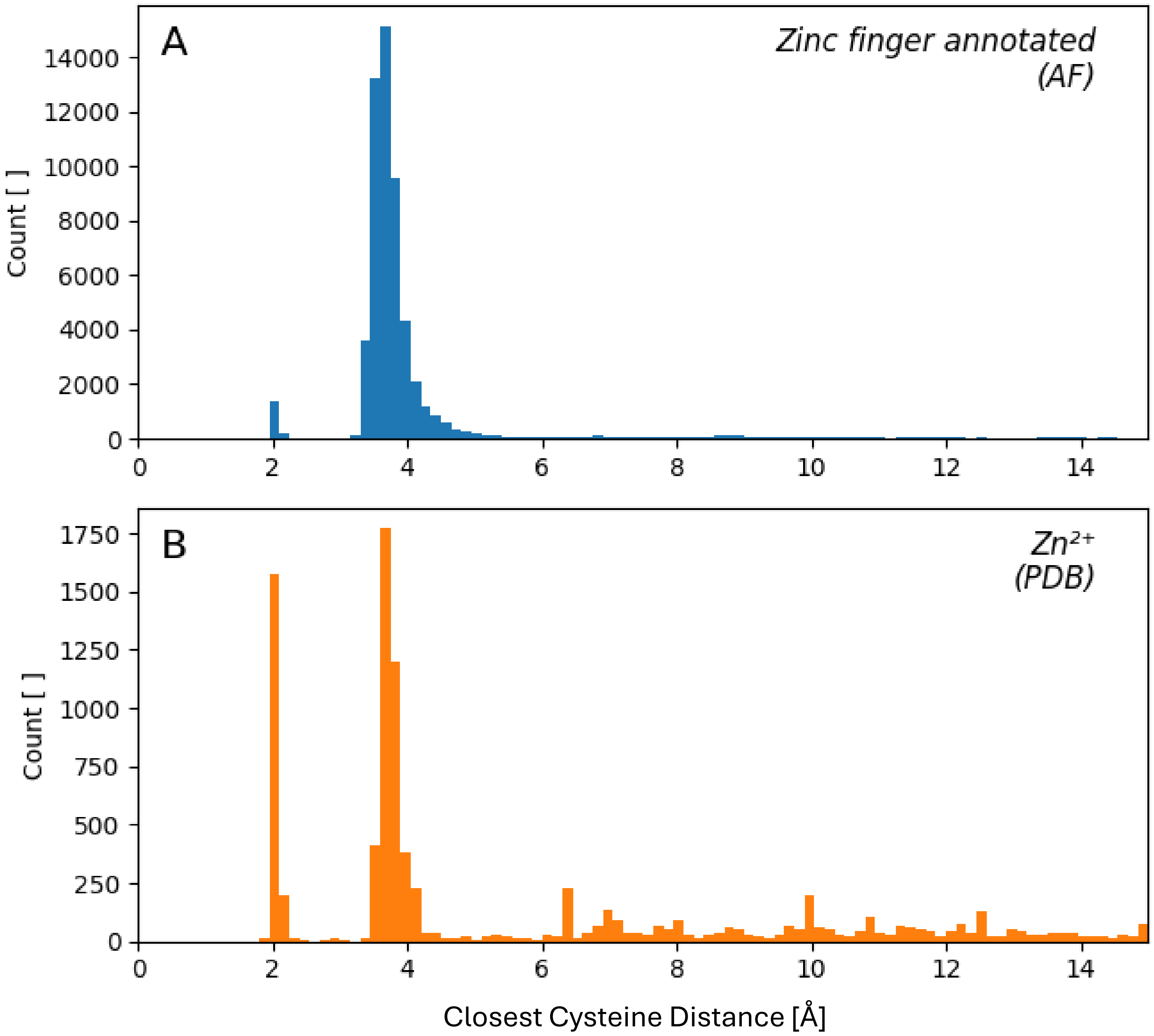


Figure 39 Comparison between the cysteine distances extracted from AF structures and PDB structures. A: Cysteine distances in proteins with from 12 model organisms which have a zinc finger annotated in their UniProt proteomes. B: Cysteine distances in PDB structures which have Zn^{2+} reported as a ligand.

Iron sulfur clusters

Iron sulfur cluster are combinations of iron and sulfur atoms which together can form several different clusters and can function as co-factors in a variety of proteins [197]. The three most commonly found clusters are [2Fe-2S], [3Fe-4S] and [4Fe-4S] (Figure 40). The first two are often found in proteins linked to one-electron-transfer reactions while the later has a more diverse set of functionalities.

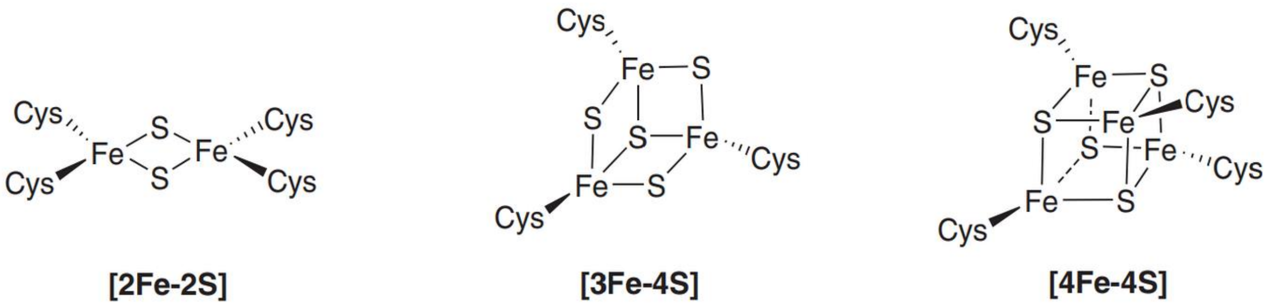


Figure 40 Three most commonly found types of iron sulfur clusters in proteins. Figure taken from [197]

As can be seen in Figure 40, these clusters are coordinated by cysteine residues. And while the sulfur atoms of the iron sulfur clusters do not show up in the AF structures, the sulfur atoms of the coordinating cysteines do. With the same approach as used with the zinc finger above, the AF structures of all 12 model organisms with at least one of the three clusters annotated in their respective UniProt entries are used to calculate a histogram of their closest cysteine distances. The resulting histogram can be seen in Figure 41.

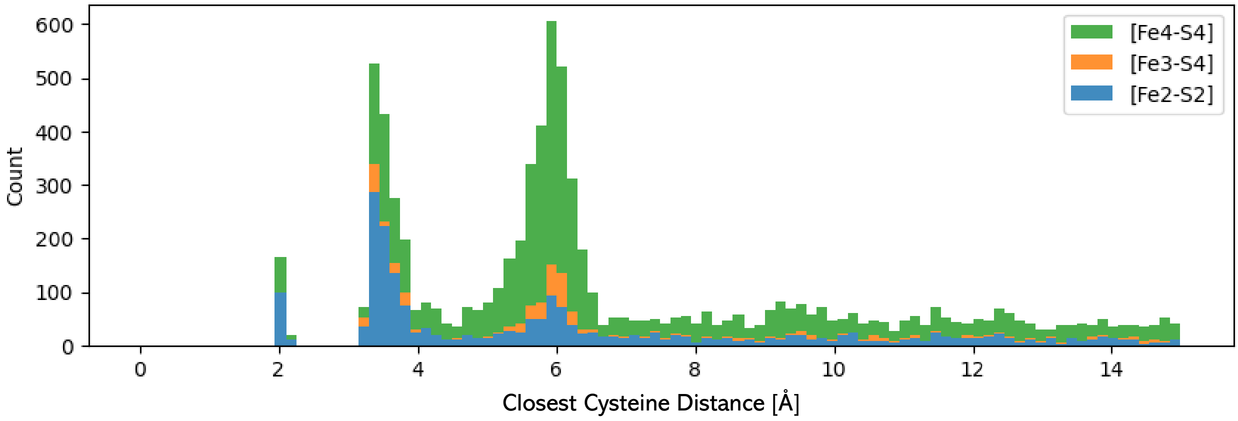


Figure 41 Cysteine distances calculated from AF structures of 12 model organisms which are linked to proteins that have at least one of the three common iron-sulfur clusters annotated.

Compared to zinc fingers, far fewer proteins have annotated iron sulfur clusters with only 667 proteins from the 12 model organisms having at least one of the three iron sulfur clusters annotated. This difference can also be observed in the y-axis scales of the two corresponding figures. Nevertheless, three distinct areas can be observed in the iron sulfur cluster histogram: a relatively small DSB ‘peak’ as well as two main cysteine distance areas. The first one between roughly 3 and 4 Å is mostly occupied by proteins containing [Fe2-S2] and [Fe4-S4] clusters whereas the second area centres around 6 Å can be mostly associated with [Fe4-S4] clusters.

The distance between the coordinating sulfur atoms is dependent on three factors: the type of iron-sulfur cluster bound, the bond length between the sulfur atoms and the iron atoms and well as the angle at which the cluster is coordinated by the cysteine residue. While the sulfur – iron bond length is not a fixed value it can be influenced by factors such as electronegativity of neighbouring residues or other surrounding atoms/molecules. Nevertheless, the margins for this value are relatively small compared to the impact the coordination angle has on the observed distance between the cysteine-sulfurs. Furthermore, when looking at the geometries of the 3 clusters in Figure 40: one can see that the [Fe2-S2]-cluster is smaller and has fewer bonds separating the coordinating cysteines compared to the [Fe4-S4]-cluster. It is therefore unsurprising that most observed distances in proteins containing [Fe2-S2]-clusters are in the first histogram cluster around 3-4 Å. Compared to the [Fe4-S4]-clusters which are mostly associated with the ~6 Å distance area. However, even in proteins containing only the relatively large [Fe4-S4]-cluster, a substantial amount of cysteine distances in the 3-4 Å area were measured. The cysteine distance measured for proteins containing the [Fe3-S4]-cluster looks similar to the [Fe4-S4] clusters although with far fewer annotation entries.

In similar fashion to the zinc finger validation above, every protein structure in the PDB which has at least one of the three iron sulfur cluster annotated as a ligand was downloaded and the closest cysteine distance was calculated. The result can be seen in Figure 42 which displays the histogram based on 1966 PDB structures. The resulting histogram pattern looks very similar to the one based on the AF structures in Figure 41. One difference that can be observed is the slight shift in both 'peaks' in the PDB based histogram compared to the AF based one. The shift is roughly 0.2 Å towards larger cysteine distances in the PDB based cysteine distance measurements. This result suggest that AF does make a slight systematic error in placing the coordinating cysteine which could be due to the missing ligands binding force exerted on the cysteine-sulfurs.

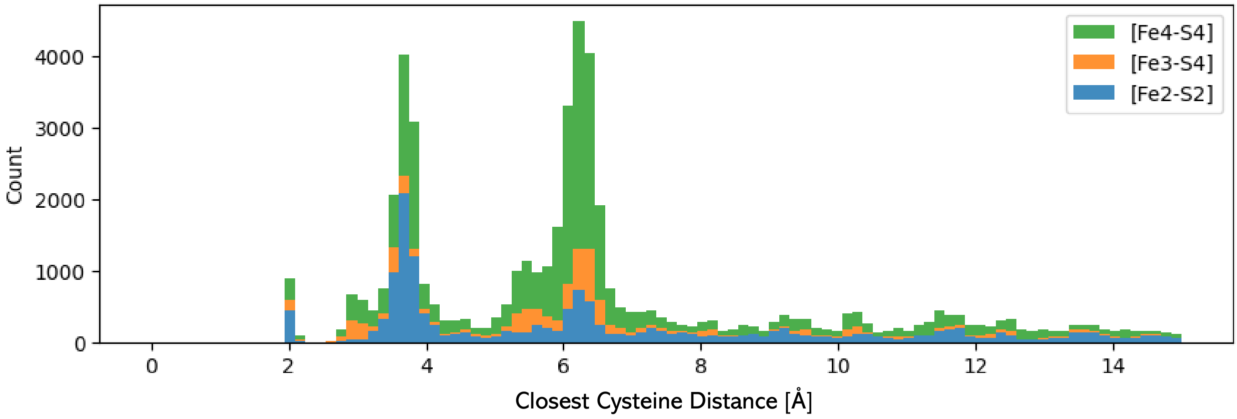


Figure 42 Cysteine distances histogram based on all PDB structures which have one of the three most common iron-sulfur clusters annotated as a ligand in their structure.

In some earlier studies on iron sulfur clusters by Morales et al. the geometry of the cysteine-sulfur and the [Fe2-S2]-cluster was measured [198]. In a ferrodoxin from Anabaena sp. (renamed to Nostoc sp.) the cysteine sulfur – iron atom bond length as well as the angle spanned by the cysteine sulfurs and the iron atom were calculated. The two S-Fe distances were measured as 2.3 Å and the angle as 106.3°. This triangle can then be solved to obtain the distance between the cysteine sulfur atoms which is 3.7 Å. This is also the cysteine distance measured most often in the PDB files displayed in Figure 42. This result strengthens the claim that the AF based predictions are slightly off from the PDB based measurements and not the other way around.

In a similar fashion, the geometries of [Fe4-S4]-clusters have also been calculated, however, no data on angle of coordination by the cysteine sulfurs is available [199]. Furthermore, the possible combinations of coordination angles of the four cysteines in combination with their effects on the resulting closest sulfur distances makes this calculation significantly more challenging compared to the one used for the [Fe2-S2]-clusters above. Nevertheless, the values for bond-lengths between the [Fe4-S4]-cluster irons and the coordinating cysteine sulfurs as well as the inner-cluster bond lengths are available and can be used to calculate a theoretical ideal cysteine sulfur distance. Assuming the cluster is coordinated by 4 evenly spaced cysteines, the resulting shape of the cysteine sulfur atoms can be described as a tetrahedron. Together with the bond length measurements for the Fe-S distances in the cluster (2.27 Å average) and the bond length measurements for the cysteine sulfurs and the iron atoms (2.72 Å average) and the tetrahedral bond angle of 109.47° the ideal distance between the cysteine sulfur atoms can be calculated as 6.91 Å. While this is not the maximum distance that can be observed, since 3 of the 4 coordinating cysteines can be closer to each other than to the 4th cysteine, the possible coordination angles that lead to cysteine distances above the 6.91 Å value are limited compared to potential smaller distances. Both the AF and the

PDB based cysteine distance histograms show drop-offs in observed distances at or shortly before this theoretical ideal value.

Ligand based cysteine distances without ligands

AF does not include ligands in their structure predictions, at least not explicitly. As is clear from the two protein features discussed above (zinc-fingers and iron sulfur clusters), AF structures can accurately predict cysteine positions in protein features that are centred around ligands. While this might not be true for each individual protein predicted by AF, it is certainly true for the 12 model organisms discussed in this chapter when summarized on their proteome level. AlphaFolds' two main features are learning from measured protein structures and multiple sequence alignments. And while this approach is not unique to AF, the amount of predictability this combination has achieved is impressive. However, one has to assume that it has its limitations. Novel proteins without available measured structures which use unknown approaches (i.e. sequences) for binding iron sulfur clusters or forming zinc fingers, will presumably not predict cysteine positions accurately without the inclusion of potential ligands into the prediction process. Maybe this will be possible in future iterations of AF. On the other hand however, as long as the sequences are close enough to other known and measured sequences, AF structure predictions can be used to identify ligand binding protein structures. When combined with sequence alignment and cysteine-distance measurements, novel zinc finger or iron sulfur cluster containing proteins could be identified in other proteins.

The human AF proteome

AF only predicts protein structures up to a length of 2700 amino acids. This means that some proteins are excluded from their proteomes. However, for human proteins longer than 2700 residues they provide fragmented structures which together cover the whole length of even the longest human proteins such as Titin at an approximate length of 34350 amino acids (UniProt ID: Q8WZ42). AF does this by splitting proteins which are 2700 residues or longer into 1400 residue long fragments. These fragments overlap by 1200 amino acids. This results in the first structure covering amino acids 0 to 1400 while the second fragment covers amino acids 201 to 1600 and so on. For counting DSBs this becomes an issue which needs addressing so not to over-count DSBs which are located in the overlapping structure areas. This was addressed by only counting the area of fragment-structures which have not been covered in previous structures. This significantly improves accuracy of DSB counting compared to counting every fragment fully or excluding fragments all together. However, this still results in an underestimation of DSBs which are formed over longer sequence-distances that would connect different fragment-structures. However, the resulting underestimation is compensated by the improved estimation quality derived from the additional coverage of extra-large proteins in the human proteome compared to the other AF proteomes which do not include proteins longer than 2700 amino acids.

Quality control of Predicted Disulfide Bonds

Before further analysis of the AF predicted DSBs, it is important to first assess the quality of their prediction. This was done by comparing the DSBs in each available PDB structure for any of the 12 model organism proteins with their respective AF predictions. A total of 17522 DSBs were found in relevant PDB files 16021 of which were correctly predicted by AF. Important to note, roughly two thirds of all PDB DSBs are from *H. sapiens* proteins (Table 5). The resulting matching quality for each individual organism can be seen in Figure 43. Out of the 1501 unmatched DBSs, 1420 where matched to AF files but not predicted as DSBs. 81 structures were not successfully matched. This was mostly due to differences between the sequences (although both using the same UniProt ID) or linking different proteins-isoforms. The 1501 unmatched DSBs could be in part due to constant changes and updates made to the UniProt database which might result in wrongfully matches PDB and AF structures. Furthermore, it is important to note that PDB files can also include wrongly predicted DSBs and in some cases the AF prediction might be closer to the correct value. This is further complicated by the fact that not all potential DSBs are constantly found in a disulfide bonded state which might make both predictions correct at different timepoints or occasions. Nevertheless, the overall overlap between the two source of disulfide bonds is good and provides the confidence in the AF predicted DSBs required for the following parts of this Chapter.

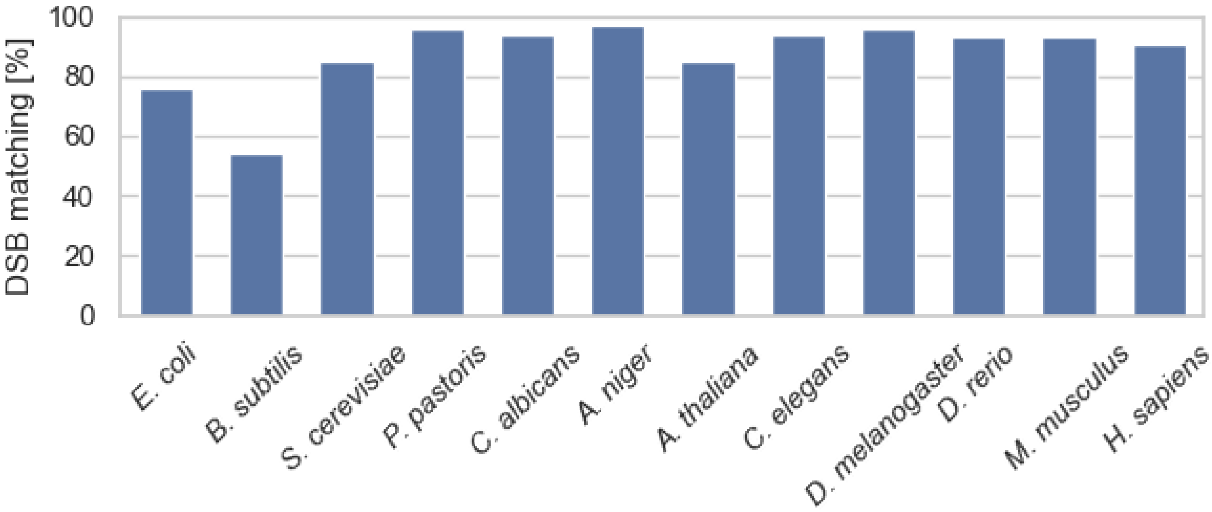


Figure 43 Comparison between the annotated DSBs in each proteome according to the UniProt database and the corresponding predictions from AF for each given DSB.

For most organisms the UniProt annotated disulfide bonds are well predicted in AlphaFold with 10 out of 12 being matched at above 85%. For the two bacteria *E. coli* and *B. subtilis* the prediction drops down to 76% and 54% respectively. This is surprising since AF structure predictions are based on multiple sequence alignments and existing protein structural knowledge. This makes it unlikely for *E. coli*, one of the most well studied organisms, to show a worse prediction accuracy for the AF structures compared to some of the other, less well studied model organisms.

The numbers of annotated DSBs for the two bacteria are relatively low with 101 and 53 respectively so any prediction above 3 Å has a big impact on the matching percentage. Furthermore, for the two bacteria 41% of the annotated DSBs with an above 3 Å AF distance prediction have a predicted distance between 3 and 4 Å compared to only 14% for *H. sapiens*. This indicates that the AF prediction for these two organisms is at the very least not far off for the non-DSB predictions of AF. Another explanation could be that these two organisms have a larger number of non-folding essential DSBs compared to the other organisms. However, further structure information would be required to validate this observation.

Cysteine and Disulfide Bond Prevalence in Selected Model Organisms

Cysteine usage

Before AF, predicting the overall DSB count of an organism required looking at the organism's cysteine usage. Cysteine is among the rarest of the amino acids found in proteins and has the seconds highest synthesis cost of all amino acids (8 ATPs in *E. coli* on glucose) [200]. This results in an evolutionary pressure to remove unnecessary cysteines from proteins which in turn corresponds to most cysteines being involved in functions or reactions that involve their unique sulfur atom and their redox activity. The percentage cysteine usage in a given organism is therefore already a good indication of an organism’s prevalence to using disulfide bonds in its proteins (Figure 44).

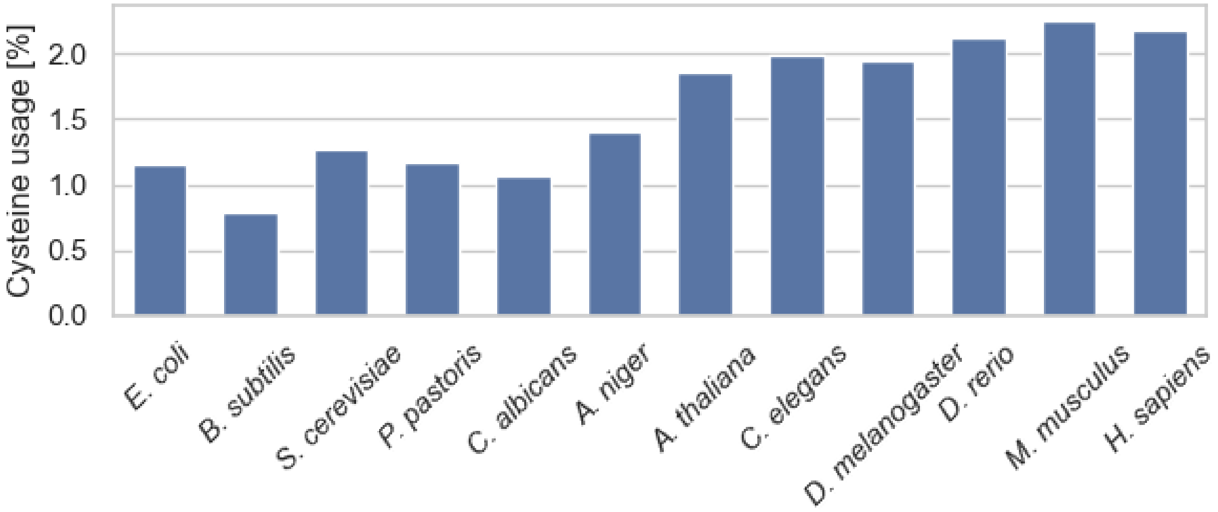


Figure 44 Cysteine usage in proteins compared to the other 19 amino acids in each of the 12 model organism proteomes

Within the microbes*, E. coli, S. cerevisiae, P. pastoris* and *C. albicans* all share a similar preference for cysteines. *B. subtilis* is the organism in this list with the most limited usage of cysteine (0.79%) and the filamentous fungi *A. niger* shows the highest affinity for it (1.41%). The remaining eukaryotic organisms all have a higher cysteine usage than the microbials, with the two mammalian organisms *M. musculus* and *H. sapiens* being the highest (2.26% and 2.18%).

Disulfides

We can observe a minimum of 0.07 disulfide bonds per protein in *B. subtilis* and the maximum value achieved by *H. sapiens* with almost 1.5 disulfide bonds per protein on average (Figure 45). For most organisms the relative prevalence for disulfide bonds aligns well with the observed cysteine usage. The biggest differences can be observed for *C. elegans* with 1.28 DSB per protein on average. When compared to *A. thaliana* which has a similar cysteine usage compared to *C. elegans*, the latter has more than two times the average DSB usage per protein.

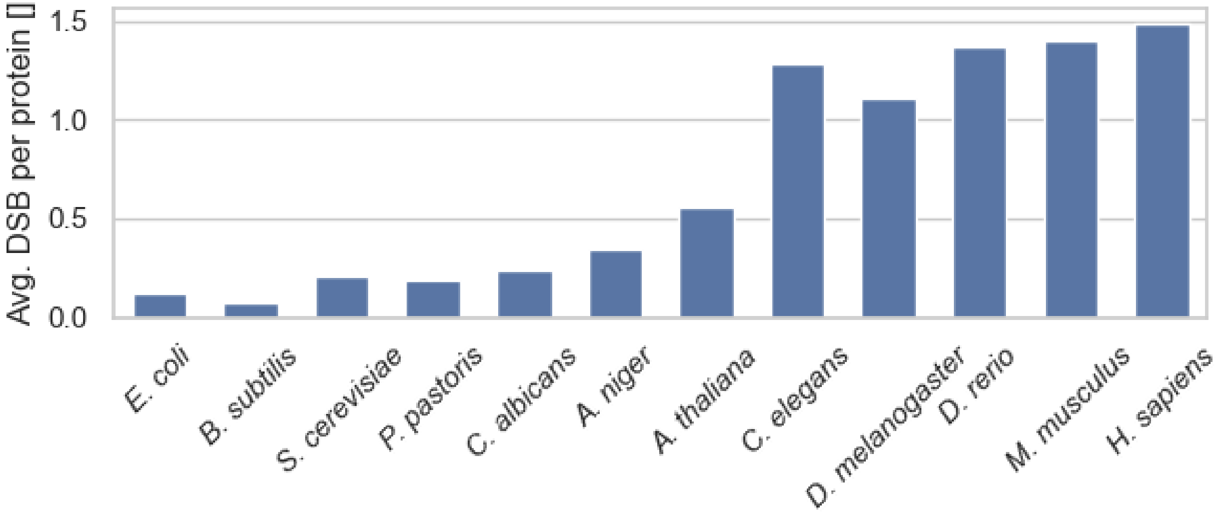


Figure 45 Average predicted DSB count per protein per organisms according to AF.

As organisms become more complex the average size of their respective proteins increases. The average DSB count per 1000 amino acids can be used as an alternative measurement to the above discussed DSBs-per-protein metric. These values are not influenced by the increased average protein size of more complex organisms and are displayed in Figure 46.

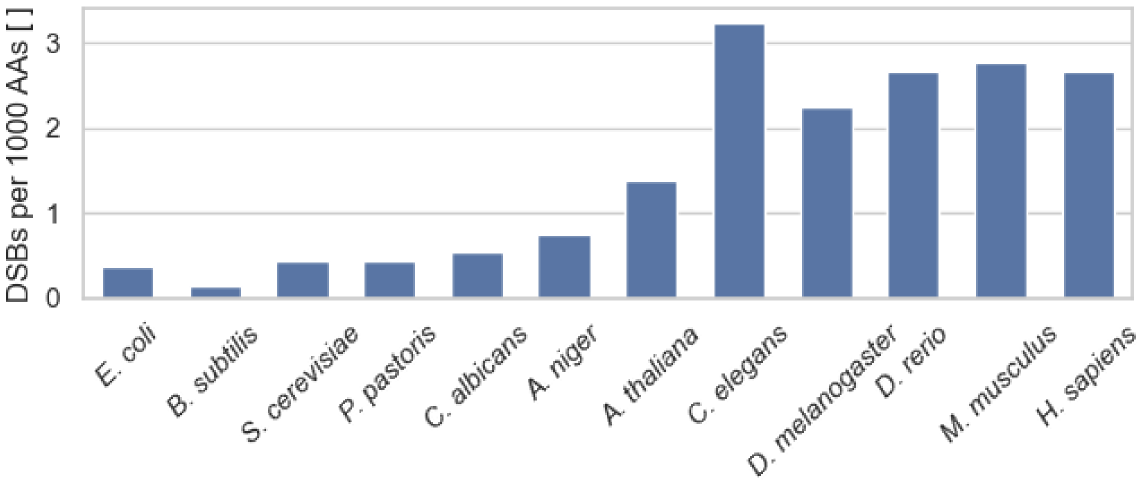


Figure 46 Average DSB count per 1000 amino acids for the 12 model organisms. Based on DSBs predicted in AF.

The two figures above show a similar trend for all organisms except *C. elegans* which stands out even more in Figure 46 compared to Figure 45. Several aspects can be identified as the source of *C. elegans* high DSB per 1000 AA value. The total amount of predicted DSBs is 25284 which is much higher compared to most of the other model organisms. When at the same time, *C. elegans* has the shortest average proteins length out of all 10 eukaryotes investigated here (397 AAs per protein, Table 5). When calculating the DSBs per 1000 AA values, these two outlying values amplify each other resulting in more than 3 DBSs per 1000 AAs on average in *C. elegans*. Another possible explanation for the comparatively large number of DSBs per 1000 amino acids identified in *C. elegans* is their large G-protein-coupled receptors (GPCRs) content. These proteins are essential for signal transmission from the outside to the inside of the cell and have been extensively studied in *C. elegans*. The functionality of this abundant group of proteins has previously been linked to DSBs and in *C. elegans* roughly 7% of their whole proteome has been predicted to code for them [201], [202]. The combination of the comparatively small average protein size and the elevated amount of DSB containing GPCR proteins can explain the observed outlier in DSBs per 1000 amino acid metric.

As mentioned above, the cysteine usage of an organisms has been commonly used as the predictor for DSB content in an organism. Following this centuries widespread adaptation of whole genome sequencing, the cysteine usage value became readily available for many organisms. Now, with the rise in structure prediction quality and the readily available AF predictions for basically any organism[148], it is important to assess if the two values – cysteine count and AF predicted DSB count - are correlating well. In Figure 47 this correlation is calculated for cysteine usage and DSBs per 1000 AAs for the 12 model organisms (same data as in Figure 44 and Figure 46).

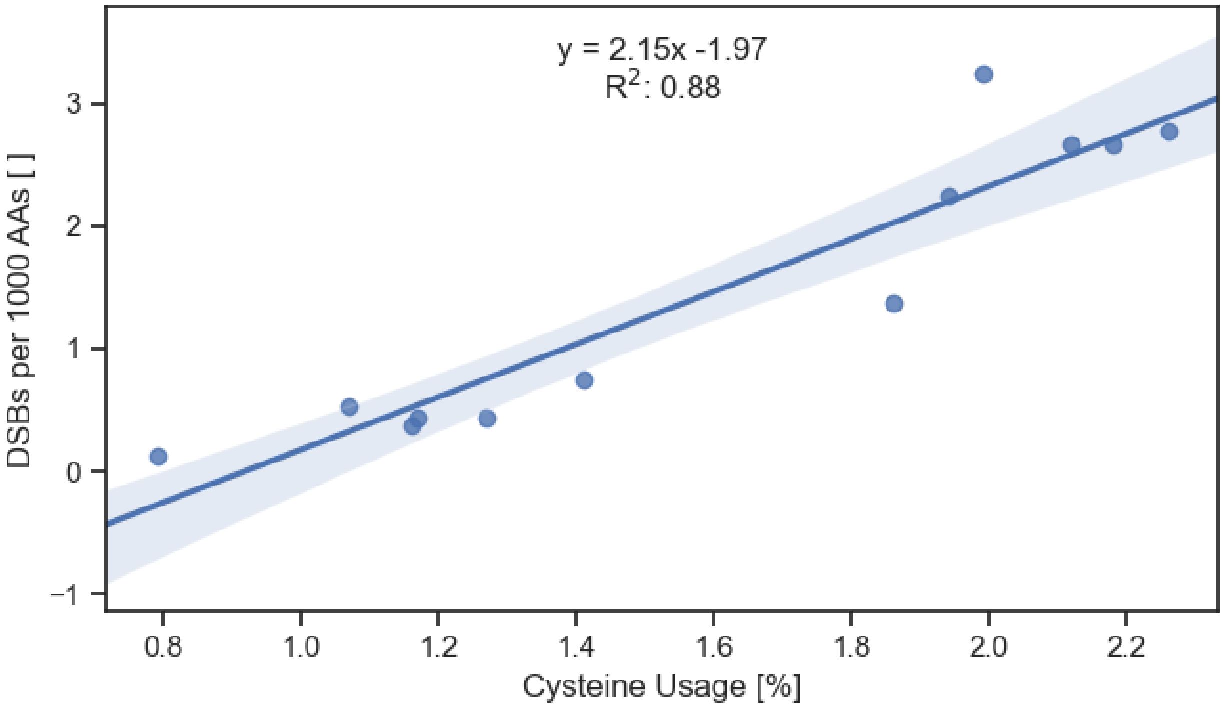


Figure 47 Correlation plot between cysteine usage and DSBs per 1000 AAs as predicted by AF for the 12 model organisms. The trendline and the coefficient of determination were calculated and are displayed. The trendline is surrounded by a 95% confidence interval which was calculated based on bootstrapping.

The DSBs per 1000 AAs values where choses for comparison since – like cysteine count – both values are unaffected by neither protein size nor proteome size. The correlation based on the 12 model organisms analysed in this chapter result in a relatively good correlation with a coefficient of determination (R^2) of 0.88. As seen in previous figures, the two organisms *A. niger* and *A. thaliana* act as a sort of transition between two distinct sets of organisms. The two most notable outliers in the correlation plot are *A. thaliana* and *C. elegans*, the first having comparatively few DSBs for its cysteine usage while the latter has comparatively many. Regardless, the overall correlation between the two values is good and might be improves by adding a more diverse set of organisms to the analysis.

Organism Complexity

The increase in average protein size from 279 for *B. subtilis* to 557 for *H. sapiens* is outweighed by the increased usage of DSBs in the more complex organisms. The trends observed in the two figures (Figure 44 and Figure 46), place the fungi organisms as an intermediate in DSB utilisation between bacteria and the other eukaryotes. The DSB data suggest a strong change in DSB utilisation between single-cell organisms and multi-cell organisms and with an organism's complexity in general. *A. niger* is right on the edge between single-cellular and multi-cellular organisms but just like the multi-cellular plant *A. thaliana* they are both not part of the animalia kingdom [203]. As such they do not exhibit the same cell types and tissues as the remaining 5 multi-cellular model organisms. This aligns well with their DSB utilisation depicted in Figure 45 and Figure 46, where the two organisms create a transition in DSB usage between the ‘simple’ and the ‘complex’ organisms.

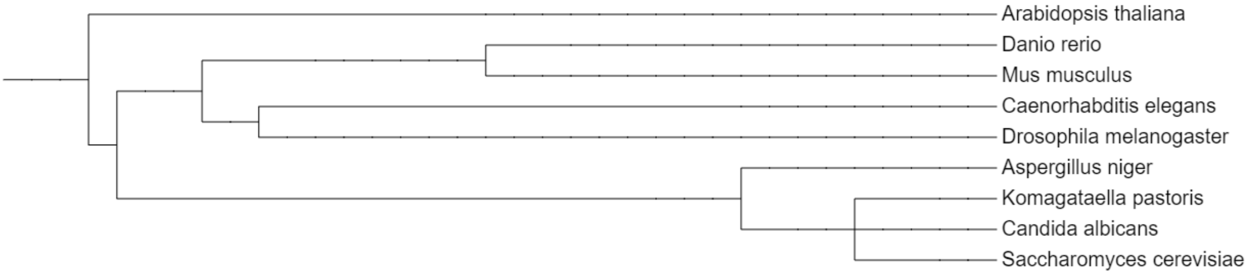


Figure 48 Phylogenetic tree for 9 model eukaryotes (H. sapiens and M. musculus group together). The tree displays the evolutionary relationships and distances between the organisms.

The classification of organism complexity is an active area of research which tackles a question which has no simple answer [204]. The phylogenetic tree for 9 of the 10 model eukaryotes is displayed in Figure 48 for reference. In general, single-celled organisms are mostly ‘simpler’ than multi-cellular organisms because the multi-celled nature dictates a larger degree of specialisation. However, within the multi-cellular organisms the complexity ranking becomes more difficult. Is a plant less complex than an animal? Is a fish less complex than a mouse? Common metrics for ranking organismal complexity are cellularity (e.g. single or multicellular), morphology (e.g. tissues and organs), developmental complexity (e.g. lifecycles), neurological complexity (e.g. neuron count), behavioural complexity (e.g. communication and problem solving) and genomic complexity (e.g. gene count) [205], [206], [207], [208]. Looking at the total gene count for the 12 model organisms discussed in this chapter (Table 5), we can observe that the predicted trends occur generally, although individual organisms can strongly deviate from the general trend" or similar.

The plant *A. thaliana* and the nematode *C. elegans* have similar gene numbers (44112 and 50989 respectively). However, *C. elegans* has twice the amount of DSBs (Table 5) compared to the plant. For mouse and human on the other hand, the two metrics seem to align well with each other. In the previous work by Miseta and Csutora, they showed how cysteine count can be used to infer an organism’s complexity [209]. However, at the time of their work (2000), widespread information of DSB counts were not available. Now, AF structure predictions and the method developed in this chapter could be used to calculate the DSB count of many more organisms in addition to the 12 discussed here. This information could be used to expand cysteine count method developed by Miseta and Csutora towards considering DSBs counts as well. It remains to be seen; however, if this would change or improve the ranking.

Disulfide bond complexity composition

More complex organisms have on average more complicated disulfide bonds. There are many different protein folding aspects that determine how difficult it is for a disulfide bond to be formed correctly, however; one of the most important factors is whether or not the DSB is formed between consecutive cysteines in the protein's amino acid sequence. Figure 49 shows this trend by displaying the ratios between consecutive and non-consecutive DSBs for the AF predicted disulfide bonds.

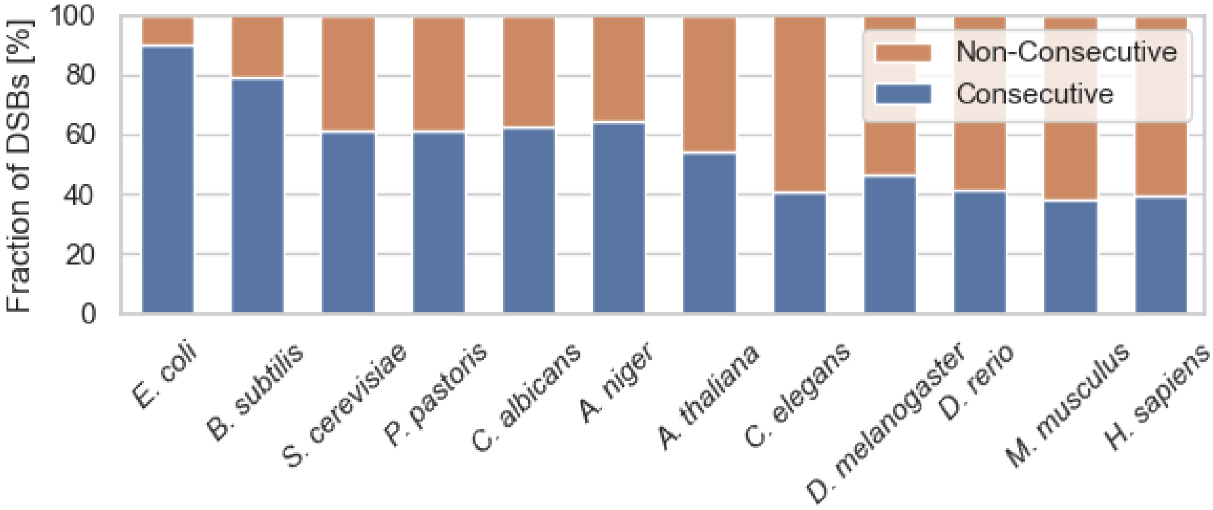


Figure 49 Ratio comparison between consecutive and non-consecutive DSBs as predicted by AF for the 12 model organisms.

As Figure 49 suggests, the average DSB complexity increases with the complexity of the respective organism. Particularly the two bacteria *E. coli* and *B. subtilis* exhibit a comparatively simple DSB composition. This is further demonstrated in Figure 50 which shows the share of DSBs which are formed in proteins with only two cysteines present. This constitutes the simplest possible setup for a DSB to form successfully during oxidative folding (see also Chapter 1 for a more thorough discussion of *E. coli* DSB folding complexity). These types of DSBs make up a significant part of the DSB proteome of the two bacteria with roughly 20% and 25% of the total AF predicted DSBs for *E. coli* and *B. subtilis* falling into that category respectively. The 4 fungi model organisms already show a sharp decrease in the fraction for these simple DSBs with only around 2-4% and the remaining 6 more complex model organisms all have below 2% of DSBs that fit this category. As previously described in the thesis introduction, the oxidative folding of substrates in *E. coli* is performed by DsbA which is a particularly strong oxidase. This makes the enzyme very efficient at forming DSBs in substrates but also more prone to errors, particularly when the cysteines are far apart in the amino acid sequence and the protein needs more time to collapse and bring the correct cysteine into special proximity. However, DSBs which are formed between the only two cysteines in a protein eliminate this possibility and are therefore ideally suited for oxidative folding via DsbA. The high fraction of these simplest DSBs in *B. subtilis* would suggest that the oxidative folding machinery in *B. subtilis* functions in a similar way to *E. coli* in this regard.

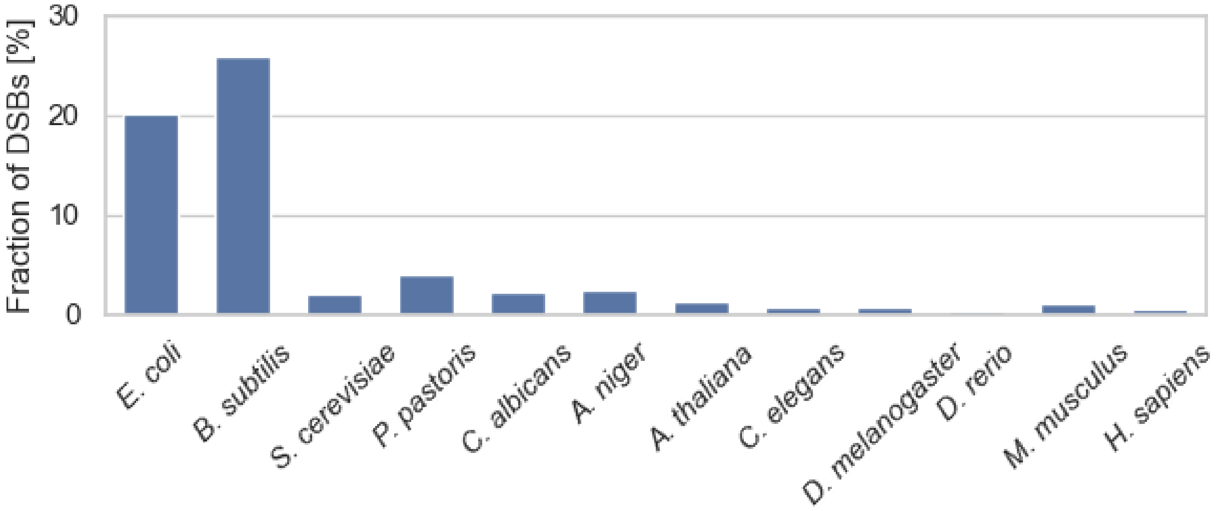


Figure 50 Share of AF predicted DSB which are formed between the only two cysteines on a given protein.

The quantitative DSB Proteome

By identifying the AF predicted disulfide bonds in each model organism's proteome, a qualitative overview of the oxidative folding demand can be drawn. However, by including cellular protein abundance into the calculation, the quantitative aspects of the systems can be investigated. For kinetic modelling as utilized in Chapter 1, absolute protein abundance proteomes are required. These proteomes are still relatively rare to find in literature, while relative quantitative proteomes are much more commonly measured and published. Furthermore, even though this Chapter investigates *model* organisms, which have more information available to them compared to other organisms, finding high protein coverage quantitative proteomes is difficult for some of the organisms investigated here.

In their 2012 paper von Wang et al. introduced their protein abundance database PaxDB [140]. In this database they collect published quantitative proteomes from all domains of life. Whenever there is more than a single quantitative proteome available for an organism, they calculate an integrated proteome that combines different proteomes into a single consensus proteome. At the time of writing this thesis (early 2023) quantitative proteomes were available for 11 of the 12 model organisms discussed in this Chapter, with *P. pastoris* being the only organisms left out. 9 out of the remaining 11 model organism have an integrated proteome based on at least 2 published quantitative proteomes available. For *C. albicans* and *A. niger* only a single quantitative proteome is available. The *C. albicans* quantitative proteome has a relatively good protein coverage with roughly 4000 proteins measured while the *A. niger* proteome only covers 800 proteins. In Figure 51 the proteome coverage estimations based on PaxDBs own estimation, based on the AF dataset and based on the reference proteome size are displayed and compared.

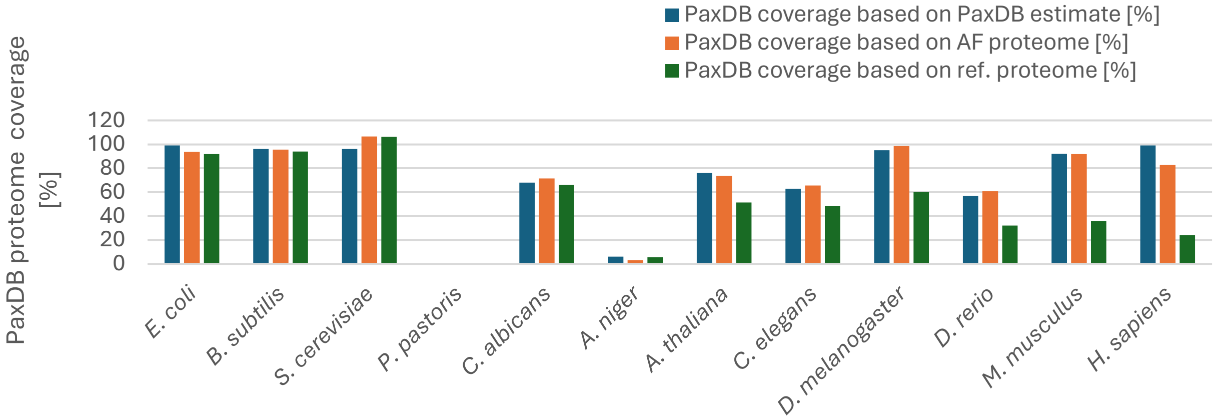


Figure 51 Comparison between proteome coverage measurements. In blue (left bars) the proteome coverage as reported by the PaxDB website. In orange (middle bars) the proteome coverage based on merging the PaxDB entries with the AF proteome entries. In green (right bars) the proteome coverage based on merging the PaxDB with the previously used reference proteomes (source UniProt). The above 100% coverage observed for S. cerevisiae is a result of protein nomenclature discrepancies between UniProt and PaxDB, where more than one UniProt entry is linked to the same PaxDB entry.

As organisms become more complex going from the left to the right in Figure 51, the consent between the three different estimates lessens. Particularly for the reference proteome based estimated. This is because, as a proteome becomes more complex, an increasing number of proteins have different isoforms and alternative splicing patterns which amplifies the variations of proteins derived from a single gene. The overlap between the PaxDB and the AF based coverage estimation on the other hand is better.

It should be noted that the proteome composition of an organism can vary significantly between different growth and stress conditions as well as different tissues. The integration of different quantitative proteomes into a single proteome is therefore connected to a loss of information and creates a proteome which might not actually be found in-vivo. Nevertheless, these integrated proteomes are ideally suited for gaining an overview of the quantitative oxidative folding requirement of an organism and for comparing this information between different organisms.

The final step combining the PaxDB quantitative proteome information with the AF predicted DSB information is the merging of the datasets based on the proteins accession numbers. Heterogeneity in the nomenclature of proteins is an unfortunate complication in this task. Figure 52 displays the amount of AF predicted DSBs which were successfully linked to a PaxDB entry.

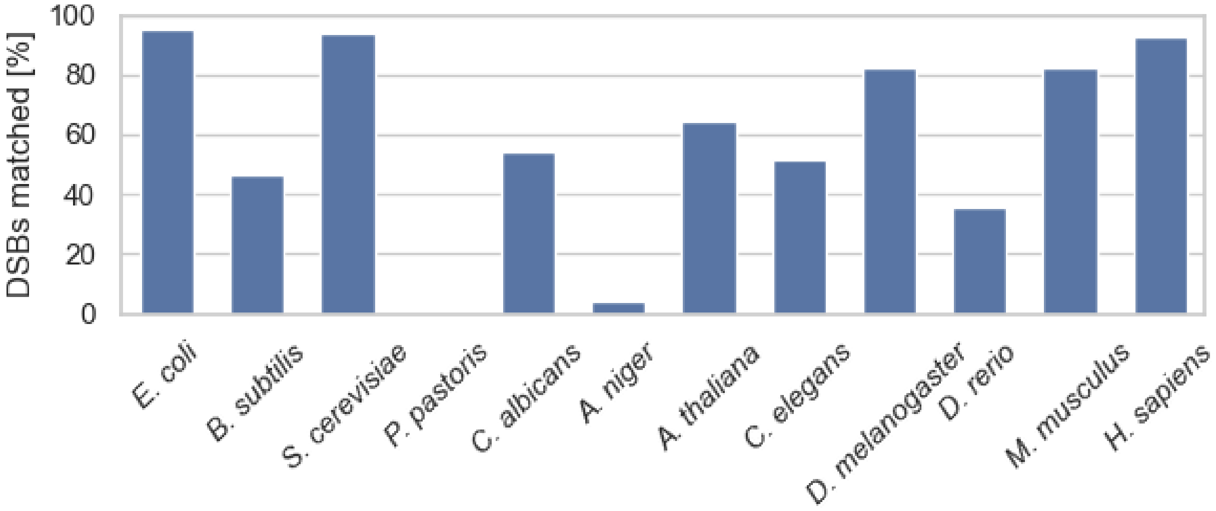


Figure 52 Fraction of DSBs which were successfully matched between AF and the PaxDB proteomes.

The figure shows that the matching between the data sources is not always ideal. Differences in nomenclature and proteome composition (i.e. which proteins to include and which proteins to count) result in roughly half of the investigated organisms having less than 80% of their proteins matched. However, the more abundant proteins (which are on average more studied and therefore more commonly have harmonised nomenclature) are disproportionally more often successfully matched. Figure 53 shows the amount of PaxDB proteins that are matched to an AF structure in ppm. A complete match should result in 1 million, i.e. all parts of the parts-per-million quantification.

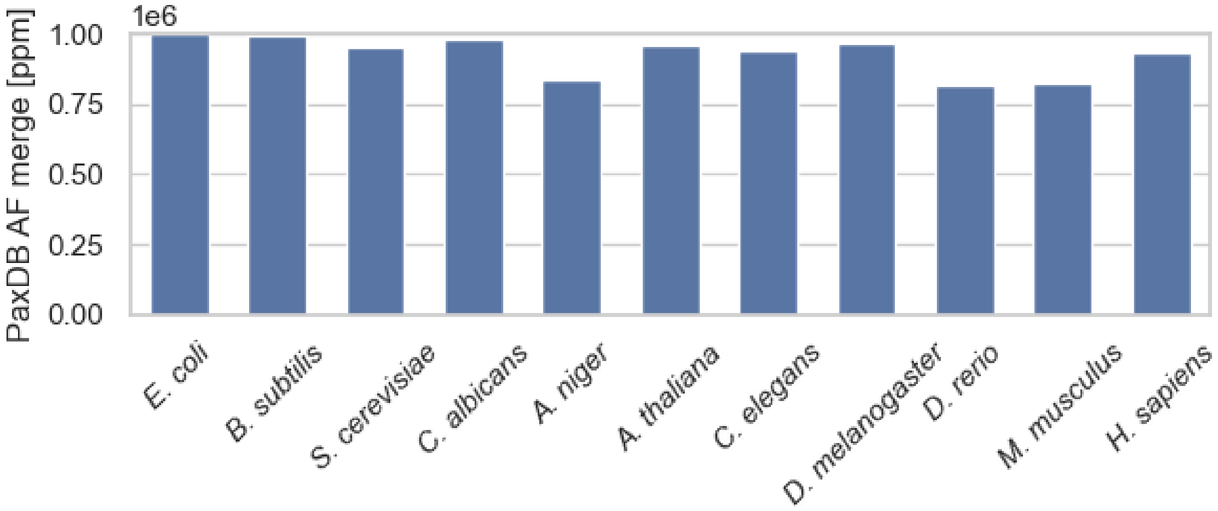


Figure 53 Quantitative result of the AF DSB and PaxDB proteome merge. Numbers close to 1000000 ppm indicate a great quantitative merge between the two data sets.

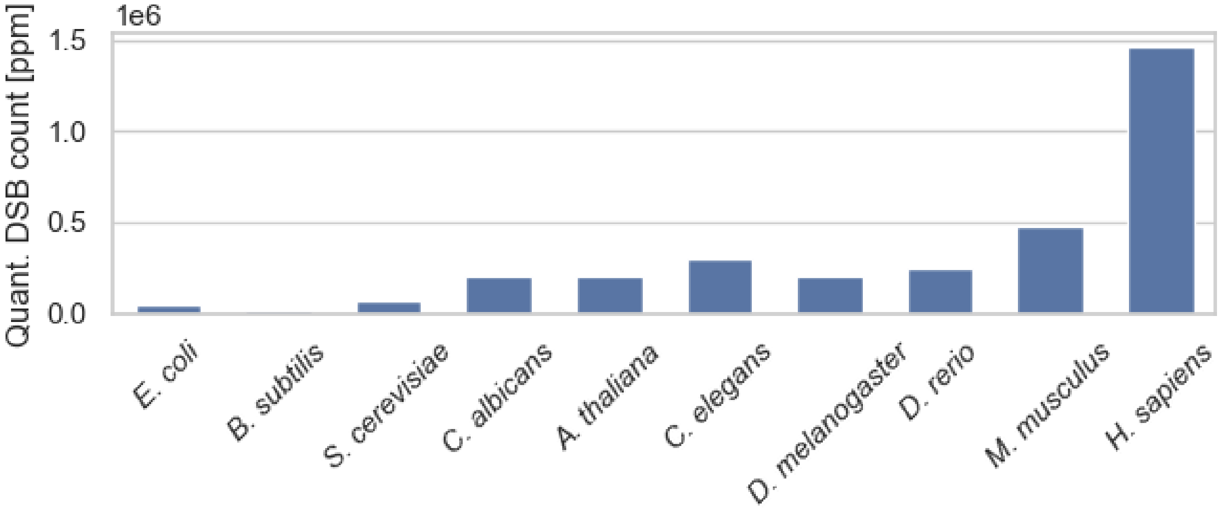


Figure 54 Quantitative estimation of DSBs in 10 model organisms based on PaxDB and AF

Once each organism's DSBs have been matched with the abundance of each respective proteins, the predicted amount of DSBs in the organism's average proteome can be calculated. This data is shown in Figure 54 for model organisms excluding *P. pastoris* (no quantitative proteome available) and *A. niger* (too little coverage in the only quantitative proteome available). The figure shows how extreme the DSB utilisation in humans is. In the above analysis *H. sapiens* is always relatively close to *M. musculus*, however, when looking at the bar for Figure 54 *H. sapiens*, it is roughly 3 times larger.

As mentioned previously, bacteria have long been known to utilise far less DSBs compared to eukaryotes, particularly mammals. Figure 54 strongly emphasises how huge the difference between *E. coli* and *H. sapiens* is. The latter's proteome has more than 30 times the amount of DSBs compared to the former.

In a consistent manner to the previous analysis, *B. subtilis* is by far the least DSB-utilizing organism of the selected model organisms. The organism is widely used for protein production, particularly enzymes such as proteases and amylases. Attempts to produce more complex, biopharmaceutically relevant proteins, have been made but have so far been unsuccessful in establishing this organism as a competitive alternative to other, more complex, organisms. The results on the organisms DSB-utilisation suggest that the high-titre production of heterologous proteins with (complex) disulfides would require a sharp change to its native growth behaviour in regards to oxidative folding.

In a similar fashion to *B. subtilis*, *S. cerevisiae* has been widely used for protein production but has so far struggled to establish itself as a common host in biopharmaceutical protein productions. Even though, *S. cerevisiae* is significantly better at producing complex disulfide bonds compared to *E. coli* and *B. subtilis* (Figure 49) the quantitative DSB utilisation suggest that it does not require particularly many of them. *P. pastoris* is one of the most promising alternatives to *S. cerevisiae* when it comes to yeast-based protein production. Unfortunately, the lack of quantitative proteome data for this organism makes it currently impossible to compare the two organisms in terms of quantitative DSB usage. What can be said is that in terms of qualitative DSB complexity the two organisms are similar. The third yeast investigated here, *C. albicans*, has a roughly two-fold higher DSB-utilisation which, combined with a similar proclivity of complex DSBs as the other two yeasts, might make this yeast a better host for high titre production of disulfide bonded proteins. Although this is being complicated by the fact that *C. albicans* is an opportunistic pathogen to humans.

The six non-microbial eukaryotic organisms all have a higher DSB complexity utilisation than the six microbial organisms. In regard to their quantitative DSB utilisation, *A. thaliana, C. elegans, D. melanogaster* as well as *D. rerio* show similar behaviour to *C. albicans* - the highest of the microbials tested. The following analysis of the DSB – enzyme type correlation will show some difference between these organisms but at least in regard to their overall DSB utilisation, they are relatively similar.

The most demanding in terms of both DSB complexity and DSB utilisation are the two mammals, *M. musculus* and *H. sapiens*. Both show a similar proclivity for complex disulfide bonds, however, *H. sapiens* has a significant higher quantitative amount of DSBs (roughly 3 times more).

The Enzyme Commission Number

In the above paragraphs the proteome wide DSB data is used to investigate the DSB composition of the model organisms as well as the oxidative folding demand of their proteomes. However, this new, more complete, knowledge about the DSB composition of a proteome can be used to investigate the oxidative folding demand and composition of other cellular or enzymatic aspects.

Enzymes provide an immensely wide range of different functions for an organism, with almost every aspect of life dominated by their activities. Classifying this diverse range of functionality into categories is not trivial. However, one of the most commonly used enzyme classifications is the

Enzyme Commission Number (EC number) which groups enzymes based on the chemical reactions that they catalyse. In this convention, enzymes are given four numbers separated by a dot and starting with 'EC', with the numbers describing a hierarchical classification and the four numbers describing progressively finer aspects of enzyme activities. The first number describes the highest enzymatic hierarchy and has 7 different classes numbered 1 to 7. These 7 EC classes are listed and briefly described in Table 4.

Table 4 Enzyme classification numbers and corresponding functions.

EC Number Classification		**Description**
1	Oxidoreductases	Enzymes that catalyse oxidation-reduction reactions
2	Transferases	Enzymes that transfer a functional group from one molecule to another
3	Hydrolases	Enzymes that catalyse the hydrolysis of various bonds
4	Lyases	Enzymes that catalyse the cleavage of C-C, C-O, C-N, and other bonds by means other than hydrolysis and oxidation
5	Isomerases	Enzymes that catalyse isomerization reactions
6	Ligases	Enzymes that catalyse the formation of a new bond via energy from ATP or a similar molecule
7	Translocases	Enzymes involved in the transport of molecules across cell membranes

The EC number classification only covers enzymes that catalyse a chemical function. Proteins involved in e.g. structural support, regulation or signal transduction, can fall outside the EC classification system and therefore do not have an EC number. Also, many proteins have unknown functions and have therefore not yet been given an EC number. This is particularly true when looking at whole proteomes which often cover many proteins which have not been intensively researched and may only have putative functions assigned to them based on structural similarities. As such, the EC system does not cover the whole proteomes investigated in this chapter, however, it still provides a useful classification of enzymes into fundamental enzymatic functionalities. The number of proteins in each proteome that has an EC number assigned to them is displayed in Figure 55. The numbers show that the total amount of annotated enzymes varies strongly between the organisms with *A. thaliana* having the highest number of annotated proteins. In plant biology, metabolic pathways (e.g. for secondary metabolites) are of central interest and EC numbers are particularly useful in classifying the roles of various enzymes in these pathways. The overall variation in EC number annotation has to be kept in mind during the following analysis of the DSB usage in different organisms and EC numbers. Figure 56 shows the annotation composition of the 7 EC categories. It shows that despite the different total annotation counts between the organisms, there seems to be relatively little variation between the different EC categories. Going from left to right (less complex to more complex) the fraction of oxidoreductases (EC 1) and lyases (EC 4) decreases while the transferases (EC 3) and hydrolases (EC 2) become more common.

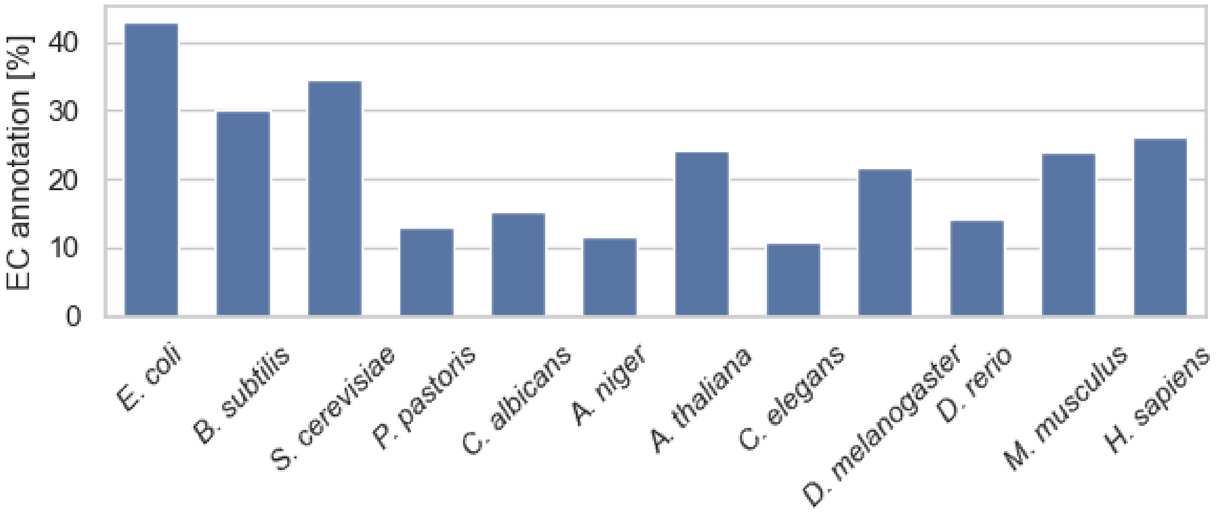


Figure 55 EC annotation coverage as a fraction of the total proteome for all 12 model organisms.

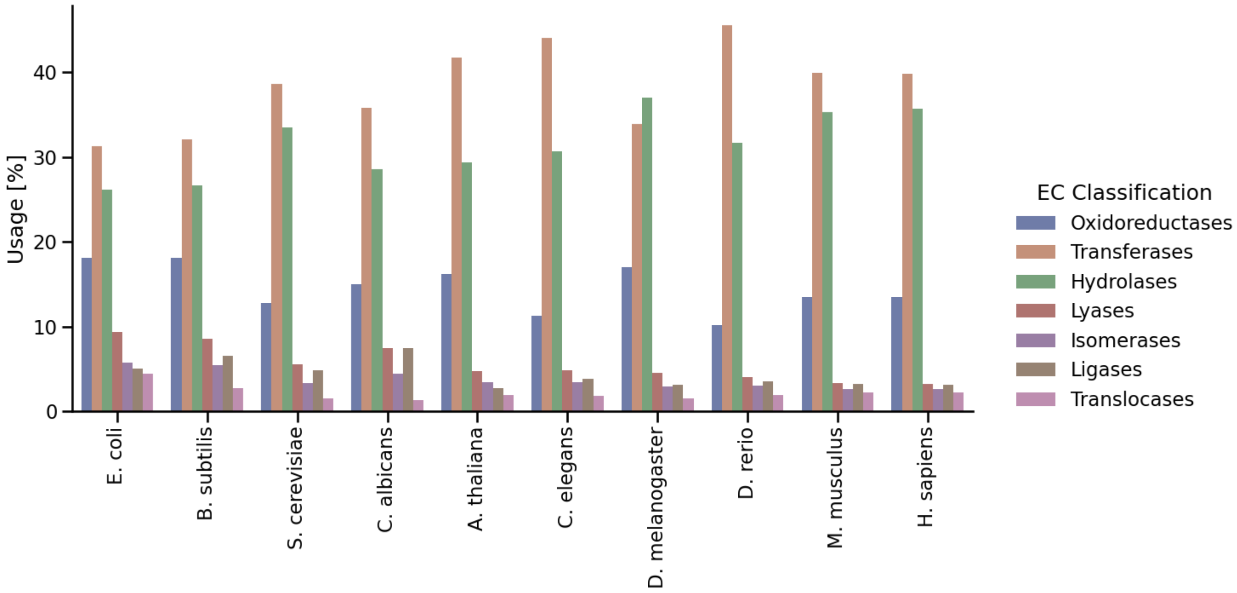


Figure 56 Fraction of the EC classification categories as percentage of all EC annotated proteins for 10 model organisms.

DSB utilisation in different Enzyme Classes

In the following analysis the prevalence for DSBs in the 7 EC categories is investigated. Once the overall EC number annotations in the 10 proteomes (excluding *P. pastoris* and *A. niger*) have been extracted, the 7 different EC classes are multiplied by each proteins respective DSB count. This provides the qualitative description which depicts how many enzymes in each category have DSBs. Secondly, based on the PaxDB data used in the previous analysis, the qualitative DSBs counts for each EC category are multiplied by the respective protein abundances, providing insight into the oxidative folding requirement of each EC group. Figure 57 displays the qualitative values and Figure 58 displays the quantitative values for all 7 EC categories. The numbers have been converted to % usage for better comparability between the numbers and the organisms.

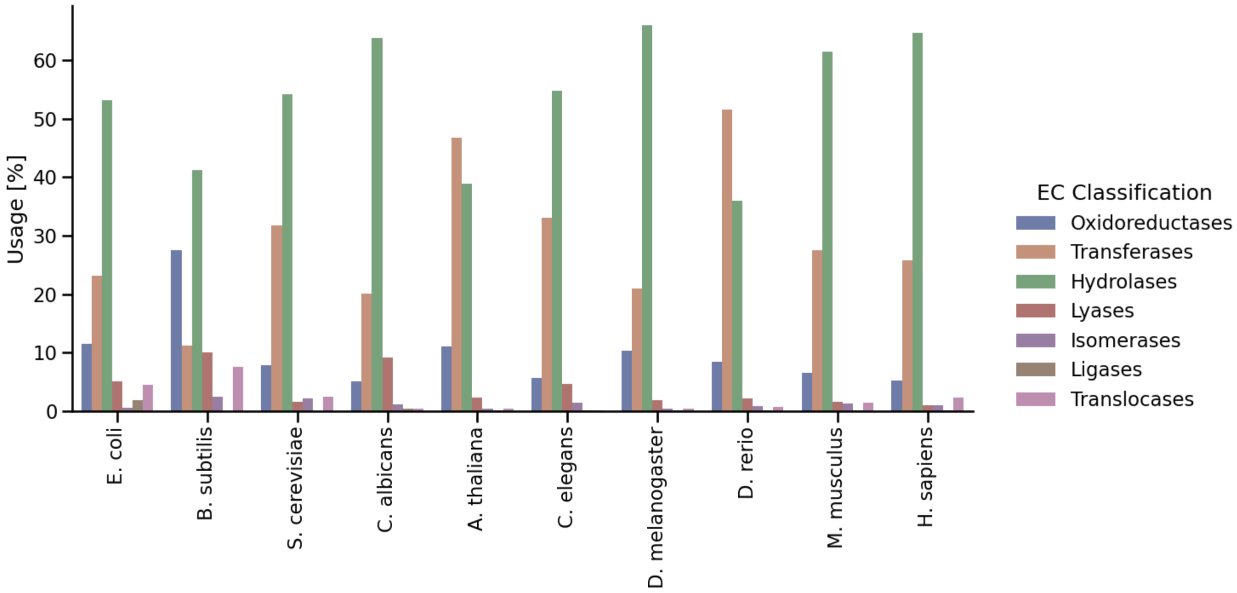


Figure 57 Qualitative DSB usage in each EC category for 10 model organisms.

Compared to the annotation count in Figure 55, the fraction of transferases and hydrolases becomes even more pronounced. Only *B. subtilis* shows a different behaviour with a higher utilisation of DSBs in the oxidoreductase category. This might be in part derived from an overall low DSB count in *B. subtilis* which makes the value more susceptible to variations based on individual proteins. The remaining 9 organisms show a reduction in oxidoreductase fraction compared to the annotation fractions. For all organisms, DSB usage is highest in hydrolases expect in *A. thaliana* and *D. rerio* where the transferases are higher. This is a clear change compared to the annotation ratios in Figure 55, where the transferases are the top category in 9 out of 10 of the organisms (*D. melanogaster* being the exception).

When looking at Figure 58 and the quantitative DSB usage in the EC categories, some of the trends observed above continue. In most organisms the hydrolases fraction has become even more dominant when protein abundance is being taken into consideration. The two previous outliers *A. thaliana* and *D. rerio* now have some of the highest hydrolases fractions with the transferase fractions now almost inconsequential towards the cells oxidative folding demand. Also, the strong DSB utilisation observed in *B. subtilis* oxidoreductases has further increased with the introduction of the quantitative perspective. And now also *E. coli* has a very substantial oxidoreductase fraction compared to before and compared to the non-bacterial organisms.

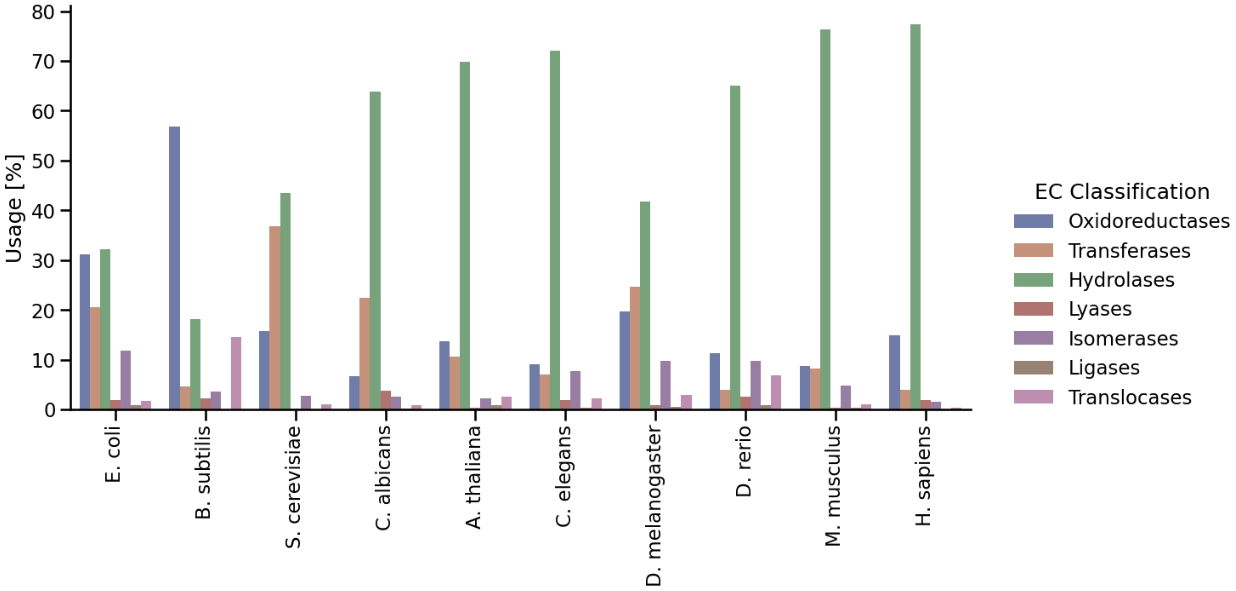


Figure 58 Quantitative DSB usage in in each EC category for 10 model organisms.

Table 5 Summary table of the key data presented in this chapter. The first of the two taxonomic IDs listed for E. coli *and* B. subtilis *are the IDs for each respective organism however, for this study a specific strain had to be selected for each organism which was K12-MG1655 and 168 for* E. coli *and* B. subtilis *respectively. The strains IDs are listed second. The PaxDB entry for C. albicans and A. niger were not based on integrated Proteomes and the A. niger proteome number is listed here but was not used in analysis.*

Organism	***E. coli***	***B. subtilis***	***S. cerevisiae***	***P. pastoris***	***C. albicans***	***A. niger***	***A. thaliana***	***C. elegans***	***D. melanogaster***	***D. rerio***	***M. musculus***	***H. sapiens***
Taxonomic ID	562/11145	1423/224308	4932	4922	5476	5061	3702	6239	7227	7955	10090	9606
Reference proteome	UP000000625	UP000001570	UP000002311	UP000094565	UP000000559	UP000006706	UP000006548	UP000001940	UP000000803	UP000000437	UP000000589	UP000005640
Genes (NCBI Entrez)	4661	4871	7074	5325	15217	28205	44112	50989	29036	100231	212434	307198
Average protein length in Proteome	309	279	480	437	443	459	402	397	495	513	505	557
DSBs annotated in Uniprot	123	61	211	20	156	215	3691	3442	4960	7990	19114	21588
Total cysteine count in proteome	15802	9719	37288	29134	31830	87253	308820	244583	292306	727865	525889	632321
Cysteine usage [%]	1.16	0.79	1.27	1.17	1.07	1.41	1.86	1.99	1.94	2.12	2.26	2.18
AF structures	4363	4183	6039	4914	5577	25325	27434	19694	13458	24664	21615	20504
AF DSB prediction	504	302	1274	940	1320	8701	15281	25284	14971	33694	30203	30501
AF DSB per Structure	0.12	0.07	0.21	0.19	0.24	0.34	0.56	1.28	1.11	1.37	1.4	1.49
AF DSBs per structure	0.12	0.07	0.21	0.19	0.24	0.34	0.56	1.28	1.11	1.37	1.4	1.49
AF DSBs per 1000 amino acids	0.37	0.13	0.44	0.44	0.53	0.75	1.38	3.24	2.25	2.66	2.77	2.67
AF consecutive DSBs	455	240	778	578	826	5616	8252	10371	6929	14033	11606	12047
AF non-consecutive DSBs	49	62	496	362	494	3085	7028	14913	8041	19661	18596	18521
AF % non-consecutive [%]	9.72	20.53	38.93	38.51	37.42	35.46	45.99	58.98	53.71	58.38	61.57	60.3
AF % consecutive [%]	90.28	79.47	61.07	61.49	62.58	64.54	54.01	41.02	46.29	41.65	38.43	39.41
AF only 2 cysteines in DSB	101	78	26	37	29	209	188	167	91	92	322	151
AF only 2 cysteines in DSB [%]	20.04	25.83	2.04	3.94	2.2	2.4	1.23	0.66	0.61	0.27	1.07	0.5
PaxDB entries	4090	4001	6440	-	3991	800	20183	12921	13264	14998	19871	19338
PaxDB coverage based on PaxDB estimate [%]	99	96	96	-	68	-	76	63	95	57	92	99
PaxDB coverage based on AF proteome [%]	93.7	95.6	106.6	-	71.6	-	73.6	65.6	98.6	60.8	91.9	94.3
PaxDB coverage based on ref. proteome [%]	92	93.9	106.3	-	66.1	-	51.3	48.3	60.1	32.1	35.9	24
PaxDB entries mapped to Uniprot ID	4046	3937	5879	-	3559	-	17720	10438	11631	13815	18170	18735
AF & PaxDB matched [ppm]	999978	997403	952056	-	982117	-	959161	936968	968481	816299	821887	935395
AF & PaxDB ppm-matched [%]	100	99.7	95.2	-	98.2	-	95.9	93.7	96.8	81.6	82.2	93.5
AF & PaxDB DSBs matched [%]	95.2	46.7	93.9	-	54.4	-	64.7	51.9	82.3	35.7	82.4	92.6
AF & PaxDB consec. DSBs in proteome [ppm]	48944	14853	44041	-	137210	-	136679	160012	112192	137201	176267	399996
AF & PaxDB non-consec. DSBs in proteome [ppm]	2556	1697	26081	-	73423	-	65064	136204	95006	106022	297381	1065172
AF & PaxDB non-consec. quantity [%]	5	10.3	37.2	-	34.9	-	32.3	46	45.9	43.6	62.8	72.8
EC total count	1881	2553	2085	726	752	2972	6640	2131	2921	3484	5165	5350
EC number coverage [%]	43.1	30	34.5	13	15.3	11.7	24.2	10.8	21.7	14.1	23.9	26.2

Discussion

The initial aim of my PhD project was to computationally investigate the oxidative folding machinery of *E. coli*. A similar approach had previously been published for *S. cerevisiae* by Beal et al. based on enzyme kinetic characterisation experiments [155]. Here, we investigate oxidate folding in *E. coli*, using a much more extensive and detailed set of proteomic datasets than was available for the yeast work. This was made possible by the availability of several high-quality large-scale data sets and publications available for *E. coli*. These datasets made it possible to model the oxidative folding machinery in *E. coli* without the need for additional experimental laboratory-based work.

An essential and obligatory part of every ITN grant (i.e. SECRETERS) are secondments. When the travel restrictions imposed by the pandemic were lifted, doing the first secondment of my project became possible. This new project would be a continuation of my research into oxidative folding pathways in microorganisms but would also bring a shift in target organism as well as a methodological shift from computational to laboratory-based work. The yeast *P. pastoris* is a commonly used protein production organism but the scientific understanding of its oxidative folding capacities had so far remained largely unexplored [210]. On my secondment in Oulu University I performed protein expression and purification of key oxidative folding enzymes (PDI) from both *H. sapiens* and *P. pastoris*, together with a suitable and well-established substrate protein BPTI. These three proteins, together with a PDI family member from *P. pastoris* (ERp38), were used to characterise both oxidative folding and isomerisation capabilities of these key enzymes from the *P. pastoris* protein folding machinery [129].

The results from the first secondment in Oulu were promising and in order to finalise them, a second secondment was started following the first. During this second secondment in Oulu University a research project was established to combine my computational knowledge with the oxidative folding expertise of both me and my two supervisors Tobias von der Haar and Lloyd Ruddock together with the newly released AlphaFold protein structures from 12 complete model organism proteomes. Together, these parts allowed me to investigate the DSB composition of 12 organisms in a novel and holistic approach.

These three research projects form the three Chapters of this thesis and will be discussed in more detail below. The final part of the discussion will aim to bring the overarching themes of this thesis together.

Quantitative oxidative folding in *E. coli*

A thorough discussion centred around the results presented in the publication together with an attempt at putting the findings of the DSB folding requirements in context with the corresponding growth and stress conditions can be found in the Discussion paragraph of the publication. The following discussion will instead focus on the data sources and ideas put in place for this analysis to become feasible.

Data collection for *E. coli* DSB-proteome estimation

Even though *E. coli* is one if not the most studied organisms in the world, there are still many aspects of its proteome, pathways and functions that we do not yet fully understand. One such aspect is oxidative folding. Thanks to recent improvements in sequencing and proteomics, the cellular composition of *E. coli* has become better understood than ever [211]. This is further highlighted by the availability of 74686 *E. coli* qualitative proteomes on UniProt (as of 11.09.2023) [131]. However, not only qualitative proteomes have been widely measured, many more quantitative proteomes have also become available [212], [213], [214], [215], [216], [217], [218]. Schmidt et al. provides a good example of a multi-levelled data source for *E. coli* grown under different growth and stress conditions combined with high protein coverage. Their thorough reporting of growth rates allows modelling approaches to add a time-variable into the calculations [218]. Another important piece of information required for modelling oxidative folding in *E. coli* are the concentrations of the oxidative folding enzymes: DsbA, DsbB, DsbC + DsbG and DsbD. Particularly DsbD - with the lowest abundance out of the five - was a limiting factor since not every quantitative proteome was able to report its quantity.

Knowing the number of proteins produced per minute is interesting but in order to model oxidative folding it is essential to know which of the over 4000 proteins in the *E. coli* proteome have DSBs. Fortunately three data sources provide proteome-wide information on DSB content. One is the UniProt database which hosts a wide collection of DSBs annotations for *E. coli* proteins and two data sources which are based on proteome wide DSB labelling techniques [219], [220]. While neither of these three data sources are complete, the overlap between all three provides a set of DSB annotations that can be used for modelling oxidative folding. However, not all DSBs are formed through oxidative folding. The cytoplasm of *E. coli* is a reducing environment and therefore unsuitable for forming stable structural DSBs in substrates. The oxidative folding environment for structural DSBs in *E. coli* is the periplasm. However, proteins can still form DSBs through functional activity in the cytoplasm. This in turn requires that DSB containing proteins must be sorted depending on cytoplasmic or periplasmic localisation. The UniProt database provides some annotation for the most important proteins, however the majority of proteins are not annotated for localisation. The approach developed by Loos et al. is, however, ideally suited for this task [221]. They trained a machine learning algorithm that can estimate protein localisation based on a proteins primary protein sequences. Since this approach covers the whole proteome, their dataset provides localisation information for all proteins with identified DSBs. The final piece of information missing is the volume of the reaction vessel, i.e. the periplasmic volume of the cells. Fortunately the data set provided by Schmidt et al. also provides measurements on the periplasmic size of the *E. coli* cells at the various growth conditions. However, for all other collected quantitative *E. coli* proteomes this information was not directly available. In order to estimate the periplasmic volume the corresponding cell sizes at different growth rates had to be estimated first. The data provided by Schmidt et al. was combined with the data published by Volkmer and Heinemann and allowed for the fitting of a linear function that can estimate the periplasmic size as a function of the (much more commonly reported) growth rate [218], [222].

Looking at this long list of data sources required for this specific modelling approach, it becomes clear why it has so far only been feasible for *E. coli*. No other organism has this large and holistic set of information available to them. Particularly when it comes to eukaryotic cells the oxidative folding machinery becomes a lot more complex. In *E. coli* we have a relatively good understanding of each oxidative folding enzymes tasks, for most other organisms this is not the case. Furthermore, protein localisation, oxidative folding compartment size (ER), oxidase and isomerase interactions, substrate specificity, quantification of less abundant enzymes and tissue specific variations all becomes increasingly complicated and unknown variables in eukaryotic organisms.

Nevertheless, as more and more high-quality data sources become available, approaches such as the one presented here in Chapter 1 will become increasingly feasible. Chapter 3 demonstrates how other sources such as predicted protein structures can be used to annotate DSBs in proteins. The PaxDB used in Chapter 3 demonstrates not only how many quantitative proteomes are already available for organisms, it also showcases how different proteomes can be combined into consensus proteomes, making modelling and comparisons easier [139]. Protein localisation prediction has been substantially improved and expanded thanks to the increasing availability of sequence data which has resulted in approaches such as the one by Loos et al. used in Chapter 1 or TOPCONS or SignalP [221], [223], [224]. And while substrate specificity of oxidative folding enzymes remains largely unknown, research such as the one presented in Chapter 2 and in Palma et al. helps by measuring oxidative and isomerisation activities of new enzymes [129].

Estimating enzyme kinetics based on folding requirements

The long list of data sources mentioned above was required to properly estimate the oxidative folding requirements of *E. coli* at the various growth conditions. Once this data was collected it was possible to calculate a theoretical 'stream' of unfolded substrate proteins that enter the periplasm. Furthermore, in order to simulate substrate folding complexity, this stream was split into three parts, depending on the approximated isomerisation requirements of the substrates. For DSBs formed between the only two cysteines in a protein, it was assumed that no isomerisation was required at all, only oxidation. For consecutive disulfide bonds in proteins with more than the two cysteines, it was estimated that 50% of the DSBs would require isomerisation. And for the few proteins with non-consecutive DSBs it was assumed that every DSB required at least one isomerisation step before reaching the correctly folded state.

Based on these assumptions and the known concentrations of folding enzymes it is possible to compute a minimal kinetic rate for each enzyme. These minimal rates correspond to how many reactions each enzyme must perform per second in order to avoid substrate accumulation. In other words, the cells grow under these conditions so their enzymes must be fast enough to fold their complete proteome within the course of the cells doubling time. Furthermore, it is possible to run these kinetic parameter estimations based on each of the growth conditions reported in quantitative proteomes. And since the resulting kinetic parameters estimations are all 'minimal' parameters, the largest will be closest to the real value. However, with this approach alone it is still not possible to estimate the maximal kinetic activity of the enzymes only their minimal-required speed. The best reason as to why this value should at least be close to the real value, is that cells are evolved to not be wasteful. Why would a cell have 100 folding catalysts if 50 could do the same job? Therefore, unless the enzymes have some other unknown functions that slow their oxidative folding capabilities down, this minimal estimated kinetic rate should be at least close to the real value. Furthermore, this approach yields a value that is almost impossible to measure in in-vitro experiments: Since all substrates are treated equally by the model (except based on their DSB pattern) the resulting kinetic parameters displays the enzymes' overall ability to fold the proteome as opposed to a single substrate. When describing the capabilities of a whole *E. coli* cell, these average-based kinetic parameters provide a better representation.

Characterising oxidative folding in *P. pastoris*

The methanotrophic yeast *P. pastoris* (often also referred to as *Komagataella phaffii*) belongs to the same order of the budding yeasts (Saccharomycetales) as *S. cerevisiae*. *P. pastoris* has become widely used in recombinant protein production due to the organisms' capacity to utilize methanol as carbon source as well as the associated strongly inducible *AOX1* promoter. In protein production, yeasts provide a similar cultivation simplicity as bacteria coupled with the advantages of eukaryotic cell PTM formation and compartmentalisation. However, in order to facilitate the predictability and characterisation of the oxidative folding capabilities of *P. pastoris* more research into the organisms oxidative folding catalysts was (and still is) required.

Comparing *P. pastoris* PDI to *H. sapiens* PDI

PDI is found in the ER of eukaryotes in high abundance and it was the first folding catalyst ever reported [225]. PDI has been linked with several cellular functions but its predominant task is the formation, breaking and rearranging of DSBs [94]. Both human and *P. pastoris* PDI (i.e. *PDI1*) have the same typical PDI structure with 4 domains, thioredoxin folds and the central -CGHC- motifs. However, while the catalytic activity of human PDI has already been well characterised before, *P. pastoris* PDI hasn't been studied in detail yet [94], [226]. To enable a first characterisation of *P. pastoris* PDI the two PDI proteins were produced in *E. coli* and purified for subsequent protein kinetic measurements and comparison. The well characterized 3 DSB containing substrate protein BPTI was produced in *E. coli* CyDisCo cells for soluble protein production followed by purification and substrate reduction resulting in purified and fully unfolded BPTI [172], [227]. Together with a second PDI family member from *P. pastoris* (ERp38 - not produced by me) these three folding catalysts were tested for their oxidative folding and isomerisation capabilities in refolding assays with unfolded BPTI. What makes BPTI a good choice for such an assay is its well characterised folding pathways. The first two oxidation steps from zero to two DSBs are predominantly reliant on a catalyst's oxidation capabilities and can also be achieved by the glutathione buffer alone. The two oxidation steps are then followed by an isomerisation step before the third and final DSB can be formed. This last step requires an isomerisation catalyst such as PDI. Fitting of a time-resolved refolding assay allows for the calculation of kinetic parameters for three steps, the two early oxidation steps and the isomerisation steps. The final oxidation step cannot be detected since the two DSB containing BPTI species (isomerised and not isomerised) cannot be differentiated in the mass-spectrometric analysis. Furthermore, the isomerisation step is significantly slower than the final oxidation step and therefore commonly regarded as a single isomerisation (plus oxidation) step.

The comparison between 7 µM hPDI and pPDI revealed that both enzymes take the whole duration of the assay time (2h) for fully folding all available BPTI. However, their specific strategies for reaching these fully folded states differ. pPDI displays the faster oxidation kinetics in steps one and two (1.76; 2.77 and 1.06 s^{-1} $µM^{-1}$) compared to hPDI (0.95; 1.25 and 0.54 s^{-1} $µM^{-1}$). BPTI isomerisation has two distinct pathways, one for either of the two dominant 2S species (N' and N*) [173]. Isomerising N* takes longer to isomerise than N' does. Fitting the double exponential curve into the assay results for the 2S to 3S isomerisation step results in two kinetic parameters for the two pathways as well as two percentages describing how many species pass through each pathways. For pPDI the fitted kinetic rates are 0.0038 and 0.048 s^{-1} $µM^{-1}$ as well as 41 and 59% respectively. For hPDI the results are 0.0031 and 0.040 s^{-1} $µM^{-1}$ as well as 60 and 40% respectively. Since the two 2S intermediates (N' and N*) cannot be differentiated via their masses, we cannot prove that the slow isomerisation step known for BPTI (i.e. N*) is also the slow step observed here, however, it seems likely [228]. If we make this assumption, it points towards a different oxidative folding strategy for pPDI compared to hPDI. pPDI is faster at oxidising compared to hPDI but it comes at the cost of creating more of the kinetically unfavoured N* species (59% compared to 40% with hPDI). Consequently hPDI is able to isomerise the more abundant N' species and achieves a higher 3S concentration faster. During the cause of the 2h isomerisation time, pPDI eventually catches up to hPDI as a result of being the faster isomerase for both 2S subspecies. These results suggest different oxidative folding strategies for *H. sapiens* and *P. pastoris*. The latter being more 'fast and loose' while the former employs a more controlled oxidative folding approach. In practice this might suggest that P. pastoris prefers to introduce DSBs more quickly, maybe to avoid accumulation of unfolded proteins in e.g. inclusion bodies. On the other hand, *H. sapiens* prefers to oxidise DSBs more carefully and avoid accumulation of partly (oxidatively) folded substrates. The limitations to these assumptions come primarily from BPTI being a non-native protein for both enzymes as well as the fact that it is only a single tested substrate. The specific interactions between the native substrates of both enzymes might exhibit a different folding characteristics and speeds. Nevertheless, these results could be indicative of the overall folding strategies of *P. pastoris* compared to *H. sapiens* and might be beneficial in research as well as future recombinant protein production attempts for producing human proteins in this yeast.

ERp38 – super oxidase

The second *P. pastoris* enzyme tested for its oxidative folding capabilities was ERp38. This PDI family member does not have the typical -CGHC- thioredoxin motifs that both hPDI and pPDI have, it has -CSHC- and CGYC- instead [166]. Its oxidative folding capabilities have never been described before and a more detailed description of it will hopefully soon be published in a manuscript currently being prepared by Arianna Palma and co-authored by myself. The results presented in Chapter 2 of this thesis describe ERp38 as a particularly strong oxidase and a weak isomerase. The measured rate constants for oxidation step 1 (2.86 s^{-1} μM^{-1}) and step 2 (3.80 & 1.18 s^{-1} μM^{-1}) are significantly faster than the ones observed for the two PDI1s discussed above. The refolding assay data even suggest that ERp38 is so ‘fast and loose’ in its oxidation catalysis that it misfolds BPTI to a non-native 3S state which is subsequently reduced back to earlier oxidative folding states. It would be interesting for future research to investigate the specific misfolding pathway ERp38 and BPTI describe as well as the specific tasks of such a strong oxidase in ER protein folding. The continuous expression of such a strong oxidase might result in more misfolded protein accumulation while the induced expression of ERp38 as a stress response might be beneficial to the cell.

Quantifying the DSB-proteome in model organisms

Protein structures can be used for identifying and annotating DSBs in proteins. However, the process obtaining protein structures is difficult and time consuming [229]. On the other hand, the ‘Anfinsen dogma’ dogma tells us that all the required information for a protein to achieve its native fold is already given in the primary sequence of the protein [225]. Protein structure prediction has therefore been an important field of research for decades. Many noteworthy prediction models have been created over the years; however, the accuracy achieved by AlphaFold 2.0 combined with the immense computational power available to Alphabet (i.e. Google) has made it possible to (reasonably) accurately predict the protein structures of whole proteomes for the first time [148]. In Chapter 3 the structure predictions for 12 model organism proteomes were used to predict many more DSBs than previously identified or annotated in literature or databanks.

How different organisms utilise DSBs

Of the 12 model organisms analysed in Chapter 3, ranging from bacteria, to fungi, plants and animals, all have DSBs. However, their utilisation of this extra covalent protein bond is very different. Bacteria have comparatively few DSBs in their proteomes with *B. subtilis* having the least among the tested organisms (~300) while animals have approximately 100-times more (~30000). Not only the amount of unique DSBs present in the respective organisms are different, the complexity of their DSB bonding patterns are different as well. Out of the detected DSBs in *E. coli* roughly 10% are formed between non-consecutive cysteines while on the other side of the spectrum, roughly 60% of the human DSBs are non-consecutive. And yet another aspect that is different is the quantitative aspect. Roughly 5% of all proteins in *E. coli* cells have DSBs while the average human protein has roughly 1.5 DSBs (i.e. 150%). The analysis in Chapter 3 emphasises the wide range of DSB utilisation between different organisms and phyla.

DSB content has previously been linked as a possible metric for ranking organism complexity [209]. The analysis presented in Chapter 3 could be used to include DSB-counts per organism into such an analysis which might increase prediction quality compared to cysteine counting methods alone [230]. Also, as mentioned above, the newly availability of large-scale distinction between consecutive and non-consecutive DSBs could increase the predictive quality of the DSB metric in organism complexity analysis even further.

Combining different facets of oxidative folding research

Oxidative folding is a central element of protein folding and function. The correct folding of native proteins is an essential part of their function and mistakes can lead to protein accumulation, degradation or in the worst cases cell death or disease states [231]. Chapter 2 demonstrates how established techniques such as BPTI refolding assays can help us characterise the oxidative folding abilities of new enzymes and compare their abilities to those of other often better-known enzymes. Chapter 2 also demonstrates how new enzymes can yield assay results that have not yet been observed before. Chapter 1 demonstrates how available high-quality data sources can be combined to infer new knowledge about already well-studied oxidative folding pathways such as the one in *E. coli*. Furthermore it highlights the possibility of quantitative proteomes as data sources for inferring kinetic parameters. Chapter 3 showcases how novel and future data sources can be used as tools to better understand the complexity and utilisation of DSBs across different species.

Research approaches such as the ones developed in Chapter 1 and 3 will become more widely used in the future. As the availability and quality of data sources both predicted and measured improves, the combination of different genome-wide quantitative data sources to yield new scientific insights will become more widely applicable. The analysis developed in Chapter 3 is a good example how new data sources can be created to help fill in the blanks in our knowledge. For example, the immense availability of genomes and proteomes has allowed us to use the cysteine counts of organisms as a useful proxy for DSB utilisation [232]. However, this approach is unable to differentiate between consecutive and non-consecutive DSBs and as Chapter 3 demonstrates, this information can be extracted from predicted protein structures. Chapter 1 displays the usefulness of combining information from different data sources to infer new insights. But at the same time it demonstrates how much data harmonisation and processing is necessary before different data sources can be combined for answering research questions.

In my opinion it will take the help of machine learning algorithms for us to catch up our available knowledge to our available data. AlphaFold has impressively demonstrated the way we can utilize such algorithms to infer new information from already available information (sequence information and measured protein structures). This data sources are of immense value to researchers working in many different fields of science. However, machine learning approaches are becoming more widespread in more and more research project include them already [233]. Although the diverse and exception rich nature of biological research represents additional challenges for these algorithms [234]. It might take different machine learning approaches such as automated machine learning to help combine different and complex biological data sources into effective prediction tools [235]. In summary, the contemporary research landscape in biology, biochemistry, biotechnology and computational biology is fast changing, promising and exciting.

Conclusion

This thesis covers a wide range of different techniques for investigating oxidative folding on either the individual protein level, the organism-level and even at the level of phylogenetic kingdoms. Together with the introductory review on current trends and developments in the field of recombinant protein productions in microbial organisms, this thesis covers many of the essential areas of contemporary techniques, challenges and innovations in the multi-faceted field of oxidative protein folding. The findings and results presented in this thesis increase our current understanding of DSB formation and might help improve future recombinant protein production projects and future research.

References

[1] A. Conesa, P. J. Punt, N. Van Luijk, and C. A. M. J. J. Van den Hondel, “The Secretion Pathway in Filamentous Fungi: A Biotechnological View,” *Fungal Genetics and Biology*, vol. 33, no. 3, pp. 155–171, Aug. 2001, doi: 10.1006/FGBI.2001.1276.

[2] J. R. Swartz, “Advances in *Escherichia coli* production of therapeutic proteins,” *Curr Opin Biotechnol*, vol. 12, no. 2, pp. 195–201, Apr. 2001, doi: 10.1016/S0958-1669(00)00199-3.

[3] N. Sutin, “Theory of electron transfer reactions: insights and hindsights,” *Progress in inorganic chemistry*, vol. 30, pp. 441–498, 2009.

[4] M. Kanehisa, “Toward understanding the origin and evolution of cellular organisms,” *Protein Science*, vol. 28, no. 11, pp. 1947–1951, Nov. 2019, doi: 10.1002/PRO.3715.

[5] L. A. Rettenbacher *et al.*, “Microbial protein cell factories fight back?,” *Trends Biotechnol*, vol. 40, no. 5, pp. 576–590, May 2022, doi: 10.1016/J.TIBTECH.2021.10.003.

[6] A. L. Demain, “Microbial biotechnology,” *Trends Biotechnol*, vol. 18, no. 1, pp. 26–31, Jan. 2000, doi: 10.1016/S0167-7799(99)01400-6.

[7] M. Gavrilescu and Y. Chisti, “Biotechnology—a sustainable alternative for chemical industry,” *Biotechnol Adv*, vol. 23, no. 7–8, pp. 471–499, 2005.

[8] J. Y. Kim, Y. G. Kim, and G. M. Lee, “CHO cells in biotechnology for production of recombinant proteins: Current state and further potential,” *Appl Microbiol Biotechnol*, vol. 93, no. 3, pp. 917–930, Feb. 2012, doi: 10.1007/S00253-011-3758-5.

[9] H. P. Sørensen and K. K. Mortensen, “Advanced genetic strategies for recombinant protein expression in *Escherichia coli*,” *J Biotechnol*, vol. 115, no. 2, pp. 113–128, Jan. 2005, doi: 10.1016/J.JBIOTEC.2004.08.004.

[10] G. L. Rosano and E. A. Ceccarelli, “Recombinant protein expression in *Escherichia coli*: Advances and challenges,” *Front Microbiol*, vol. 5, no. APR, p. 79503, Apr. 2014, doi: 10.3389/FMICB.2014.00172.

[11] S. Jozefczuk *et al.*, “Metabolomic and transcriptomic stress response of *Escherichia coli*,” *Mol Syst Biol*, vol. 6, no. 1, p. 364, Jan. 2010, doi: 10.1038/MSB.2010.18.

[12] A. M. Feist and B. Palsson, “The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*,” *Nature Biotechnology 2008 26:6*, vol. 26, no. 6, pp. 659–667, Jun. 2008, doi: 10.1038/nbt1401.

[13] A. Khodayari and C. D. Maranas, “A genome-scale *Escherichia coli* kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains,” *Nature Communications 2016 7:1*, vol. 7, no. 1, pp. 1–12, Dec. 2016, doi: 10.1038/ncomms13806.

[14] T. Baba *et al.*, “Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection,” *Mol Syst Biol*, vol. 2, no. 1, p. 2006.0008, Jan. 2006, doi: 10.1038/MSB4100050.

[15] Y. Yamada, M. Matsuda, K. Maeda, and K. Mikata, “The Phylogenetic Relationships of Methanol-assimilating Yeasts Based on the Partial Sequences of 18S and 26S Ribosomal RNAs: The

Proposal of Komagataella Gen. Nov. (Saccharomycetaceae)," *Biosci Biotechnol Biochem*, vol. 59, no. 3, pp. 439–444, 1995, doi: 10.1271/BBB.59.439.

[16] M. Inan and M. M. Meagher, "Non-repressing carbon sources for alcohol oxidase (AOX1) promoter of *Pichia pastoris*," *J Biosci Bioeng*, vol. 92, no. 6, pp. 585–589, Jan. 2001, doi: 10.1016/S1389-1723(01)80321-2.

[17] R. A. Rader and E. S. Langer, "Fifteen years of progress: Biopharmaceutical industry survey results," *Pharmaceutical Technology Europe*, vol. 30, no. 7, pp. 10–12, 2018.

[18] N. K. Tripathi and A. Shrivastava, "Recent Developments in Bioprocessing of Recombinant Proteins: Expression Hosts and Process Development," *Front Bioeng Biotechnol*, vol. 7, p. 496566, Dec. 2019, doi: 10.3389/FBIOE.2019.00420.

[19] G. Walsh, "Biopharmaceutical benchmarks 2018," *Nature Biotechnology 2018 36:12*, vol. 36, no. 12, pp. 1136–1145, Dec. 2018, doi: 10.1038/nbt.4305.

[20] L. Sanchez-Garcia, L. Martín, R. Mangues, N. Ferrer-Miralles, E. Vázquez, and A. Villaverde, "Recombinant pharmaceuticals from microbial cells: A 2015 update," *Microb Cell Fact*, vol. 15, no. 1, pp. 1–7, Feb. 2016, doi: 10.1186/S12934-016-0437-3.

[21] A. Sandomenico, J. P. Sivaccumar, and M. Ruvo, "Evolution of *Escherichia coli* expression system in producing antibody recombinant fragments," *Int J Mol Sci*, vol. 21, no. 17, pp. 1–39, 2020, doi: 10.3390/ijms21176324.

[22] S. K. Gupta and P. Shukla, "Microbial platform technology for recombinant antibody fragment production: A review," *Crit Rev Microbiol*, vol. 43, no. 1, pp. 31–42, 2017, doi: 10.3109/1040841X.2016.1150959.

[23] A. Lakowitz, T. Godard, R. Biedendieck, and R. Krull, "Mini review: Recombinant production of tailored bio-pharmaceuticals in different Bacillus strains and future perspectives," *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 126, pp. 27–39, 2018, doi: 10.1016/j.ejpb.2017.06.008.

[24] Y. Liu and H. Huang, "Expression of single-domain antibody in different systems," *Appl Microbiol Biotechnol*, vol. 102, no. 2, pp. 539–551, 2018, doi: 10.1007/s00253-017-8644-3.

[25] O. Spadiut, S. Capone, F. Krainer, A. Glieder, and C. Herwig, "Microbials for the production of monoclonal antibodies and antibody fragments," *Trends Biotechnol*, vol. 32, no. 1, pp. 54–60, Jan. 2014, doi: 10.1016/j.tibtech.2013.10.002.

[26] B. Owens, "Faster, deeper, smaller-the rise of antibody-like scaffolds," *Nat Biotechnol*, vol. 35, no. 7, pp. 602–603, Jul. 2017, doi: 10.1038/nbt0717-602.

[27] R. Simeon and Z. Chen, "In vitro-engineered non-antibody protein therapeutics," *Protein Cell*, vol. 9, no. 1, pp. 3–14, 2018, doi: 10.1007/s13238-017-0386-6.

[28] H. K. Binz *et al.*, "Design and characterization of MP0250, a tri-specific anti-HGF/anti-VEGF DARPin® drug candidate," *MAbs*, vol. 9, no. 8, pp. 1262–1269, 2017, doi: 10.1080/19420862.2017.1305529.

[29] M. N. Baeshen *et al.*, "Production of biopharmaceuticals in *E. Coli*: Current scenario and future perspectives," *J Microbiol Biotechnol*, vol. 25, no. 7, pp. 953–962, Jul. 2015, doi: 10.4014/jmb.1412.12079.

[30] A. F. Jozala *et al.*, “Biopharmaceuticals from microorganisms: from production to purification,” *Brazilian Journal of Microbiology*, vol. 47, no. Suppl 1, pp. 51–63, Dec. 2016, doi: 10.1016/j.bjm.2016.10.007.

[31] L. Sanchez-Garcia, L. Martín, R. Mangues, N. Ferrer-Miralles, E. Vázquez, and A. Villaverde, “Recombinant pharmaceuticals from microbial cells: A 2015 update,” *Microb Cell Fact*, vol. 15, no. 1, pp. 1–7, 2016, doi: 10.1186/s12934-016-0437-3.

[32] W. Zhu, R. Xu, G. Gong, L. Xu, Y. Hu, and L. Xie, “Medium optimization for high yield production of human serum albumin in *Pichia pastoris* and its efficient purification,” *Protein Expr Purif*, vol. 181, pp. 1–11, 2021, doi: 10.1016/j.pep.2021.105831.

[33] M. W. T. Werten, G. Eggink, M. A. Cohen Stuart, and F. A. de Wolf, “Production of protein-based polymers in *Pichia pastoris*,” *Biotechnol Adv*, vol. 37, no. 5, pp. 642–666, 2019, doi: 10.1016/j.biotechadv.2019.03.012.

[34] A. Vieira Gomes, T. Souza Carmo, L. Silva Carvalho, F. Mendonça Bahia, and N. Parachin, “Comparison of Yeasts as Hosts for Recombinant Protein Production,” *Microorganisms*, vol. 6, no. 2, p. 38, 2018, doi: 10.3390/microorganisms6020038.

[35] L. Gifre, A. Arís, À. Bach, and E. Garcia-Fruitós, “Trends in recombinant protein use in animal production,” *Microb Cell Fact*, vol. 16, no. 1, p. 40, 2017, doi: 10.1186/s12934-017-0654-4.

[36] J. M. van Dijl and M. Hecker, “*Bacillus subtilis*: From soil bacterium to super-secreting cell factory,” *Microb Cell Fact*, vol. 12, no. 1, pp. 1–6, 2013, doi: 10.1186/1475-2859-12-3.

[37] R. Singh, M. Kumar, A. Mittal, and P. K. Mehta, “Microbial enzymes: industrial progress in 21st century,” *3 Biotech*, vol. 6, no. 2, p. 174, 2016, doi: 10.1007/s13205-016-0485-8.

[38] H. H. Su, J. C. Chen, and P. T. Chen, “Production of recombinant human epidermal growth factor in *Bacillus subtilis*,” *J Taiwan Inst Chem Eng*, vol. 106, pp. 86–91, 2020, doi: 10.1016/j.jtice.2019.10.024.

[39] V. B. Yadwad, S. Wilson, and O. P. Ward, “Production of human epidermal growth factor by an ampicillin resistant recombinant *Escherichia coli* strain,” *Biotechnol Lett*, vol. 16, no. 9, pp. 885–890, Sep. 1994, doi: 10.1007/BF00128619.

[40] J. Valdés *et al.*, “Physiological study in *Saccharomyces cerevisiae* for overproduction of a homogeneous human epidermal growth factor molecule,” *Biotecnologia Aplicada*, vol. 26, no. 2, pp. 166–167, 2009.

[41] S. Eissazadeh, H. Moeini, M. G. Dezfouli, S. Heidary, R. Nelofer, and M. P. Abdullah, “Production of recombinant human epidermal growth factor in *Pichia pastoris*,” *Brazilian Journal of Microbiology*, vol. 48, no. 2, pp. 286–293, Apr. 2017, doi: 10.1016/j.bjm.2016.10.017.

[42] B. Şahin, S. Öztürk, P. Çalık, and T. H. Özdamar, “Feeding strategy design for recombinant human growth hormone production by *Bacillus subtilis*,” *Bioprocess Biosyst Eng*, vol. 38, no. 10, Oct. 2015, doi: 10.1007/s00449-015-1426-3.

[43] I. Guerrero Montero *et al.*, “*Escherichia coli* ‘TatExpress’ strains export several g/L human growth hormone to the periplasm by the Tat pathway,” *Biotechnol Bioeng*, vol. 116, no. 12, pp. 3282–3291, 2019, doi: 10.1002/bit.27147.

[44] M. S. Hahm and B. H. Chung, “Secretory expression of human growth hormone in *Saccharomyces cerevisiae* using three different leader sequences,” *Biotechnology and Bioprocess Engineering*, vol. 6, no. 4, pp. 306–309, Aug. 2001, doi: 10.1007/BF02931995.

[45] P. Çalik *et al.*, “Effect of co-substrate sorbitol different feeding strategies on human growth hormone production by recombinant *Pichia pastoris*,” *Journal of Chemical Technology and Biotechnology*, vol. 88, no. 9, pp. 1631–1640, Sep. 2013, doi: 10.1002/jctb.4011.

[46] R. Kunaparaju, M. Liao, and N. A. Sunstrom, “Epi-CHO, an episomal expression system for recombinant protein production in CHO cells,” *Biotechnol Bioeng*, vol. 91, no. 6, pp. 670–677, Sep. 2005, doi: 10.1002/bit.20534.

[47] R. Breitling, D. Gerlach, M. Hartmann, and D. Behnke, “Secretory expression in *Escherichia coli* and *Bacillus subtilis* of human interferon α genes directed by staphylokinase signals,” *MGG Molecular & General Genetics*, vol. 217, no. 2–3, pp. 384–391, Jun. 1989, doi: 10.1007/BF02464908.

[48] K. R. Babu, S. Swaminathan, S. Marten, N. Khanna, and U. Rinas, “Production of interferon-α in high cell density cultures of recombinant *Escherichia coli* and its single step purification from refolded inclusion body proteins,” *Appl Microbiol Biotechnol*, vol. 53, no. 6, pp. 655–660, Jun. 2000, doi: 10.1007/s002530000318.

[49] J. Chu, S. Zhang, and Y. Zhuang, “Fermentation Process Optimization of Recombinant *Saccharomyces cerevisiae* for the Production of Human Interferon-α2a,” *Applied Biochemistry and Biotechnology - Part A Enzyme Engineering and Biotechnology*, vol. 111, no. 3, pp. 129–137, 2003, doi: 10.1385/ABAB:111:3:129.

[50] S. Katla, K. N. R. Yoganand, S. Hingane, C. T. Ranjith Kumar, B. Anand, and S. Sivaprakasam, “Novel glycosylated human interferon alpha 2b expressed in glycoengineered *Pichia pastoris* and its biological activity: N-linked glycoengineering approach,” *Enzyme Microb Technol*, vol. 128, pp. 49–58, Sep. 2019, doi: 10.1016/j.enzmictec.2019.05.007.

[51] Y. Shiga *et al.*, “Efficient production of N-terminally truncated biologically active human interleukin-6 by bacillus brevis,” *Biosci Biotechnol Biochem*, vol. 64, no. 3, pp. 665–669, Jan. 2000, doi: 10.1271/bbb.64.665.

[52] A. Gaciarz *et al.*, “Efficient soluble expression of disulfide bonded proteins in the cytoplasm of *Escherichia coli* in fed-batch fermentations on chemically defined minimal media,” *Microb Cell Fact*, vol. 16, no. 1, pp. 1–12, 2017, doi: 10.1186/s12934-017-0721-x.

[53] Y. Guisez *et al.*, “Production and purification of recombinant human interleukin-6 secreted by the yeast *Saccharomyces cerevisiae*,” *Eur J Biochem*, vol. 198, no. 1, pp. 217–222, May 1991, doi: 10.1111/j.1432-1033.1991.tb16004.x.

[54] H. Li, Y. Wang, A. Xu, S. Li, S. Jin, and D. Wu, “Large-scale production, purification and bioactivity assay of recombinant human interleukin-6 in the methylotrophic yeast *Pichia pastoris*,” *FEMS Yeast Res*, vol. 11, no. 2, pp. 160–167, Mar. 2011, doi: 10.1111/j.1567-1364.2010.00701.x.

[55] Y. Zhang, Y. Katakura, H. Ohashi, and S. Shirahata, “Efficient and inducible production of human interleukin 6 in Chinese hamster ovary cells using a novel expression system,” *Cytotechnology*, vol. 25, no. 1–3, pp. 53–60, 1997, doi: 10.1023/a:1007972002180.

[56] S. W. Lee, N. H. Kang, and J. W. Choi, "Functional Secretion of Granulocyte Colony Stimulating Factor in *Bacillus subtilis* and Its Thermogenic Activity in Brown Adipocytes," *Biotechnology and Bioprocess Engineering*, vol. 24, no. 2, pp. 298–307, Mar. 2019, doi: 10.1007/s12257-019-0127-1.

[57] V. Dasari, N. Venkaiah Chowdary, and B. Adibhatla Kali Satya, "A novel process for production of recombinant human g-csf," WO2008096368A2, 2008 [Online]. Available: https://patents.google.com/patent/WO2008096368A2/en

[58] C. S. Bae, D. S. Yang, J. Lee, and Y. H. Park, "Improved process for production of recombinant yeast-derived monomeric human G-CSF," *Appl Microbiol Biotechnol*, vol. 52, no. 3, pp. 338–344, Sep. 1999, doi: 10.1007/s002530051529.

[59] A. Bahrami, S. A. Shojaosadati, R. Khalilzadeh, A. R. Saeedinia, E. Vasheghani Farahani, and J. Mohammadian-Mosaabadi, "Production of Recombinant Human Granulocyte-Colony Stimulating Factor by *Pichia pastoris*," *Iran J Biotechnol*, vol. 5, no. 3, pp. 162–169, 2007.

[60] L. Rotondaro, L. Mazzanti, A. Mele, and G. Rovera, "High-level Expression of a cDNA for Human Granulocyte Colony-Stimulating Factor in Chinese Hamster Ovary Cells: Effect of 3'-Noncoding Sequences," *Applied Biochemistry and Biotechnology - Part B Molecular Biotechnology*, vol. 7, no. 3, pp. 231–240, Jun. 1997, doi: 10.1007/BF02740814.

[61] J. Olmos-Soto and R. Contreras-Flores, "Genetic system constructed to overproduce and secrete proinsulin in *Bacillus subtilis*," *Appl Microbiol Biotechnol*, vol. 62, no. 4, pp. 369–373, Sep. 2003, doi: 10.1007/s00253-003-1289-4.

[62] C. S. Shin, M. S. Hong, C. S. Bae, and J. Lee, "Enhanced production of human mini-proinsulin in fed-batch cultures at high cell density of *Escherichia coli* BL21(DE3)[pET-3aT2M2]," *Biotechnol Prog*, vol. 13, no. 3, pp. 249–257, 1997, doi: 10.1021/bp970018m.

[63] T. Kjeldsen, "Yeast secretory expression of insulin precursors," *Appl Microbiol Biotechnol*, vol. 54, no. 3, pp. 277–286, 2000, doi: 10.1007/s002530000402.

[64] C. Gurramkonda *et al.*, "Application of simple fed-batch technique to high-level secretory production of insulin precursor using *Pichia pastoris* with subsequent purification and conversion to human insulin," *Microb Cell Fact*, vol. 9, no. 1, p. 31, 2010, doi: 10.1186/1475-2859-9-31.

[65] S. C. O. Pak, S. M. N. Hunt, M. J. Sleigh, and P. P. Gray, "Expression of Recombinant Human Insulin in Chinese Hamster Ovary Cells is Complicated by Intracellular Insulin-Degrading Enzymes," in *New Developments and New Applications in Animal Cell Technology*, Dordrecht, 2006, pp. 59–67. doi: 10.1007/0-306-46860-3_10.

[66] M. M. Müller, "Post-Translational Modifications of Protein Backbones: Unique Functions, Mechanisms, and Challenges," *Biochemistry*, vol. 57, no. 2, pp. 177–185, 2018, doi: 10.1021/acs.biochem.7b00861.

[67] B. Macek, K. Forchhammer, J. Hardouin, E. Weber-Ban, C. Grangeasse, and I. Mijakovic, "Protein post-translational modifications in bacteria," *Nat Rev Microbiol*, vol. 17, no. 11, pp. 651–664, 2019, doi: 10.1038/s41579-019-0243-0.

[68] G. D. Tredwell, R. Aw, B. Edwards-Jones, D. J. Leak, and J. G. Bundy, "Rapid screening of cellular stress responses in recombinant *Pichia pastoris* strains using metabolite profiling," *J Ind Microbiol Biotechnol*, vol. 44, no. 3, pp. 413–417, 2017, doi: 10.1007/s10295-017-1904-5.

[69] C. Hetz, K. Zhang, and R. J. Kaufman, "Mechanisms, regulation and functions of the unfolded protein response," *Nature Reviews Molecular Cell Biology 2020 21:8*, vol. 21, no. 8, pp. 421–438, May 2020, doi: 10.1038/s41580-020-0250-z.

[70] M. Valkonen, M. Penttilä, and M. Saloheimo, "Effects of inactivation and constitutive expression of the unfolded-protein response pathway on protein production in the yeast *Saccharomyces cerevisiae*," *Appl Environ Microbiol*, vol. 69, no. 4, pp. 2065–2072, Apr. 2003, doi: 10.1128/AEM.69.4.2065-2072.2003.

[71] M. Ellis, P. Patel, M. Edon, W. Ramage, R. Dickinson, and D. P. Humphreys, "Development of a high yielding *E. coli* periplasmic expression system for the production of humanized Fab' fragments," *Biotechnol Prog*, vol. 33, no. 1, pp. 212–220, Jan. 2017, doi: 10.1002/btpr.2393.

[72] A. Quesada-Ganuza *et al.*, "Identification and optimization of PrsA in *Bacillus subtilis* for improved yield of amylase," *Microbial Cell Factories 2019 18:1*, vol. 18, no. 1, pp. 1–16, Sep. 2019, doi: 10.1186/S12934-019-1203-0.

[73] L. Navone *et al.*, "Disulfide bond engineering of AppA phytase for increased thermostability requires co-expression of protein disulfide isomerase in *Pichia pastoris*," *Biotechnol Biofuels*, vol. 14, no. 1, pp. 1–14, 2021, doi: 10.1186/s13068-021-01936-8.

[74] M. Martínez-Alonso *et al.*, "Rehosting of bacterial chaperones for high-quality protein production," *Appl Environ Microbiol*, vol. 75, no. 24, pp. 7850–7854, Dec. 2009, doi: 10.1128/AEM.01532-09.

[75] R. Levy, R. Weiss, G. Chen, B. L. Iverson, and G. Georgiou, "Production of correctly folded fab antibody fragment in the cytoplasm of *Escherichia coli* trxB gor mutants via the coexpression of molecular chaperones," *Protein Expr Purif*, vol. 23, no. 2, pp. 338–347, 2001, doi: 10.1006/prep.2001.1520.

[76] J. Lobstein, C. A. Emrich, C. Jeans, M. Faulkner, P. Riggs, and M. Berkmen, "Erratum to: SHuffle, a novel *Escherichia coli* protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm," 2016. doi: 10.1186/s12934-016-0512-9.

[77] G. Ren, N. Ke, and M. Berkmen, "Use of the SHuffle strains in production of proteins," *Curr Protoc Protein Sci*, vol. 2016, pp. 5.26.1-5.26.21, 2016, doi: 10.1002/cpps.11.

[78] T. R. H. M. Kouwen and J. M. Van Dijl, "Applications of thiol-disulfide oxidoreductases for optimized in vivo production of functionally active proteins in Bacillus," *Appl Microbiol Biotechnol*, vol. 85, no. 1, pp. 45–52, 2009, doi: 10.1007/s00253-009-2212-4.

[79] F. Hatahet, V. D. Nguyen, K. E. H. Salo, and L. W. Ruddock, "Disruption of reducing pathways is not essential for efficient disulfide bond formation in the cytoplasm of *E. coli*.," *Microb Cell Fact*, vol. 9, no. 1, p. 67, 2010, doi: 10.1186/1475-2859-9-67.

[80] A. A. Sohail, M. Gaikwad, P. Khadka, M. J. Saaranen, and L. W. Ruddock, "Production of extracellular matrix proteins in the cytoplasm of *E. coli*: Making giants in tiny factories," *Int J Mol Sci*, vol. 21, no. 3, p. 688, 2020, doi: 10.3390/ijms21030688.

[81] G. Guidi, L. Di Tucci, and M. D. Santambrogio, “ProFAX: A hardware acceleration of a protein folding algorithm,” *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a Better Tomorrow, RTSI 2016*, Nov. 2016, doi: 10.1109/RTSI.2016.7740584.

[82] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr, “The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain,” *Proceedings of the National Academy of Sciences*, vol. 47, no. 9, pp. 1309–1314, 1961.

[83] M. E. Feder, D. A. Parsell, and S. L. Lindquist, “The Stress Response and Stress Proteins,” *Cell Biology of Trauma*, pp. 177–191, Aug. 2020, doi: 10.1201/9781003067801-16.

[84] I. Braakman and D. N. Hebert, “Protein folding in the endoplasmic reticulum,” *Cold Spring Harb Perspect Biol*, vol. 5, no. 5, p. a013201, 2013.

[85] C. M. Dobson, “Protein misfolding, evolution and disease,” *Trends Biochem Sci*, vol. 24, no. 9, pp. 329–332, Sep. 1999, doi: 10.1016/S0968-0004(99)01445-0.

[86] F. U. Hartl, “Protein Misfolding Diseases,” *https://doi.org/10.1146/annurev-biochem-061516-044518*, vol. 86, pp. 21–26, Jun. 2017, doi: 10.1146/ANNUREV-BIOCHEM-061516-044518.

[87] A. Kumar, A. Singh, and Ekavali, “A review on Alzheimer’s disease pathophysiology and its management: an update,” *Pharmacological Reports*, vol. 67, no. 2, pp. 195–203, Apr. 2015, doi: 10.1016/J.PHAREP.2014.09.004.

[88] S. B. Prusiner, “Prion diseases and the BSE crisis,” *Science (1979)*, vol. 278, no. 5336, pp. 245–251, Oct. 1997, doi: 10.1126/SCIENCE.278.5336.245.

[89] J. Tyedmers, A. Mogk, and B. Bukau, “Cellular strategies for controlling protein aggregation,” *Nature Reviews Molecular Cell Biology 2010 11:11*, vol. 11, no. 11, pp. 777–788, Oct. 2010, doi: 10.1038/nrm2993.

[90] C. Maas, S. Hermeling, B. Bouma, W. Jiskoot, and M. F. B. G. Gebbink, “A role for protein misfolding in immunogenicity of biopharmaceuticals,” *Journal of Biological Chemistry*, vol. 282, no. 4, pp. 2229–2236, Jan. 2007, doi: 10.1074/jbc.M605984200.

[91] F. Baneyx and M. Mujacic, “Recombinant protein folding and misfolding in *Escherichia coli*” *Nature Biotechnology 2004 22:11*, vol. 22, no. 11, pp. 1399–1408, Nov. 2004, doi: 10.1038/nbt1029.

[92] J. Kopp and O. Spadiut, “Inclusion Bodies: Status Quo and Perspectives,” *Methods Mol Biol*, vol. 2617, pp. 1–13, 2023, doi: 10.1007/978-1-0716-2930-7_1.

[93] J. W. H. WONG and P. J. HOGG, “Analysis of disulfide bonds in protein structures,” *Journal of Thrombosis and Haemostasis*, vol. 8, no. 10, p. 2345, Oct. 2010, doi: 10.1111/j.1538-7836.2010.03894.x.

[94] F. Hatahet and L. W. Ruddock, “Protein Disulfide Isomerase: A Critical Evaluation of Its Function in Disulfide Bond Formation,” *Antioxid Redox Signal*, vol. 11, no. 11, pp. 2807–2850, May 2009, doi: 10.1089/ars.2009.2466.

[95] T. Fujimoto, K. Inaba, and H. Kadokura, “Methods to identify the substrates of thiol-disulfide oxidoreductases,” *Protein Science*, vol. 28, no. 1, pp. 30–40, Jan. 2019, doi: 10.1002/PRO.3530.

[96] G. Carruth and E. Ehrlich, “Bond energies,” *Volume Library*, vol. 1, 2002.

[97] J. W. H. Wong, S. Y. W. Ho, and P. J. Hogg, “Disulfide Bond Acquisition through Eukaryotic Protein Evolution,” *Mol Biol Evol*, vol. 28, no. 1, pp. 327–334, Jan. 2011, doi: 10.1093/MOLBEV/MSQ194.

[98] J. L. Martin, “Thioredoxin—a fold for all reasons,” *Structure*, vol. 3, no. 3, pp. 245–250, 1995.

[99] A. Ingles-Prieto *et al.*, “Conservation of protein structure over four billion years,” *Structure*, vol. 21, no. 9, pp. 1690–1697, 2013.

[100] H. Eklund *et al.*, “Conformational and functional similarities between glutaredoxin and thioredoxins.,” *EMBO J*, vol. 3, no. 7, pp. 1443–1449, Jul. 1984, doi: 10.1002/J.1460-2075.1984.TB01994.X.

[101] G. Ren *et al.*, “Properties of the thioredoxin fold superfamily are modulated by a single amino acid residue,” *Journal of Biological Chemistry*, vol. 284, no. 15, pp. 10150–10159, Apr. 2009, doi: 10.1074/jbc.M809509200.

[102] G. Krause, J. Lundstrom, J. L. Barea, C. P. De la Cuesta, and A. Holmgren, “Mimicking the active site of protein disulfide-isomerase by substitution of proline 34 in *Escherichia coli* thioredoxin,” *Journal of Biological Chemistry*, vol. 266, no. 15, pp. 9494–9500, May 1991, doi: 10.1016/S0021-9258(18)92848-6.

[103] A. Zapun, J. C. A. Bardwell, and T. E. Creighton, “The Reactive and Destabilizing Disulfide Bond of DsbA, a Protein Required for Protein Disulfide Bond Formation in K/vot,” *Biochemistry*, vol. 32, pp. 5083–5092, 1993, Accessed: Sep. 11, 2023. [Online]. Available: https://pubs.acs.org/sharingguidelines

[104] C. W. Gruber, M. Čemažar, B. Heras, J. L. Martin, and D. J. Craik, “Protein disulfide isomerase: the structure of oxidative folding,” *Trends Biochem Sci*, vol. 31, no. 8, pp. 455–464, Aug. 2006, doi: 10.1016/J.TIBS.2006.06.001.

[105] H. Kadokura, H. Tian, T. Zander, J. C. A. Bardwell, and J. Beckwith, “Snapshots of DsbA in Action: Detection of Proteins in the Process of Oxidative Folding,” *Science (1979)*, vol. 303, no. 5657, pp. 534–537, Jan. 2004, doi: 10.1126/SCIENCE.1091724.

[106] D. Su, C. Berndt, D. E. Fomenko, A. Holmgren, and V. N. Gladyshev, “A conserved cis-proline precludes metal binding by the active site thiolates in members of the thioredoxin family of proteins,” *Biochemistry*, vol. 46, no. 23, pp. 6903–6910, Jun. 2007, doi: 10.1021/BI700152B.

[107] S. Dai *et al.*, “Structural snapshots along the reaction pathway of ferredoxin–thioredoxin reductase,” *Nature 2007 448:7149*, vol. 448, no. 7149, pp. 92–96, Jul. 2007, doi: 10.1038/nature05937.

[108] S. M. Marino and V. N. Gladyshev, “Cysteine Function Governs Its Conservation and Degeneration and Restricts Its Utilization on Protein Surfaces,” *J Mol Biol*, vol. 404, no. 5, pp. 902–916, Dec. 2010, doi: 10.1016/J.JMB.2010.09.027.

[109] T. Fujimoto, K. Inaba, and H. Kadokura, “Methods to identify the substrates of thiol-disulfide oxidoreductases,” *Protein Science*, vol. 28, no. 1, pp. 30–40, Jan. 2019, doi: 10.1002/PRO.3530.

[110] J. Regeimbal and J. C. A. Bardwell, “DsbB catalyzes disulfide bond formation de novo,” *Journal of Biological Chemistry*, vol. 277, no. 36, pp. 32706–32713, 2002.

[111] M. Bader, W. Muse, D. P. Ballou, C. Gassner, and J. C. A. Bardwell, “Oxidative protein folding is driven by the electron transport system,” *Cell*, vol. 98, no. 2, pp. 217–227, Jul. 1999, doi: 10.1016/S0092-8674(00)81016-8.

[112] C. S. Sevier and C. A. Kaiser, “Ero1 and redox homeostasis in the endoplasmic reticulum,” *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1783, no. 4, pp. 549–556, Apr. 2008, doi: 10.1016/J.BBAMCR.2007.12.011.

[113] B. P. Tu and J. S. Weissman, “The FAD- and O(2)-dependent reaction cycle of Ero1-mediated oxidative protein folding in the endoplasmic reticulum,” *Mol Cell*, vol. 10, no. 5, pp. 983–994, Nov. 2002, doi: 10.1016/S1097-2765(02)00696-2.

[114] E. Gross, D. B. Kastner, C. A. Kaiser, and D. Fass, “Structure of Ero1p, source of disulfide bonds for oxidative protein folding in the cell,” *Cell*, vol. 117, no. 5, pp. 601–610, May 2004, doi: 10.1016/S0092-8674(04)00418-0.

[115] C. S. Sevier, H. Qu, N. Heldman, E. Gross, D. Fass, and C. A. Kaiser, “Modulation of cellular disulfide-bond formation and the ER redox environment by feedback regulation of Ero1,” *Cell*, vol. 129, no. 2, pp. 333–344, Apr. 2007, doi: 10.1016/J.CELL.2007.02.039.

[116] C. S. Sevier, J. W. Cuozzo, A. Vala, F. Åslund, and C. A. Kaiser, “A flavoprotein oxidase defines a new endoplasmic reticulum pathway for biosynthetic disulphide bond formation.,” *Nat Cell Biol*, vol. 3, no. 10, pp. 874–882, Oct. 2001, doi: 10.1038/NCB1001-874.

[117] A. Alon, E. J. Heckler, C. Thorpe, and D. Fass, “QSOX contains a pseudo-dimer of functional and degenerate sulfhydryl oxidase domains,” *FEBS Lett*, vol. 584, no. 8, pp. 1521–1525, Apr. 2010, doi: 10.1016/J.FEBSLET.2010.03.001.

[118] K. L. Hoober, S. L. Sheasley, H. F. Gilbert, and C. Thorpe, “Sulfhydryl oxidase from egg white. A facile catalyst for disulfide bond formation in proteins and peptides.,” *J Biol Chem*, vol. 274, no. 32, pp. 22147–22150, Aug. 1999, doi: 10.1074/JBC.274.32.22147.

[119] L. Debarbieux and J. Beckwith, “Electron avenue: Pathways of disulfide bond formation and isomerization,” *Cell*, vol. 99, no. 2, pp. 117–119, Oct. 1999, doi: 10.1016/S0092-8674(00)81642-6.

[120] A. Zapun, T. E. Creighton, and J. C. A. Bardwell, “The reactive and destabilizing disulfide bond of DsbA, a protein required for protein disulfide bond formation in vivo,” *Biochemistry*, vol. 32, no. 19, pp. 5083–5092, 1993, doi: 10.1021/BI00070A016.

[121] N. Ke and M. Berkmen, “Production of Disulfide-Bonded Proteins in *Escherichia coli*,” *Curr Protoc Mol Biol*, vol. 108, no. 1, pp. 16.1B.1-16.1B.21, Oct. 2014, doi: 10.1002/0471142727.MB1601BS108.

[122] J. Lundström and A. Holmgren, “Determination of the reduction-oxidation potential of the thioredoxin-like domains of protein disulfide-isomerase from the equilibrium with glutathione and thioredoxin,” *Biochemistry*, vol. 32, no. 26, pp. 6649–6655, 1993, doi: 10.1021/BI00077A018.

[123] N. J. Darby, E. Penka, and R. Vincentelli, “The multi-domain structure of protein disulfide isomerase is essential for high catalytic efficiency,” *J Mol Biol*, vol. 276, no. 1, pp. 239–247, Feb. 1998, doi: 10.1006/JMBI.1997.1504.

[124] C. L. Andersen, A. Matthey-Dupraz, D. Missiakas, and S. Raina, “A new *Escherichia coli* gene, dsbG, encodes a periplasmic protein involved in disulphide bond formation, required for recycling DsbA/DsbB and DsbC redox proteins,” *Mol Microbiol*, vol. 26, no. 1, pp. 121–132, Oct. 1997, doi: 10.1046/J.1365-2958.1997.5581925.X.

[125] S. Gleiter and J. C. A. Bardwell, “Disulfide bond isomerization in prokaryotes,” *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1783, no. 4, pp. 530–534, Apr. 2008, doi: 10.1016/J.BBAMCR.2008.02.009.

[126] A. Raturi and B. Mutus, “Characterization of redox state and reductase activity of protein disulfide isomerase under different redox environments using a sensitive fluorescent assay,” *Free Radic Biol Med*, vol. 43, no. 1, pp. 62–70, Jul. 2007, doi: 10.1016/J.FREERADBIOMED.2007.03.025.

[127] D. Missiakas, F. Schwager, and S. Raina, “Identification and characterization of a new disulfide isomerase-like protein (DsbD) in *Escherichia coli*.,” *EMBO J*, vol. 14, no. 14, pp. 3415–3424, Jul. 1995, doi: 10.1002/J.1460-2075.1995.TB07347.X.

[128] A. Rozhkova *et al.*, “Structural basis and kinetics of inter- and intramolecular disulfide exchange in the redox catalyst DsbD,” *EMBO J*, vol. 23, no. 8, pp. 1709–1719, Apr. 2004, doi: 10.1038/SJ.EMBOJ.7600178.

[129] A. Palma *et al.*, “Biochemical analysis of *Komagataella phaffii* oxidative folding proposes novel regulatory mechanisms of disulfide bond formation in yeast,” *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–13, Aug. 2023, doi: 10.1038/s41598-023-41375-z.

[130] M. Mahmud, M. S. Kaiser, T. M. McGinnity, and A. Hussain, “Deep Learning in Mining Biological Data,” *Cognit Comput*, vol. 13, no. 1, pp. 1–33, Jan. 2021, doi: 10.1007/S12559-020-09773-X.

[131] A. Bateman *et al.*, “UniProt: the Universal Protein Knowledgebase in 2023,” *Nucleic Acids Res*, vol. 51, no. D1, pp. D523–D531, Jan. 2023, doi: 10.1093/NAR/GKAC1052.

[132] B. Boeckmann *et al.*, “The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003,” *Nucleic Acids Res*, vol. 31, no. 1, pp. 365–370, Jan. 2003, doi: 10.1093/NAR/GKG095.

[133] N. L. Anderson and N. G. Anderson, “The human plasma proteome: history, character, and diagnostic prospects,” *Mol Cell Proteomics*, vol. 1, no. 11, pp. 845–867, 2002, doi: 10.1074/MCP.R200007-MCP200.

[134] C. Lindemann *et al.*, “Strategies in relative and absolute quantitative mass spectrometry based proteomics,” *Biol Chem*, vol. 398, no. 5–6, pp. 687–699, May 2017, doi: 10.1515/HSZ-2017-0104.

[135] F. Calderón-Celis, J. R. Encinar, and A. Sanz-Medel, “Standardization approaches in absolute quantitative proteomics with mass spectrometry,” *Mass Spectrom Rev*, vol. 37, no. 6, pp. 715–737, Nov. 2018, doi: 10.1002/MAS.21542.

[136] S. Maaß and D. Becher, “Methods and applications of absolute protein quantification in microbial systems,” *J Proteomics*, vol. 136, pp. 222–233, Mar. 2016, doi: 10.1016/J.JPROT.2016.01.015.

[137] E. W. Deutsch *et al.*, “The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition,” *Nucleic Acids Res*, vol. 45, no. Database issue, p. D1100, Jan. 2017, doi: 10.1093/NAR/GKW936.

[138] Y. Perez-Riverol *et al.*, “The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences,” *Nucleic Acids Res*, vol. 50, no. D1, pp. D543–D552, Jan. 2022, doi: 10.1093/NAR/GKAB1038.

[139] M. Wang, C. J. Herrmann, M. Simonovic, D. Szklarczyk, and C. von Mering, “Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines,” *Proteomics*, vol. 15, no. 18, pp. 3163–3168, Sep. 2015, doi: 10.1002/PMIC.201400441.

[140] M. Wang *et al.*, “PaxDb, a database of protein abundance averages across all three domains of life,” *Mol Cell Proteomics*, vol. 11, no. 8, pp. 492–500, Aug. 2012, doi: 10.1074/MCP.O111.014704.

[141] M. Jaskolski, Z. Dauter, and A. Wlodawer, “A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits,” *FEBS J*, vol. 281, no. 18, pp. 3985–4009, Sep. 2014, doi: 10.1111/FEBS.12796.

[142] J. M. Rondeau and H. Schreuder, “Protein Crystallography and Drug Discovery,” *The Practice of Medicinal Chemistry: Fourth Edition*, pp. 511–537, Jan. 2015, doi: 10.1016/B978-0-12-417205-0.00022-5.

[143] H. M. Berman *et al.*, “The Protein Data Bank,” *Nucleic Acids Res*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/NAR/28.1.235.

[144] B. Al-Lazikani, J. Jung, Z. Xiang, and B. Honig, “Protein structure prediction,” *Curr Opin Chem Biol*, vol. 5, no. 1, pp. 51–56, Feb. 2001, doi: 10.1016/S1367-5931(00)00164-2.

[145] J. T. Ngo and J. Marks, “Computational complexity of a problem in molecular structure prediction,” *Protein Engineering, Design and Selection*, vol. 5, no. 4, pp. 313–321, Jun. 1992, doi: 10.1093/PROTEIN/5.4.313.

[146] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, “Critical assessment of methods of protein structure prediction (CASP)—Round XIII,” *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1011–1020, Dec. 2019, doi: 10.1002/PROT.25823.

[147] A. W. Senior *et al.*, “Improved protein structure prediction using potentials from deep learning,” *Nature 2020 577:7792*, vol. 577, no. 7792, pp. 706–710, Jan. 2020, doi: 10.1038/s41586-019-1923-7.

[148] J. Jumper *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nature 2021 596:7873*, vol. 596, no. 7873, pp. 583–589, Jul. 2021, doi: 10.1038/s41586-021-03819-2.

[149] P. B. Moore, W. A. Hendrickson, R. Henderson, and A. T. Brunger, “The protein-folding problem: Not yet solved,” *Science (1979)*, vol. 375, no. 6580, p. 507, Feb. 2022, doi: 10.1126/SCIENCE.ABN9422.

[150] K. Hübner, S. Sahle, and U. Kummer, “Applications and trends in systems biology in biochemistry,” *FEBS J*, vol. 278, no. 16, pp. 2767–2857, Aug. 2011, doi: 10.1111/J.1742-4658.2011.08217.X.

[151] F. T. Bergmann *et al.*, “COPASI and its applications in biotechnology,” *J Biotechnol*, vol. 261, pp. 215–220, Nov. 2017, doi: 10.1016/J.JBIOTEC.2017.06.1200.

[152] H. Bisswanger, *Enzyme kinetics: principles and methods*. John Wiley & Sons, 2017.

[153] R. A. Alberty, “Enzyme kinetics,” *Adv Enzymol Relat Areas Mol Biol*, vol. 17, pp. 1–64, 1956.

[154] F. J. Bruggeman and H. V. Westerhoff, “The nature of systems biology,” *Trends Microbiol*, vol. 15, no. 1, pp. 45–50, Jan. 2007, doi: 10.1016/J.TIM.2006.11.003.

[155] D. M. Beal, E. L. Bastow, G. L. Staniforth, T. Von Der Haar, R. B. Freedman, and M. F. Tuite, “Quantitative Analyses of the Yeast Oxidative Protein Folding Pathway in Vitro and in Vivo,” *Antioxid Redox Signal*, vol. 31, no. 4, pp. 261–274, Aug. 2019, doi: 10.1089/ARS.2018.7615/SUPPL_FILE/SUPP_TABLE1.CSV.

[156] L. Rettenbacher and T. Von Der Haar, “SIMULATING OXIDATIVE PROTEIN FOLDING Results - Native Results - CyDisCo References : Acknowledgments : Abstract Methods Results - Native Results - CyDisCo Outlook References : Acknowledgments :,” vol. 34, no. 2016, 2020.

[157] M. Delic, M. Valli, A. B. Graf, M. Pfeffer, D. Mattanovich, and B. Gasser, “The secretory pathway: exploring yeast diversity,” *FEMS Microbiol Rev*, vol. 37, no. 6, pp. 872–914, Nov. 2013, doi: 10.1111/1574-6976.12020.

[158] E. Mössner, M. Huber-Wunderlich, and R. Glockshuber, “Characterization of *Escherichia coli* thioredoxin variants mimicking the active-sites of other thiol/disulfide oxidoreductases,” *Protein Science*, vol. 7, no. 5, pp. 1233–1244, May 1998, doi: 10.1002/PRO.5560070519.

[159] J. L. Pan and J. C. A. Bardwell, “The origami of thioredoxin-like folds,” *Protein Science*, vol. 15, no. 10, pp. 2217–2227, Oct. 2006, doi: 10.1110/PS.062268106.

[160] C. Appenzeller-Herzog and L. Ellgaard, “The human PDI family: Versatility packed into a single fold,” *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1783, no. 4, pp. 535–548, Apr. 2008, doi: 10.1016/J.BBAMCR.2007.11.010.

[161] S. I. Nishikawa, J. L. Brodsky, and K. Nakatsukasa, “Roles of Molecular Chaperones in Endoplasmic Reticulum (ER) Quality Control and ER-Associated Degradation (ERAD),” *The Journal of Biochemistry*, vol. 137, no. 5, pp. 551–555, May 2005, doi: 10.1093/JB/MVI068.

[162] T. Higo, M. Hattori, T. Nakamura, T. Natsume, T. Michikawa, and K. Mikoshiba, “Subtype-Specific and ER Lumenal Environment-Dependent Regulation of Inositol 1,4,5-Trisphosphate Receptor Type 1 by ERp44,” *Cell*, vol. 120, no. 1, pp. 85–98, Jan. 2005, doi: 10.1016/J.CELL.2004.11.048.

[163] M. P. Mattson, “Pathways towards and away from Alzheimer’s disease,” *Nature 2004 430:7000*, vol. 430, no. 7000, pp. 631–639, Aug. 2004, doi: 10.1038/nature02621.

[164] Y. Zhang, E. Baig, and D. B. Williams, “Functions of ERp57 in the Folding and Assembly of Major Histocompatibility Complex Class I Molecules,” *Journal of Biological Chemistry*, vol. 281, no. 21, pp. 14622–14631, May 2006, doi: 10.1074/JBC.M512073200.

[165] S. J. Kang and P. Cresswell, “Calnexin, Calreticulin, and ERp57 Cooperate in Disulfide Bond Formation in Human CD1d Heavy Chain,” *Journal of Biological Chemistry*, vol. 277, no. 47, pp. 44838–44844, Nov. 2002, doi: 10.1074/JBC.M207831200.

[166] M. Delic, M. Valli, A. B. Graf, M. Pfeffer, D. Mattanovich, and B. Gasser, “The secretory pathway: exploring yeast diversity,” *FEMS Microbiol Rev*, vol. 37, no. 6, pp. 872–914, Nov. 2013, doi: 10.1111/1574-6976.12020.

[167] J. M. Cregg, J. L. Cereghino, J. Shi, and D. R. Higgins, “Recombinant protein expression in *Pichia pastoris*,” *Molecular Biotechnology 2000 16:1*, vol. 16, no. 1, pp. 23–52, 2000, doi: 10.1385/MB:16:1:23.

[168] R. F. Goldberger, C. J. Epstein, and C. B. Anfinsen, “Acceleration of Reactivation of Reduced Bovine Pancreatic Ribonuclease by a Microsomal System from Rat Liver,” *Journal of Biological Chemistry*, vol. 238, no. 2, pp. 628–635, Feb. 1963, doi: 10.1016/S0021-9258(18)81309-6.

[169] A. Warsame, R. Vad, T. Kristensen, and T. B. Øyen, “Characterization of a gene encoding a *Pichia pastoris* protein disulfide isomerase,” *Biochem Biophys Res Commun*, vol. 281, no. 5, pp. 1176–1182, 2001, doi: 10.1006/bbrc.2001.4479.

[170] B. Brooks and M. Karplus, “Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor.,” *Proceedings of the National Academy of Sciences*, vol. 80, no. 21, pp. 6571–6575, Nov. 1983, doi: 10.1073/PNAS.80.21.6571.

[171] J. Horel *et al.*, “Reexamination of the Folding of BPTI: Predominance of Native Intermediates,” *Science (1979)*, vol. 253, no. 5026, pp. 1386–1393, 1991, doi: 10.1126/SCIENCE.1716783.

[172] “The Disulfide Folding Pathway of BPTI”, doi: 10.1126/science.1373519.

[173] R. Mousa, S. Lansky, G. Shoham, and N. Metanis, “BPTI folding revisited: switching a disulfide into methylene thioacetal reveals a previously hidden path,” *Chem Sci*, vol. 9, no. 21, pp. 4814–4820, May 2018, doi: 10.1039/C8SC01110A.

[174] C. M. Grant, “Role of the glutathione/glutaredoxin and thioredoxin systems in yeast growth and response to stress conditions,” *Mol Microbiol*, vol. 39, no. 3, pp. 533–541, Feb. 2001, doi: 10.1046/J.1365-2958.2001.02283.X.

[175] D. M. Townsend, K. D. Tew, and H. Tapiero, “The importance of glutathione in human disease,” *Biomedicine & Pharmacotherapy*, vol. 57, no. 3–4, pp. 145–155, May 2003, doi: 10.1016/S0753-3322(03)00043-X.

[176] A. Meister and M. E. Anderson, “Glutathione,” *Annu Rev Biochem*, vol. 52, no. 1, pp. 711–760, 1983.

[177] I. Rebrin and R. S. Sohal, “Pro-oxidant shift in glutathione redox state during aging,” *Adv Drug Deliv Rev*, vol. 60, no. 13–14, pp. 1545–1552, Oct. 2008, doi: 10.1016/J.ADDR.2008.06.001.

[178] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nature Methods 2020 17:3*, vol. 17, no. 3, pp. 261–272, Feb. 2020, doi: 10.1038/s41592-019-0686-2.

[179] L. W. Ruddock, T. R. Hirst, and R. B. Freedman, “pH-dependence of the dithiol-oxidizing activity of DsbA (a periplasmic protein thiol:disulphide oxidoreductase) and protein disulphide-isomerase: studies with a novel simple peptide substrate,” *Biochemical Journal*, vol. 315, no. 3, pp. 1001–1005, May 1996, doi: 10.1042/BJ3151001.

[180] D. G. Smyth, O. O. Blumenfeld, and W. Konigsberg, “Reactions of N-ethylmaleimide with peptides and amino acids,” *Biochemical Journal*, vol. 91, no. 3, pp. 589–595, Jun. 1964, doi: 10.1042/BJ0910589.

[181] J. S. Weissman and P. S. Kim, “A kinetic explanation for the rearrangement pathway of BPTI folding,” *Nature Structural Biology 1995 2:12*, vol. 2, no. 12, pp. 1123–1130, 1995, doi: 10.1038/nsb1295-1123.

[182] A. R. Karala, A. K. Lappi, and L. W. Ruddock, “Modulation of an Active-Site Cysteine pKa Allows PDI to Act as a Catalyst of both Disulfide Bond Formation and Isomerization,” *J Mol Biol*, vol. 396, no. 4, pp. 883–892, Mar. 2010, doi: 10.1016/J.JMB.2009.12.014.

[183] K. Araki and K. Inaba, “Structure, Mechanism, and Evolution of Ero1 Family Enzymes,” *https://home.liebertpub.com/ars*, vol. 16, no. 8, pp. 790–799, Feb. 2012, doi: 10.1089/ARS.2011.4418.

[184] Y. Sato and K. Inaba, “Disulfide bond formation network in the three biological kingdoms, bacteria, fungi and mammals,” *FEBS J*, vol. 279, no. 13, pp. 2262–2271, Jul. 2012, doi: 10.1111/J.1742-4658.2012.08593.X.

[185] Y. Ma, C. J. Lee, and J. S. Park, “Strategies for Optimizing the Production of Proteins and Peptides with Multiple Disulfide Bonds,” *Antibiotics 2020, Vol. 9, Page 541*, vol. 9, no. 9, p. 541, Aug. 2020, doi: 10.3390/ANTIBIOTICS9090541.

[186] M. Kesik-Brodacka, “Progress in biopharmaceutical development,” *Biotechnol Appl Biochem*, vol. 65, no. 3, pp. 306–322, May 2018, doi: 10.1002/BAB.1617.

[187] G. Bulaj, “Formation of disulfide bonds in proteins and peptides,” *Biotechnol Adv*, vol. 23, no. 1, pp. 87–92, Jan. 2005, doi: 10.1016/J.BIOTECHADV.2004.09.002.

[188] B. Gasser *et al.*, “Protein folding and conformational stress in microbial cells producing recombinant proteins: A host comparative overview,” *Microb Cell Fact*, vol. 7, no. 1, pp. 1–18, Apr. 2008, doi: 10.1186/1475-2859-7-11/FIGURES/3.

[189] C. Appenzeller-Herzog and L. Ellgaard, “The human PDI family: Versatility packed into a single fold,” *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1783, no. 4, pp. 535–548, Apr. 2008, doi: 10.1016/J.BBAMCR.2007.11.010.

[190] T. Kluyver *et al.*, “Jupyter Notebooks-a publishing format for reproducible computational workflows.,” *Elpub*, vol. 2016, pp. 87–90, 2016.

[191] W. Mckinney, “pandas: a Foundational Python Library for Data Analysis and Statistics”, Accessed: Sep. 11, 2023. [Online]. Available: http://pandas.sf.net

[192] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, “The NumPy array: A structure for efficient numerical computation,” *Comput Sci Eng*, vol. 13, no. 2, pp. 22–30, Mar. 2011, doi: 10.1109/MCSE.2011.37.

[193] M. L. Waskom, “seaborn: statistical data visualization”, doi: 10.21105/joss.03021.

[194] N. Ari and M. Ustazhanov, “Matplotlib in python,” *Proceedings of the 11th International Conference on Electronics, Computer and Computation, ICECCO 2014*, Dec. 2014, doi: 10.1109/ICECCO.2014.6997585.

[195] M. an Sun, Y. Wang, Q. Zhang, Y. Xia, W. Ge, and D. Guo, “Prediction of reversible disulfide based on features from local structural signatures,” *BMC Genomics*, vol. 18, no. 1, pp. 1–10, Apr. 2017, doi: 10.1186/S12864-017-3668-8/FIGURES/5.

[196] S. Iuchi, “Three classes of C2H2 zinc finger proteins,” *Cellular and Molecular Life Sciences*, vol. 58, no. 4, pp. 625–635, 2001, doi: 10.1007/PL00000885/METRICS.

[197] M. Fontecave, “Iron-sulfur clusters: ever-expanding roles,” *Nature Chemical Biology 2006 2:4*, vol. 2, no. 4, pp. 171–174, 2006, doi: 10.1038/nchembio0406-171.

[198] R. Morales *et al.*, “Refined X-ray structures of the oxidized, at 1.3 Å, and reduced, at 1.17 Å, [2Fe-2S] ferredoxin from the cyanobacterium Anabaena PCC7119 show redox-linked

conformational changes," *Biochemistry*, vol. 38, no. 48, pp. 15764–15773, Nov. 1999, doi: 10.1021/BI991578S.

[199] H. Sticht, G. Wildegger, D. Bentrop, B. Darimont, R. Sterner, and P. Rösch, "An NMR-Derived Model for the Solution Structure of Oxidized Thermotoga maritima 1[Fe4-S4] Ferredoxin," *Eur J Biochem*, vol. 237, no. 3, pp. 726–735, May 1996, doi: 10.1111/J.1432-1033.1996.0726P.X.

[200] C. Kaleta, S. Schäuble, U. Rinas, and S. Schuster, "Metabolic costs of amino acid and protein production in *Escherichia coli*," *Biotechnol J*, vol. 8, no. 9, pp. 1105–1114, Sep. 2013, doi: 10.1002/BIOT.201200267.

[201] M. Wheatley *et al.*, "Lifting the lid on GPCRs: the role of extracellular loops," *Br J Pharmacol*, vol. 165, no. 6, pp. 1688–1703, Mar. 2012, doi: 10.1111/J.1476-5381.2011.01629.X.

[202] L. Frooninckx *et al.*, "Neuropeptide GPCRs in *C. Elegans*," *Front Endocrinol (Lausanne)*, vol. 3, no. DEC, p. 37617, Dec. 2012, doi: 10.3389/FENDO.2012.00167.

[203] R. J. Bleichrodt, M. Hulsman, H. A. B. Wösten, and M. J. T. Reinders, "Switching from a unicellular to multicellular organization in an *Aspergillus niger* hypha," *mBio*, vol. 6, no. 2, Mar. 2015, doi: 10.1128/MBIO.00111-15.

[204] O. Tenaillon, O. K. Silander, J. P. Uzan, and L. Chao, "Quantifying Organismal Complexity using a Population Genetic Approach," *PLoS One*, vol. 2, no. 2, p. e217, Feb. 2007, doi: 10.1371/JOURNAL.PONE.0000217.

[205] A. K. Seth, E. Izhikevich, G. N. Reeke, and G. M. Edelman, "Theories and measures of consciousness: An extended framework," *Proc Natl Acad Sci U S A*, vol. 103, no. 28, pp. 10799–10804, Jul. 2006, doi: 10.1073/PNAS.0604347103.

[206] G. Bell and A. O. Mooers, "Size and complexity among multicellular organisms," *Biological Journal of the Linnean Society*, vol. 60, no. 3, pp. 345–363, Mar. 1997, doi: 10.1111/J.1095-8312.1997.TB01500.X.

[207] I. Mikhalevich, R. Powell, and C. Logan, "Is behavioural flexibility evidence of cognitive complexity? How evolution can inform comparative cognition," *Interface Focus*, vol. 7, no. 3, Jun. 2017, doi: 10.1098/RSFS.2016.0121.

[208] R. J. Taft, M. Pheasant, and J. S. Mattick, "The relationship between non-protein-coding DNA and eukaryotic complexity," *BioEssays*, vol. 29, no. 3, pp. 288–299, Mar. 2007, doi: 10.1002/BIES.20544.

[209] A. Miseta and P. Csutora, "Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms," *Mol Biol Evol*, vol. 17, no. 8, pp. 1232–1239, Aug. 2000, doi: 10.1093/OXFORDJOURNALS.MOLBEV.A026406.

[210] G. D. Barone *et al.*, "Industrial Production of Proteins with *Pichia pastoris — Komagataella phaffii*," *Biomolecules 2023, Vol. 13, Page 441*, vol. 13, no. 3, p. 441, Feb. 2023, doi: 10.3390/BIOM13030441.

[211] M. Kim, N. Rai, V. Zorraquino, and I. Tagkopoulos, "Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*," *Nature Communications 2016 7:1*, vol. 7, no. 1, pp. 1–12, Oct. 2016, doi: 10.1038/ncomms13090.

[212] B. Soufi, K. Krug, A. Harst, and B. Macek, "Characterization of the *E. coli* proteome and its modifications during growth and ethanol stres," *Front Microbiol*, vol. 6, no. FEB, p. 128127, Feb. 2015, doi: 10.3389/FMICB.2015.00103.

[213] K. Peebo, K. Valgepea, A. Maser, R. Nahku, K. Adamberg, and R. Vilu, "Proteome reallocation in *Escherichia coli* with increasing specific growth rate," *Mol Biosyst*, vol. 11, no. 4, pp. 1184–1193, Mar. 2015, doi: 10.1039/C4MB00721B.

[214] J. R. Wiśniewski and D. Rakus, "Multi-enzyme digestion FASP and the 'Total Protein Approach'-based absolute quantification of the *Escherichia coli* proteome," *J Proteomics*, vol. 109, pp. 322–331, Sep. 2014, doi: 10.1016/J.JPROT.2014.07.012.

[215] L. Arike, K. Valgepea, L. Peil, R. Nahku, K. Adamberg, and R. Vilu, "Comparison and applications of label-free absolute proteome quantification methods on *Escherichia coli*," *J Proteomics*, vol. 75, no. 17, pp. 5437–5448, Sep. 2012, doi: 10.1016/J.JPROT.2012.06.020.

[216] K. Valgepea, K. Adamberg, R. Nahku, P. J. Lahtvee, L. Arike, and R. Vilu, "Systems biology approach reveals that overflow metabolism of acetate in *Escherichia coli* is triggered by carbon catabolite repression of acetyl-CoA synthetase," *BMC Syst Biol*, vol. 4, no. 1, pp. 1–13, Dec. 2010, doi: 10.1186/1752-0509-4-166.

[217] Y. Taniguchi *et al.*, "Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells," *Science (1979)*, vol. 329, no. 5991, pp. 533–538, Jul. 2010, doi: 10.1126/SCIENCE.1188308.

[218] A. Schmidt *et al.*, "The quantitative and condition-dependent *Escherichia coli* proteome," *Nature Biotechnology 2015 34:1*, vol. 34, no. 1, pp. 104–110, Jan. 2016, doi: 10.1038/nbt.3418.

[219] Z. L. Chen *et al.*, "A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides," *Nature Communications 2019 10:1*, vol. 10, no. 1, pp. 1–12, Jul. 2019, doi: 10.1038/s41467-019-11337-z.

[220] S. Lu *et al.*, "Mapping native disulfide bonds at a proteome scale," *Nature Methods 2015 12:4*, vol. 12, no. 4, pp. 329–331, Feb. 2015, doi: 10.1038/nmeth.3283.

[221] M. S. Loos *et al.*, "Structural basis of the subcellular topology landscape of *Escherichia coli*," *Front Microbiol*, vol. 10, pp. 1–22, Jul. 2019, doi: 10.3389/FMICB.2019.01670.

[222] B. Volkmer and M. Heinemann, "Condition-Dependent Cell Volume and Concentration of *Escherichia coli* to Facilitate Data Conversion for Systems Biology Modeling," *PLoS One*, vol. 6, no. 7, p. e23126, 2011, doi: 10.1371/JOURNAL.PONE.0023126.

[223] F. Teufel *et al.*, "SignalP 6.0 predicts all five types of signal peptides using protein language models," *Nature Biotechnology 2022 40:7*, vol. 40, no. 7, pp. 1023–1025, Jan. 2022, doi: 10.1038/s41587-021-01156-3.

[224] K. D. Tsirigos, C. Peters, N. Shu, L. Käll, and A. Elofsson, "The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides," *Nucleic Acids Res*, vol. 43, no. W1, pp. W401–W407, 2015, doi: 10.1093/NAR/GKV485.

[225] C. J. Epstein, R. F. Goldberger, and C. B. Anfinsen, "The Genetic Control of Tertiary Protein Structure: Studies With Model Systems," *Cold Spring Harb Symp Quant Biol*, vol. 28, no. 0, pp. 439–449, Jan. 1963, doi: 10.1101/SQB.1963.028.01.060.

[226] R. Noiva, “Protein disulfide isomerase: The multifunctional redox chaperone of the endoplasmic reticulum,” *Semin Cell Dev Biol*, vol. 10, no. 5, pp. 481–493, Oct. 1999, doi: 10.1006/SCDB.1999.0319.

[227] A. Gąciarz and L. W. Ruddock, “Complementarity determining regions and frameworks contribute to the disulfide bond independent folding of intrinsically stable scFv,” *PLoS One*, vol. 12, no. 12, p. e0189964, Dec. 2017, doi: 10.1371/JOURNAL.PONE.0189964.

[228] N. J. Darby, P. E. Morin, G. Talbo, and T. E. Creighton, “Refolding of bovine pancreatic trypsin inhibitor via non-native disulphide intermediates,” *J Mol Biol*, vol. 249, no. 2, pp. 463–477, Jan. 1995, doi: 10.1006/JMBI.1995.0309.

[229] N. E. Chayen and E. Saridakis, “Protein crystallization: from purified protein to diffraction-quality crystal,” *Nature Methods 2008 5:2*, vol. 5, no. 2, pp. 147–153, Jan. 2008, doi: 10.1038/nmeth.f.203.

[230] Y. M. Go, J. D. Chandler, and D. P. Jones, “The cysteine proteome,” *Free Radic Biol Med*, vol. 84, pp. 227–245, Jul. 2015, doi: 10.1016/J.FREERADBIOMED.2015.03.022.

[231] E. Y. Chen, C. C. Bartlett, T. W. Loo, and D. M. Clarke, “The ΔF508 Mutation Disrupts Packing of the Transmembrane Segments of the Cystic Fibrosis Transmembrane Conductance Regulator,” *Journal of Biological Chemistry*, vol. 279, no. 38, pp. 39620–39627, Sep. 2004, doi: 10.1074/JBC.M407887200.

[232] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science (1979)*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/SCIENCE.AAA8415.

[233] D. T. Jones, “Setting the standards for machine learning in biology,” *Nature Reviews Molecular Cell Biology 2019 20:11*, vol. 20, no. 11, pp. 659–660, Sep. 2019, doi: 10.1038/s41580-019-0176-5.

[234] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities,” *Information Fusion*, vol. 50, pp. 71–91, Oct. 2019, doi: 10.1016/J.INFFUS.2018.09.012.

[235] X. He, K. Zhao, and X. Chu, “AutoML: A survey of the state-of-the-art,” *Knowl Based Syst*, vol. 212, p. 106622, Jan. 2021, doi: 10.1016/j.knosys.2020.106622.

[236] M. Haslbeck and J. Buchner, “In vitro refolding of proteins,” in *Oxidative Folding of Proteins: Basic Principles, Cellular Regulation and Engineering*, The Royal Society of Chemistry, 2018, pp. 129–151.