Technische Universität München

Lehrstuhl für Bioverfahrenstechnik

Optimization and Modeling of Protein Refolding Conditions

Bernd Rainer Kurt Anselment

Vollständiger Abdruck der von der Fakultät für Maschinenwesen der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigten Dissertation.

Vorsitzender		Univ.-Prof. Dr.-Ing. Andreas Kremling
Prüfer der Dissertation:	1.	Univ.-Prof. Dr.-Ing. Dirk Weuster-Botz
	2.	Univ.-Prof. Dr. rer. nat. Johannes Buchner

Die Dissertation wurde am 08.03.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Maschinenwesen am 26.6.2012 angenommen.

Acknowledgements

This thesis was realized at the Institute of Biochemical Engineering of the Technische Universität München with the help and support of many people.

Foremost, I would like to thank my advisor Prof. Dr-Ing. Dirk Weuster-Botz for his excellent supervision and the possibility of working on this project at his institute. Especially the freedom to develop and pursue my own ideas is greatly appreciated.

The support of the project partner (Department Chemie, Center for Integrated Protein Science) supervised by Prof. Dr. rer. nat. Johannes Buchner (who was also the co-examiner) was fundamental for this project. Especially helpful were the practical tips and discussions about refolding with Martin Haslbeck, Tetyana Dashivets and Eva Herold. Measurement data for several proteins were provided by Elisabeth Meyer and Danae Baerend.

I also would like to thank Prof. Dr. Andreas Kremling for taking over the position of the president of the jury.

This interdisciplinary research was supported financially by the International Graduate School of Science and Engineering (IGSSE) of the Technische Universität München. IGGSE also made a research stay at the Mayo Lab of the California Institute of Technology, Pasadena, CA possible.

Furthermore, my thanks goes to my students Jacqueline Fries, Veronika Schoemig, Christopher Kesten, Ludwig Klermund and Lars Janoschek for their experimental assistance.

I am very grateful to my colleagues at the Institute of Biochemical Engineering. It was a great time in an excellent working atmosphere. We also shared some good times and had fun outside the laboratory. I would like to especially thank Dirk Hebel, Hannes Link, Ilka Sührer, Martin Demler, Michael Weiner and Yilei Fu.

Finally, I would like to acknowledge my friends and my family for their vital support.

Table of Contents

1 Introduction

Proteins are one of the most important classes of biological macromolecules and involved in virtually all cellular processes. Their central role is reflected in the molecular basis of life, the cellular machinery that translates genetic information (DNA) into proteins. Nucleotide sequences exist for every protein, which contain the required information to synthesize the protein out of amino acids. Despite the modest number of amino acids (20), protein diversity is overwhelming. This variety arises from the multitude of possible combinations. A polypeptide with 60 amino acids offers 20^{60} ($1.153 \cdot 10^{78}$) different theoretical species, all unique in their sequence (Mountain *et al.*, 1999).

The function of a protein is determined by its three-dimensional structure, built by folding of the linear polypeptide sequence into a compact structure. Folding is energy driven and the native, biologically active structure is generally considered the most stable configuration under physiological conditions (Dill and Chan, 1997). Although the structure of the folded protein is encoded in its amino acid sequence (Anfinsen, 1972), protein folding is a complex problem due to the immense conformational space (Levinthal, 1969). Despite intense research, a prediction from the sequence is not feasible today and experimental studies are necessary to study structure and function.

The structural and functional diversity and the vital importance make proteins a very interesting target for science and industry. Modern molecular biology methods allow an easy and comprehensive analysis of genes and the corresponding proteins. Proteins of interest are usually produced in host organisms like *Escherichia coli*, which are easy to handle and feature well-established genetic tools. Besides the expression of genes from other organisms, which are simply transferred to the host organism, it is also possible to modify the nucleotide sequence and generate new variants. Thus, the enormous set of natural proteins is extended by a novel set of designed molecules with customized functionality in features like activity, stability or binding (protein engineering). Two classes of proteins are especially focused on for industrial applications: enzymes as natural catalysts (white biotechnology) and proteins for therapeutic applications (red biotechnology). For both, target proteins have to be produced pure, cost-efficiently and on a large scale (Mountain *et al.*, 1999; Voet and Voet, 2004).

2 Thesis Motivation and Objectives

2.1 Motivation

Protein expression in *Escherichia coli* (*E. coli)* is a standard low cost and high yield production process for recombinant proteins (Graumann and Premstaller, 2006). However, *in vivo* solubility is often limited. Approximately 40 % of the proteins overexpressed in *E. coli* are insoluble (Mayer and Buchner, 2004) and require solubilization and subsequent refolding in order to obtain the biologically active native structure. This refolding step represents a bottleneck in process development, as optimal refolding conditions have to be determined in large screening experiments (Clark, 2001; Middelberg, 2002).

Standard refolding screens described in the literature are primarily based on statistical methods (fractional factorial designs) and limited in some core characteristics. Either a limited number of buffer components that affect refolding are analyzed (Boyle *et al.*, 2009) or their interdependence is not sufficiently considered (Cowan *et al.*, 2008; Willis *et al.*, 2005). Another shortcoming of those studies is often the lack of further optimization of suitable refolding conditions (Armstrong *et al.*, 1999; Hofmann *et al.*, 1995). In addition, previous studies either do not include an analysis of the variable effects (Hofmann *et al.*, 1995; Lin *et al.*, 2006) or perform only a regression analysis of the most important variables (Armstrong *et al.*, 1999; Tobbell *et al.*, 2002; Willis *et al.*, 2005). A comprehensive model which connects the refolding success and the composition of the refolding buffer is missing.

Stochastic search methods like genetic algorithms (GA) offer the potential to combine screening and optimization in one step. In contrast to statistic screening methods, like the fractional factorial screens, they are not based on a simplified process model. GAs are able to identify optimal solutions with limited experimental effort in complex search spaces. The major advantage of stochastic search strategies lies in the multi-objective optimization of complex search spaces (Bianchi *et al.*, 2008). Protein refolding, with its variety of interacting variables (protein, pH, ionic strength, additives, redox agents), is presumed to be such a complicated problem.

2.2 Objectives

This thesis investigated the application of a stochastic search method on the problem of protein refolding. The aim was to provide a robust, standardized, one-step optimization strategy which allows an experimenter to optimize the refolding conditions in a series of parallel experiments. Acquired data should be used to model the coherence of refolding conditions and refolding yields and deduce trends.

The individual objectives of this thesis were to:

- Establish a standard experimental design approach for protein refolding based on a genetic algorithm (GA).
- Optimize the refolding conditions of a variety of well-characterized model proteins partially in cooperation with the project partner (Department Chemie, Center for Integrated Protein Science, Technische Universität München).
- Evaluate the performance of this approach and compare it to standard statistical design of experiments (DOE) strategies.
- Analyze the experimental data and build black box models that model refolding success as a function of the composition of the refolding buffer.

3 Theoretical Background

3.1 Production of recombinant proteins

3.1.1 Protein expression systems

Today, protein production is almost exclusively performed recombinantly, which means that the protein is heterologously expressed in a host organism. Because of the immense structural and functional diversity of proteins, the expression system has to adapted to the protein under study. Standard expression systems include bacteria, yeast, filamentous fungi and mammalian cell cultures. In the following sections, applications, advantages and disadvantages are outlined briefly.

Escherichia coli (E. coli) is the standard microorganism for the production of recombinant proteins. Efficient tools for genetic manipulation, high growth rates, high content of recombinant protein (up to 50 % of the dry cell mass) and cheap and easy cultivation in defined media make *E. coli* the primary choice for host organisms. Applications range from high-throughput screening to large-scale production processes (Andersen and Krummen, 2002; Graumann and Premstaller, 2006; Schmidt, 2004). However, *E. coli* lacks several eukaryotic properties with severe consequences for the expression of eukaryotic and especially mammalian proteins. Posttranslational modifications like glycosylation are challenging and recombinant proteins expressed in high titers are often prone to aggregation. Consequently, the expression of soluble protein is limited and protein aggregates, so called inclusion bodies (IBs), are often observed inside the cells (Choe *et al.*, 2006). This solubility problem poses a major challenge and will be discussed in detail in section 3.1.2.

Yeast expression systems are usually applied if the protein cannot be produced in soluble form in *E. coli* or posttranslational modifications are required. Both, *Pichia pastoris* and *Saccharomyces cerevisiae* are established host organisms that enable moderate protein titers (up to 15 g L^{-1}) with a relatively straightforward downstream processing (Cregg *et al.*, 2000; Gerngross, 2004; Graumann and Premstaller, 2006). Filamentous fungi expression systems, for example *Aspergillus niger* are largely comparable to yeasts. They offer an efficient secretion system and moderate process costs (slightly higher than *E. coli*). Consequently, extracellular proteins with disulfide bonds and other proteins, which are inadequate for expression in *E. coli,* are often produced with either yeast or

fungi. The latter are mainly used for industrial enzymes, for which costs are critical, and not for therapeutic proteins (Gerngross, 2004).

Mammalian cell cultures, specifically Chinese Hamster Ovary (CHO) cells, are almost exclusively used for therapeutic proteins due to the higher costs. Today, about 70 % of all therapeutic proteins on the market are produced in CHO cells, while most of the rest is expressed in *E. coli* (Hacker *et al.*, 2009; De Jesus and Wurm, 2011). As mammalian cell cultures are closer related to the human than the previously discussed expression systems, post-translational modifications are in general a minor problem. Recent advances in the field significantly improved both productivity (50 pg cell^{-1} day^{-1} to 60 pg cell^{-1} day^{-1}) and harvest concentrations (1 g L^{-1} to 5 g L^{-1}) for CHO processes (Hacker *et al.*, 2009).

Concluding remarks

In conclusion, *E. coli* is the primary choice for expression systems and close to ideal with respect to costs and most practical consideration. If the expression cannot be realized in *E. coli,* either yeast and fungi (enzymes) or CHO cells (therapeutic proteins) are typically studied as alternatives.

3.1.2 Expression in bacterial hosts – the issue of protein solubility

Cytoplasmic expression in *E. coli*

E. coli is the standard host organism for the expression of non-glycosylated peptides and proteins. In principle, three different expression strategies exist, each with unique advantages and disadvantages (Table 3.1).

Table 3.1: Strategies for protein expression in *E. coli* (Andersen and Krummen, 2002). (IB) inclusion body.

Expression route	Advantages	Limitations
Cytoplasmic	Highest yields, IB formation enriches the protein	IB formation makes refolding necessary
Periplasmic	Disulfide bridging, natural secretion signals	Empirical and often inefficient translocation, typically low yields
Secretory	Easy product separation, reasonably efficient for peptides	Secretion machinery not fully understood, inefficient for proteins

This thesis focuses on the cytoplasmic expression strategy, as it is used for most industrial processes. Pivotal for the wide-spread application are very high productivities and product yields. A good example is the production of human interferon-γ in high cell density cultivation. In this case, a fed-batch processes with a maximal biomass of 127 g L^{-1} cell dry weight (CDW) enabled the production of 42.5 g L^{-1} product in 17 h cultivation time with a productivity of 2.5 g L^{-1} h^{-1} and a specific yield of 0.33 $g_{product}$ g_{CDW}^{-1} (Koolaee *et al.*, 2006).

However, approximately 40 % of the proteins overexpressed in *E. coli* exhibit a low *in vivo* solubility and form IB (Mayer and Buchner, 2004). The exact mechanism of misfolding and aggregation is not clearly understood, but several factors are assumed to contribute to IB formation. One of the main factors is the difference between pro- and eukaryotic proteins. This applies to the protein size, as only 13 % of *E. coli* proteins possess more than 500 residues (roughly 500 kDa) compared to 38 % in *Saccharomyces cerevisiae* (Hartl and Hayer-Hartl, 2002). In addition, the complexity is lower for prokaryotic proteins. Multiple domains, oligomeric structure and multiple disulfide bonds are far more common in eukaryotes. The second factor that contributes to misfolding is the difference between the prokaryotic host and eukaryotes. Translational and post-translational machineries and folding modulators (chaperones and foldases) are partly or completely unalike. Finally, the reductive conditions of the bacterial cytoplasm promote misfolding and aggregation for disulfide-bridged proteins (Choe *et al.*, 2006; Graumann and Premstaller, 2006).

In the light of this problem two different process strategies are pursued.

On the one hand, it is possible to optimize the soluble expression in order to obtain more functional protein. In the last decades considerable progress was realized in this subject (Makino *et al.*, 2011). Protein expression as fusion proteins to solubilizing partners like the maltose-binding protein or glutathione-S-transferase (Cho *et al.*, 2008; Rabhi-Essafi *et al.*, 2007) or the co-expression of various chaperones or folding assistants (de Marco *et al.*, 2007) enables higher expression rates for many proteins. Furthermore, mutant strains of *E. coli* allow a more efficient expression of disulfide-bridged proteins (Bessette *et al.*, 1999) or glycosylated proteins (Wacker *et al.*, 2002), though the yields are still low.

On the other hand, many industrial processes are based on the insoluble expression of the protein of interest in IBs. (Table 3.2). Insulin and insulin analogs are probably the best-known products. In addition, various other therapeutic proteins including growth hormones, growth factors, interferons and interleukins are all produced in IBs. The

general advantages are the enrichment of the product, the protection from proteolysis and the ability to produce proteins that are toxic to *E. coli* cells (Choe *et al.*, 2006). Such aspects compensate for the additional effort required in the downstream processing. For IB based processes two additional processing steps become necessary: Solubilization and the subsequent refolding, which will be discussed in detail in section 3.2.

Table 3.2: Overview of therapeutic peptides and proteins produced in *E. coli* (Choe *et al.*, 2006; Rabhi-Essafi *et al.*, 2007).

Product	**Remarks**	**Companies**
Asparaginase	-	Merck
B-type natriuretic peptide	Inclusion bodies	Scios/Johnson & Johnson
Cholera toxin subunit B	-	SBL Vaccine
Granulocyte-colony stimulating factor	Inclusion bodies	Amgen
Human Growth Hormone	Inclusion bodies or periplamic	Genentech, Eli Lilly, Pfizer, Schwartz Pharma, Novo Nordisk
Insulin and analogs	Inclusion bodies	Eli Lilly, Aventis
Interferon alfacon-1	Inclusion bodies	Valeant
Interferon α-2a	-	Hoffmann-LaRoche, Schering
Interferon β-1b	Inclusion bodies	Schering AG, Chiron
Interferon γ-1b	Inclusion bodies	Genentec, Intermune
Interleukin 11	-	Genetics Institute
Interleukin 2	-	Chiron
Interleukin-1 receptor antagonist	-	Amgen
Parathyroid Hormone	Inclusion bodies	Eli Lilly
Pertussis toxin	-	Chiron
Salmon Calcitonin	Secretion	Unigene
Tissue Plasminogen activator	Inclusion bodies	Roche
Tumor necrosis factor alpha	-	Boehringer Ingelheim

Insoluble protein expression in inclusion bodies (IBs)

Before detailing the typical production process with IBs, it is important to characterize the properties of IBs (Figure 3.1). IBs are protein aggregates located in the cytoplasm or rarely the periplasm. The composition and the amount of IBs vary significantly. Both are influenced by the growth conditions (temperature, medium and other process parameters), the induction (system, concentration, time), the expression system and the protein of interest (Choe *et al.*, 2006). In general, IBs comprise the target protein (up to 95 %) and contaminants. For example, inclusion bodies of *β*-lactamase contained 35 % to 95 % protein of interest, 5 % to 50 % polypeptides, 1 % to 13 % phospholipids and traces of nucleic acids (Valax and Georgiou, 1993). However, in most cases washing steps enable an efficient depletion of contaminants. It was shown that most contaminants absorb onto the IBs after cell disruption. They are generally not incorporated in the IB and thus easy to remove (Clark, 2001; Middelberg, 2002; Valax and Georgiou, 1993). The physical properties of IBs again vary according to process conditions and the protein under study. In general, a size distribution between 0.35 µm and 1.28 µm (diameter) and a density between 1.034 g cm^{-3} and 1.260 g cm^{-3} are observed (Jin *et al.*, 1994; Taylor *et al.*, 1986). Consequently, a fractionation of IBs from insoluble cell debris is possible but in some cases challenging (Clark, 2001; Middelberg, 2002).

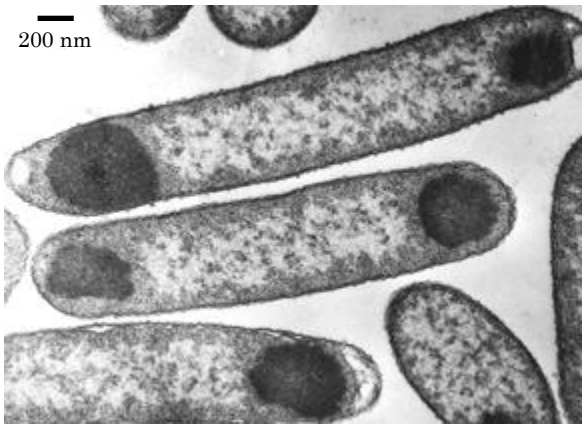


Figure 3.1: Electron micrograph of *E. coli* cells containing cytosolic inclusion bodies (http://web.mit.edu/king-lab/www/research/Scott/Scott-Research.html; Feb 2012; Betts and King, 1998;).

In the past, it was assumed that IBs contain only misfolded, inactive protein. However, recent work revealed native-like secondary structures and active protein in IBs of several proteins. Instead of being homogenous, IBs seem to consist of a distribution of misfolded and partially or fully folded protein species (Doglia *et al.*, 2008; Jevsevar *et al.*, 2005; Oberg *et al.*, 1994; Ventura and Villaverde, 2006).

Protein production processes that are based on IBs rely on the conversion from the aggregated, non-functional protein to the soluble and active form and are structured in three parts: preprocessing, solubilization and refolding (Figure 3.2). After the cultivation, cells are disrupted and soluble impurities and other (lighter) particles are depleted via centrifugation or filtration. The insoluble IBs are enriched as pellets and washed with detergents to remove other contaminants (see above). In some cases, the requirements for downstream processing may contraindicate detergent usage (Choe *et al.*, 2006). For the solubilization, the IBs are dissolved in solutions with high concentrations of chaotropes (urea or guanidine hydrochloride, Gdn·HCl). Hereby, the chaotropes break up the non-covalent bonds in the IB and the protein aggregates dissolve. Furthermore, reducing agents like dithiothreitol (DTT) are added to break up disulfide bonds via reduction. They facilitate the effective dissolution of IBs with disulfide-bridged proteins. Finally, the remaining insoluble material is removed by a fractionation step (usually centrifugation). Refolding or renaturation of the correctly folded bioactive product requires the removal of chaotropes and reducing agents. This step is critical both in process development and economic evaluations and strongly dependent on the protein (Choe *et al.*, 2006). Protein refolding will be discussed in detail in the next section.

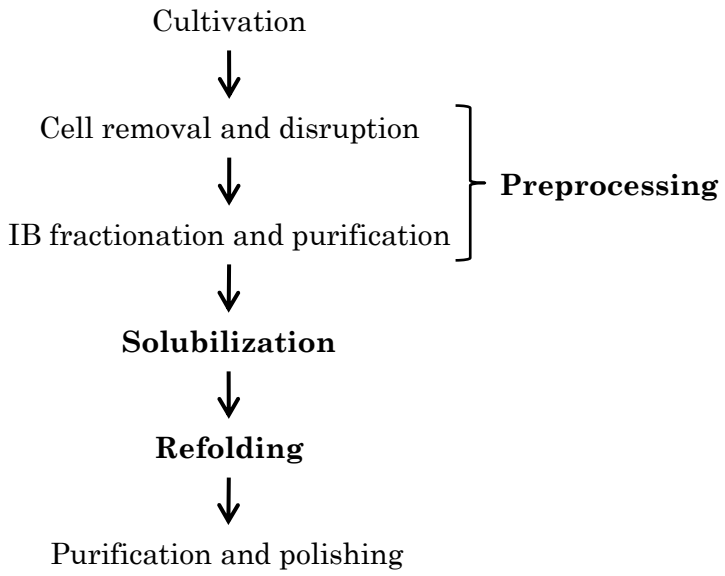


Figure 3.2: Traditional processing scheme for the production of recombinant proteins in insoluble form. Preprocessing typically involves mechanical cell disruption, centrifugation steps and preliminary purification. Subsequently, the inclusion bodies (IBs) are solubilized with chaotropes. Finally, refolding reconstitutes the native protein structure which is necessary for biological activity.

IB based processes were subject to great variety of innovations in the last decades, improving various process steps, ranging from cell disruption, IB extraction, solubilization and purification to refolding (Choe *et al.*, 2006; Crisman and Randolph, 2009; Jungbauer and Kaar, 2007; Middelberg, 2002; Qoronfleh *et al.*, 2007; Singh and Panda, 2005). Naturally, these improvements affect the competitiveness to alternative strategies and the economic performance (Freydell *et al.*, 2011; Lee *et al.*, 2006).

Concluding remarks

The expression of recombinant proteins in *E. coli* is often associated with misfolding and aggregation to IBs. IBs enable a high enrichment of the protein of interest (of up to 95 % Valax and Georgiou, 1993) and are the basis for a variety of products, especially therapeutic proteins (Graumann and Premstaller, 2006). In comparison to soluble expression, IB based processes require two additional steps: A solubilization step with chaotropes and the subsequent refolding of the native protein structure.

3.2 Protein refolding

Protein refolding or renaturation reconstitutes the native structure of the protein of interest after the IB solubilization in chaotropes. It is generally considered the bottleneck of IB based processes, since reaction conditions that enable efficient refolding drastically vary depending on the protein of interest (Jungbauer and Kaar, 2007; Lilie *et al.*, 1998). Furthermore, misfolding and aggregation represent competing side reactions that can severely reduce the refolding yield. Therefore, refolding conditions have to be optimized experimentally for each protein (Basu *et al.*, 2011; Clark, 2001; Middelberg, 2002; Rudolph and Lilie, 1996). This chapter details protein refolding and is structured in: refolding methods (3.2.1), parameters (3.2.2), reaction kinetics (3.2.3) and the analysis of folded proteins (3.2.4).

3.2.1 Refolding methods

In order to obtain biologically active proteins out of the solubilized IBs, the chaotropes have to be removed from the protein containing solution. Additionally, for oxidative protein refolding, the redox environment may be altered to enable disulfide bond formation. Both processing steps are usually combined and several different methods exist for efficient protein refolding.

Dilution

Dilution of the unfolded protein into an appropriate refolding buffer is straightforward and the simplest and most commonly used refolding method. The main applications are small-scale refolding studies and high-throughput screening experiments (Mannall *et al.*, 2009; Trésaugues *et al.*, 2004; Willis *et al.*, 2005). Large-scale dilution is also used in industry, mainly because of the simplicity of the processing scheme (Jungbauer and Kaar, 2007). After the dilution of the IBs in the refolding buffer, the solution is stirred at a controlled temperature. Subsequently, the protein is harvested after a fixed time. However, large-scale dilution has serious drawbacks in terms of the reaction vessels (large volumes, uniform mixing) and further processing (additional concentration steps).

In order to avoid aggregation and low refolding yields, it is decisive to maintain low protein concentrations. Therefore, good mixing and a slow addition of the protein containing solution are required. A final protein concentration of 10 mg L^{-1} to 100 mg L^{-1} is applied in most processes (Jungbauer and Kaar, 2007). Furthermore, a step-wise dilution of the protein, the so called pulse renaturation, is possible and often enables higher yields and final protein concentrations. Thus, pulse renaturation or other

implementations, like fed-batch or continuous processes, facilitate more economic dilution processes (De Bernardez Clark *et al.*, 1999; Katoh and Katoh, 2000; Lilie *et al.*, 1998).

Dialysis

Buffer exchange and thus removal of chaotropes is also possible through dialysis. In comparison to dilution, dialysis enables refolding with very low protein concentrations and a complete exchange of the buffer is possible. However, refolding yields can be negatively affected by non-specific protein adsorption to the dialysis membrane (West *et al.*, 1998). Furthermore, slow buffer exchange kinetics can induce aggregation of refolding intermediates, especially if the protein folding rate is low (Basu *et al.*, 2011; Tsumoto *et al.*, 2003b). Comparable to pulse renaturation for dilution, stepwise dialysis enables improved refolding yields (Tsumoto *et al.*, 2010). Due to the disadvantages of dialysis, dilution is nevertheless the preferred method for most applications.

Matrix-assisted refolding

On-column refolding provides an alternative to dilution, especially for proteins with slow refolding kinetics or a high tendency for aggregation (Jungbauer *et al.*, 2004). The immobilization of the protein on the chromatography matrix enables a spatial isolation of the proteins. Thus, intermolecular interactions of folding intermediates and consequently aggregation are reduced (Schmoeger *et al.*, 2010). Several different chromatography methods are used for on-column refolding.

Immobilized metal affinity chromatography (IMAC) refolding is based on the immobilization of the denatured protein, which has a functional tag, onto the matrix and the subsequent dilution of the denaturant to promote refolding. Hence, IMAC refolding is restricted to proteins, whose function or structure are not affected by the tag (Jungbauer *et al.*, 2004). In addition, the column material restricts the choice of refolding buffers (pH, detergents and redox agents). IMAC refolding is in particular interesting for screening applications and proteins that are difficult to refold by dilution.

Size exclusion chromatography (SEC) has been used since the 1990s for refolding (Werner *et al.*, 1994). In general, an optimization of the buffer system is necessary for efficient refolding. Furthermore, the refolding yield is dependent on the matrix composition (Fahey *et al.*, 2000; Jungbauer *et al.*, 2004). In comparison to dilution, refolding yields are often higher, but many proteins show identical performance for both methods. However, SEC incorporates a fractionation of different molecule sizes. Hence, a

depletion of contaminants is possible and advantageous for this method (Middelberg, 2002).

Ion exchange chromatography (IEC) is also used for refolding. The applications are generally comparable to SEC. However, IEC was reported to be more efficient for crude samples (Jungbauer *et al.*, 2004; Kweon *et al.*, 2004; Li *et al.*, 2003).

Furthermore, it is possible to mimic *in vivo* folding conditions by immobilization of folding catalysts onto the chromatographic support. This method might improve *in vitro* refolding yields and extend the range of proteins that can be refolded from IBs (Jungbauer *et al.*, 2004; Middelberg, 2002). However, only a few examples have been published so far (Altamirano *et al.*, 1999; Altamirano *et al.*, 2001; Tsumoto *et al.*, 2003a) and the costs of immobilized chaperonin systems and oxidoreductases hinder an industrial application.

Concluding remarks

In summary, dilution is the standard method for protein refolding (Jungbauer and Kaar, 2007). Especially for high-throughput screening experiments, the simplicity of the process outweighs the disadvantage of low protein concentrations (Mannall *et al.*, 2009; Trésaugues *et al.*, 2004; Willis *et al.*, 2005). Chromatography is an important alternative, as it enables higher final protein concentrations and yields for many proteins (Middelberg, 2002). Dialysis is mainly considered as a niche strategy.

3.2.2 Parameters in protein refolding

Protein folding normally proceeds *in vivo*, after or during translation. The refolding reaction, which is carried out *in vitro* from the solubilized protein, takes place under drastically different conditions. The protein is typically diluted in a buffered solution comprising various small molecule additives and defined redox conditions. Major differences to the cell are the absence of molecular crowding and chaperone systems and the remainder of the denaturant. Additionally, little is known about the functional relationships between refolding yield and process conditions. Therefore, process design is based on rough guidelines and the parameters have to be optimized experimentally for each protein (Basu *et al.*, 2011; Jungbauer and Kaar, 2007; Lilie *et al.*, 1998; Middelberg, 2002). The different parameters that influence protein refolding are detailed in the following.

Protein concentration and temperature

High concentrations of the protein of interest promote aggregation, hence refolding yields are decreased. Aggregation is the result of the exposure of normally inaccessible, hydrophobic core residues that become exposed on the surface of folding intermediates. If the protein concentration is high, hydrophobic interactions between these residues become more probable and aggregation occurs (Fischer *et al.*, 1993; Rudolph *et al.*, 1979). Therefore, refolding is usually performed at low protein concentrations between 10 mg L^{-1} and 100 mg L^{-1} (Jungbauer and Kaar, 2007; Lilie *et al.*, 1998) and most refolding screens are based on similar ranges (Armstrong *et al.*, 1999; Trésaugues *et al.*, 2004; Willis *et al.*, 2005). The relation of applied protein concentration and obtained refolding yield is not universal, as some proteins are more prone to aggregation and misfolding. Furthermore, the refolding method has an influence on this concentration dependency (Jungbauer *et al.*, 2004; Middelberg, 2002; Tsumoto *et al.*, 2003b).

High local protein concentrations have to be avoided in the refolding process. This is especially important during the initial phase, in which the denatured protein is diluted in the refolding buffer or loaded on the chromatography column. Therefore, mixing is an important process parameter on large industrial scales (Clark, 2001; Jungbauer and Kaar, 2007) and high-throughput screening in µL-volumes (Mannall *et al.*, 2009).

For most proteins, higher yields and less aggregation are observed at lower refolding temperatures (Mattingly *et al.*, 1995; Wang and Engel, 2009; Xie and Wetlaufer, 1996). While high temperatures seem to promote aggregation, lower temperatures decrease the folding speed and hydrophobic interactions of folding intermediates. However, it was also observed that high refolding temperatures may improve refolding for stable proteins like lysozyme (Sakamoto *et al.*, 2004). For screening experiments, protein concentration and temperature are usually standardized and kept constant at low levels (Cowieson *et al.*, 2006; Trésaugues *et al.*, 2004; Vincentelli *et al.*, 2004; Willis *et al.*, 2005).

pH of the refolding buffer

Native proteins show an increasing solubility with increasing distance from the isoelectric point (pI). The pH of the solution determines the total charge of the dissolved protein. Highly charged proteins are less prone to aggregation, as repulsive interactions raise the energy barrier for protein-protein interactions and thus for aggregation. In contrast, proteins near the pI have both negative and positive charges. An anisotropic distribution of positive and negative charges can result in dipole formation, making protein-protein interactions much more favorable (Chi *et al.*, 2003). However, guidelines

for protein solubility are not generally transferable to protein refolding. For oxidative refolding, an alkaline pH is required for the formation of thiolate ions and native disulfide bonds. Suitable conditions have to be evaluated experimentally and a prediction on the basis of the pI is not feasible. In most refolding screens, the pH is varied from slightly acidic (pH 6.0) to alkaline (pH 9.5) (Armstrong *et al.*, 1999; Cowan *et al.*, 2008; Tobbell *et al.*, 2002; Trésaugues *et al.*, 2004).

Refolding additives

Numerous additives either promote refolding by stabilizing the native structure or inhibiting aggregation. According to Hamada *et al.* (2009) additives can be grouped in three classes: **Denaturants**, including guanidine, urea, strong ionic detergents and other chaotropes, which bind to the protein (folding intermediates) and prevent aggregation. **Stabilizers**, including sugars and polyhydric alcohols (glycerol), which stabilize the native state during refolding through preferential hydration. **Mixed class** additives, which combine characteristics of denaturants and stabilizers. This group contains all other refolding additives: various detergents and non-detergent surfactants, ionic liquids, arginine, other amino acids and derivates and amphiphilic polymers like polyethylene glycol (PEG).

Prior to discussing the most important additives in the following, it is important to note that the focus of this literature review lies on the refolding application. Many of the above-mentioned additives are commonly used to stabilize proteins and suppress non-native aggregation (Chi *et al.*, 2003). This information is incorporated as the same effect (aggregation) is circumvented. However, protein refolding exhibits some differences: a background of denaturants is usually present and the starting point is the unfolded protein, not the native protein.

Guanidine hydrochloride (Gdn·HCl)

Denaturing chemicals like Gdn·HCl or urea are used for the solubilization of IBs (see section 3.1.2). They disrupt both intra- and intermolecular interactions, enabling IB solubilization and concomitant protein denaturation. If the protein is refolded via dilution, residual amounts of guanidine or urea remain. These non-denaturing residual concentrations enable the refolding of proteins that are otherwise very difficult to refold (Hevehan and De Bernardez Clark, 1997; Lilie *et al.*, 1998). The underlying mechanism is the solubilization of solvent exposed hydrophobic regions in misfolded species or folding intermediates. Both molecular dynamics simulation (O'Brien *et al.*, 2007) and

thermodynamic measurements (Arakawa and Timasheff, 1984) show, that guanidine interacts with the peptide backbone and negatively charged residues. Thus, aggregation-prone species or folding intermediates are stabilized.

Detergents, non-detergent surfactants and ionic liquids

Detergents enable higher refolding yields for many proteins (Wetlaufer and Xie, 1995; Yasuda *et al.*, 1998). The underlying mechanism is an increased solubilization of folding intermediates, as hydrophobic moieties are shielded by the detergent from the hydrophilic solvent. Consequently, aggregation is suppressed (Lilie *et al.*, 1998). The impact on refolding is strongly dependent on the protein concentration, the concentration of the detergent and the critical micellar concentration (CMC) of the detergent (Tandon and Horowitz, 1987). Strong detergents like sodium dodecyl sulfate (SDS) function as a denaturant (see Gdn·HCl). Specifically, SDS strongly binds to the protein (Takagi *et al.*, 1975), resulting in an overall negative charge. Thus, aggregation is suppressed as protein-protein interactions become energetically disfavored. Refolding conditions have to be selected carefully, as higher SDS concentrations cause denaturation, while high concentrations of other detergents are less critical.

Non-detergent surfactants, mainly non-detergent sulfobetaines, consist of a hydrophilic head group and a hydrophobic tail. However, this tail is very short compared to above-mentioned detergents. Consequently, no micelles are formed, even at concentrations of up to 1 M. Like detergents, non-detergent sulfobetaines prevent protein aggregation by interacting with folding intermediates (Vuillard *et al.*, 1998).

Ionic liquids are a recent class of refolding additives consisting of an organic cation and an either organic or inorganic anion. Most important representatives for refolding applications are N-alkyl- and N-hydroxyalkyl-N-methyl-imidazolium chlorides (Buchfink *et al.*, 2010; Lange *et al.*, 2005). Ionic liquids suppress protein aggregation and are more or less denaturing, depending on the cation. Therefore, their mode of action incorporates properties of denaturants (Gdn·HCl) and stabilizers (see below).

Cosolvent sugars and glycerol

Glycerol and several sugars act as stabilizers of the native state during refolding. Their function can be explained by their influence on the water molecules at the protein surface. Two concepts are important: preferential hydration and the Wyman linkage function (Gekko and Timasheff, 1981; Timasheff, 1998). The Wyman linkage function is the differential binding of a ligand in a two-state equilibrium, which shifts the

equilibrium towards the state with greater affinity or binding. Preferential hydration is usually interpreted as a negative binding. Protein stabilizers like glycerol are preferentially excluded from the protein surface and water molecules are enriched in this area (Figure 3.3). This can be considered as a negative binding of the cosolvents, since the surface contacts between protein and glycerol are minimized leading to higher local concentrations of water molecules near the protein surface. The exposed surface area of unfolded proteins is larger than the native state. Therefore, the degree of preferential exclusion is higher. Hence, a high negative binding of the unfolded state has the effect of favoring the native state (Chi *et al.*, 2003; Timasheff, 2002). Although entropy is most probably involved, specific changes in the entropy of the water / bounded water and the protein surface residues seem to be largely neglected and the simplified model of the preferential binding is generally used to explain the additive function (Arakawa *et al.*, 2007; Hamada *et al.*, 2009; Timasheff, 2002).

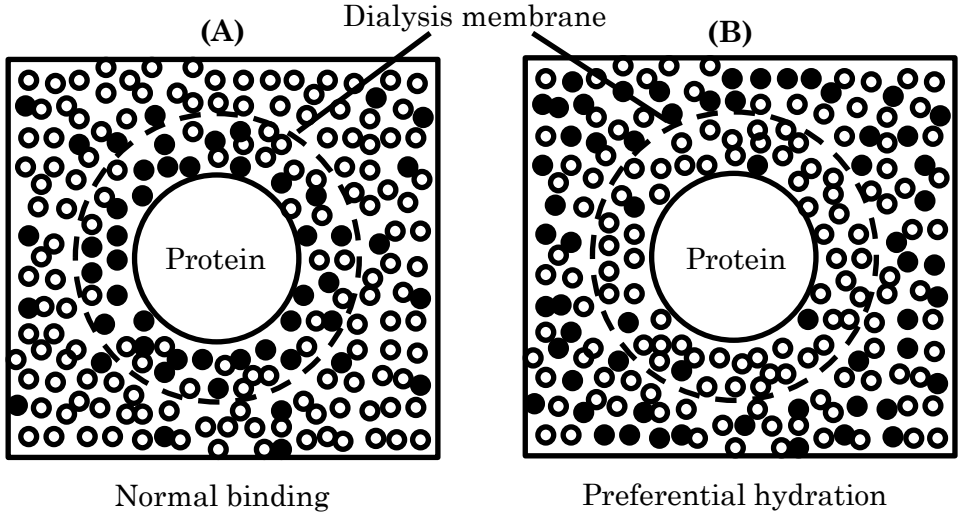


Figure 3.3: Preferential hydration. (A) normal binding of ligands to a protein. (B) preferential hydration of proteins in aqueous solutions with cosolvents like glycerol observed in dialysis equilibrium experiments (Gekko and Timasheff, 1981). (○) water molecules, (●) diffusible molecule either normal ligands (A) or glycerol (B).

Mechanistic knowledge of the effect of cosolvents on protein structure and folding is best for glycerol, that is also the most used refolding cosolvent (Phan *et al.*, 2011). According to Vagenende *et al.* (2009), the preferential hydration of proteins in glycerol-water mixtures originates from spatial orientation of glycerol molecules on the protein surface through electrostatic interactions. These interactions disfavor the larger exposed surface areas of the unfolded protein and thus bias the native state. In addition, the amphiphilic

glycerol shields hydrophobic areas and stabilizes aggregation-prone intermediates, comparable to the above-mentioned effect of detergents.

Polyethylene glycol (PEG)

Amphiphilic polymers like PEG are used for the stabilization of proteins by chemical modification (PEGylation; Roberts *et al.*, 2002) and serve as important refolding additives as well. The underlying mechanism is the preferential protein hydration (compare cosolvents). However, in contrast to the electrostatic interactions of glycerol, steric exclusion of the PEG from the protein surface is mainly responsible for this effect. Thus, the effect varies dependent on the molecular size of the PEG. For PEGs with molecular weights between 200 g mol^{-1} and 6000 g mol^{-1}, the magnitude of preferential hydration increased with increasing PEG size (Bhat and Timasheff, 1992). PEG is considered to be a mixed class additive, because it can also bind to non-polar residues.

Salts and ionic strength

Various salts act similar to the above-mentioned additives by preferential hydration. In this case, the exclusion of the salt molecules is based on the perturbation of the surface water tension. A cosolvent that increases the surface tension of water, will be depleted at the protein surface. The stabilizing effect on proteins is related to the salting-out effect described by the Hofmeister series (Kunz *et al.*, 2004).

Next to the influence on surface tension, which is generally observed at high concentrations (M), salts act as electrolytes. Hence, the ionic strength of the solution influences refolding by modulating the strength of electrostatic interactions between charged groups (Chi *et al.*, 2003). Consequently, the effects are very complex, as both intra- and intermolecular interactions between proteins are affected. Additionally, all other refolding additives with charged groups are affected, too. Combined with the impact of the pH (see above), this generates a network of very complex interactions.

Literature on the effects of the ionic strength of the refolding buffer is rather sparse. Most reviews mention ionic strength as an important factor, but experimental values are not given (Lilie *et al.*, 1998; Rudolph and Lilie, 1996; Wang, 2005). The most common salt for ionic strength variation is NaCl, which is used in concentrations of 50 mM to 500 mM for refolding experiments (Cabrita and Bottomley, 2004). Human growth hormone was reported to refold nearly independently from the ionic strength. However, only low ionic strengths of up to 200 mM were examined (Kim and Lee, 2000).

Arginine and other amino acids (mixed class)

The amino acid arginine is the most important refolding additive and commonly used for standard refolding protocols and screens in concentrations of up to 750 mM (Phan *et al.*, 2011). Arginine increases the solubility of aggregation-prone folding intermediates. Although, arginine has a guanidine group, it exhibits no denaturing or destabilizing effects on the native structure (Hamada *et al.*, 2009). Despite intense research on the mode of action, the exact mechanism is so far unsolved (Arakawa *et al.*, 2007; Tsumoto *et al.*, 2004a; Tsumoto *et al.*, 2004b). Arginine interacts with aromatic and charged protein residues and stabilizes unfolded intermediates. In addition, aqueous arginine solutions show a tendency for self-association and the formation of arginine clusters. The planar guanidine group is probably pivotal for this effect (Shukla and Trout, 2010). Other amino acids and alkyl- or amide derivates have a positive effect on refolding as well, but arginine usage is predominant (Hamada *et al.*, 2009; Phan *et al.*, 2011).

Concluding remarks on the classification of refolding additives

Refolding additives have complex effects on the protein and the solvent water (Gekko and Timasheff, 1981; Timasheff, 1998). Stabilizing agents (glycerol) mainly act through preferential hydration. In contrast, denaturants (guanidine, strong detergents) suppress aggregation mainly by the opposite effect. Denaturants bind to the protein and shield aggregation-prone hydrophobic moieties of folding intermediates. Hence, the different size (solvent accessible surface) of native (small, compact) and unfolded (large, diffuse) protein states plays an important role for both stabilizers and denaturants. Denaturants preferentially bind to the larger unfolded state. Whereas stabilizers are preferentially excluded from the protein surface, therefore, the smaller native state is favored. However, the underlying mechanisms are usually more complex and a mixture of effects is observed.

Redox agents

Proteins with disulfide bonds complicate the refolding process. Next to the correct noncovalent secondary and tertiary structure of the protein, the covalent disulfide bonds have to be reformed after IB solubilization. Correct disulfide bond formation is biased, as the native structure is generally most stable. However, proteins with many cysteine residues are difficult targets for refolding, as the number of possible combinations increases dramatically with the number of cysteine residues present (Table 3.3).

Table 3.3: Statistics of the disulfide bond (1 to j) formation in proteins with varying number of cysteine residues (1 to 2n) (Galat, 1982).

Disulfide bonds	Cysteine residues	Possible combinations, maximum of disulfide bonds	Possible combinations, partial formation allowed
1	2	1	1
2	4	3	9
4	8	105	763
8	16	2 027 025	46 306 735
j	2n	$\frac{(2n)!}{2^j \cdot (2n-2j)! \cdot (j)!}$	$\sum_{1}^{j} \frac{(2n)!}{2^j \cdot (2n-2j)! \cdot (j)!}$

IB solubilization is routinely performed under reductive conditions, in order to dissolve possible wrongly formed disulfide bonds in the IB. Hence, upon disulfide bridge formation, an oxidation of the cysteine residues is necessary (Figure 3.4).

$$RSH + RSH \rightarrow RSSR + 2\,H^+ + 2\,e^-$$

Figure 3.4: Oxidation for disulfide bond formation.

Although it is possible to use molecular oxygen for this oxidation, the yields for air oxidation are very low (Sela *et al.*, 1957). Therefore, thiols with low molecular weight are usually added to the refolding buffer. Common reagents include: reduced and oxidized glutathione (GSH, GSSG), cysteine and cystine, 2-mercaptoethanol, dithiothreitol (DTT) and tris-carboxyethyl-phosphine (TCEP). Typical molar ratios vary between 1:1 and 1:10 for the reduced and oxidized form, respectively (Lilie *et al.*, 1998). Thiols enable a rapid reshuffling of disulfide bonds, as the thiol-disulfide exchange is fast and reversible (Figure 3.5). Hence, redox agents often increase the yield of correct protein disulfide formation (Rudolph and Lilie, 1996). An alkaline pH is necessary for thiol-disulfide exchange, as the reaction mechanism is based on a nucleophilic attack of the thiolate anion.

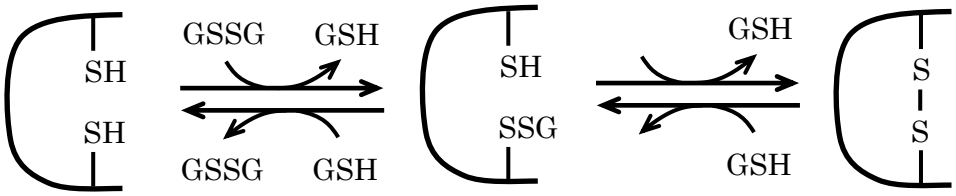


Figure 3.5: Disulfide bond formation and oxido-shuffling with glutathione in the reduced (GSH) and oxidized state (GSSG) (Voet and Voet, 2004).

Chaperones

The impact of folding catalysts was already mentioned in case of column chromatography using immobilized chaperones, which effectively mimics *in vivo* conditions (see 3.2.1). Chaperones and other folding helpers are also used for dilution experiment as supplements to the refolding buffer. They often increase refolding yields and enable refolding of challenging proteins (Mayer and Buchner, 2004; Schwarz *et al.*, 1996; Vallejo and Rinas, 2004). The bacterial GroEL / GroES chaperone complex is the most common system for *in vitro* refolding applications (Ayling and Baneyx, 1996). However, the high costs of chaperones hinder an industrial application (Jungbauer and Kaar, 2007).

REFOLD database

The REFOLD database (http://refold.med.monash.edu.au; Feb 2012; Amin *et al.*, 2006; Buckle *et al.*, 2005;) is a repository for refolding data with the information of approximately 1100 refolding experiments. Experimental data are extracted from the primary literature and dependent on contributors. Thus, data quality and filtering are an issue. Nevertheless, the database offers far more information than literature reviews and is especially valuable for the design of refolding screens. For this thesis, the possibility to extract quantitative data about the refolding buffer composition was especially valuable.

Concluding remarks

A variety of parameters affect refolding by either suppressing aggregation or stabilizing folding intermediates or the native structure. The underlying mechanisms are only roughly understood. Interdependencies and the requirements for disulfide-bridged proteins further complicate the picture and hinder a prediction of suitable refolding conditions.

3.2.3 Refolding – a kinetic competition between folding and aggregation

Refolding of the denatured protein to the biologically active, native state requires the formation of secondary, super-secondary, tertiary and quaternary structures (for oligomers) out of the denatured, highly flexible polypeptide chain. Although the native structure is encoded in its amino acid sequence (Anfinsen, 1972) and is generally the most stable structure under physiological conditions (Dill and Chan, 1997), protein folding is a complex problem due to the immense conformational space (Levinthal, 1969).

While exact mechanisms and pathways of protein folding remain controversial, it was proven that most proteins undergo different intermediate conformations before achieving their native structure (Dill *et al.*, 2008; Lindorff-Larsen *et al.*, 2011; Onuchic and Wolynes, 2004; Sosnick and Hinshaw, 2011). These intermediate states are more or less unstable and subjected to nonspecific hydrophobic interactions and incorrect interactions of partially structured regions. Hence, aggregation may occur, which is generally considered a second (or higher) order reaction. This poses the central problem for *in vitro* refolding: Because cellular chaperone systems are absent, folding intermediates readily aggregate without supplementation of refolding additives (Jungbauer *et al.*, 2004).

In vitro refolding is a competition between the correct folding pathway and misfolding and aggregation (Figure 3.6). Intermediate (I) formation and correct folding (N) are typically described as a first order reaction, while the aggregation (A) has a higher order (Kiefhaber *et al.*, 1991).

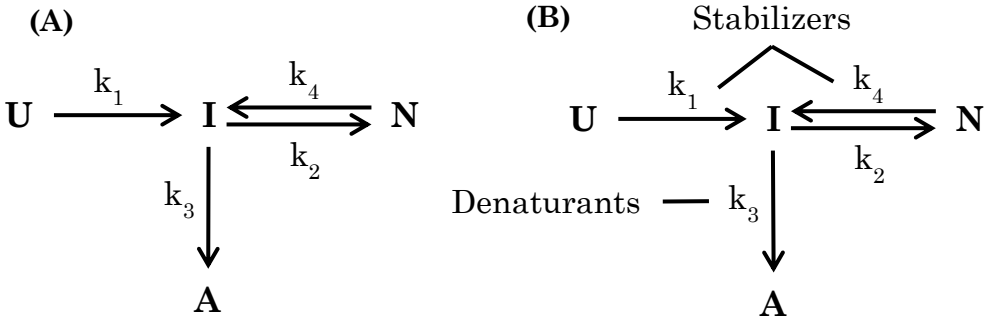


Figure 3.6: (A) Refolding kinetics with unfolded (U), intermediate (I), aggregated (A) and native (N) protein and rate constants (k_i) (Kiefhaber *et al.*, 1991). (B) Schematic influence of refolding additives. Stabilizers like glycerol stabilize the native state, while denaturants like guanidine prevent aggregation of folding intermediates (denaturing and destabilizing effects of denaturants are omitted for clarity).

The model can be further simplified for practical purposes. Typically, intermediate formation from the unfolded state is considered instantaneous. Thus, k_1 (compare Figure 3.6) is neglected. In addition, the reverse reaction from the native state (k_4) is also neglected ($k_2 >> k_4$). Therefore, the simplified model is a straightforward competition between first and higher order reactions. For batch processes refolding can be described by the following equations.

Unfolded (U) and native (N) protein concentrations can by described as:

$$\frac{dU}{dt} = -(k_2 U + k_3 U^n) \qquad \textit{(Equation 1)}$$

with	U	*unfolded protein*
	t	*time*
	n	*reaction order*
	k_2, k_3	*folding, aggregation rate constants*

$$\frac{dN}{dt} = k_2 U \qquad \textit{(Equation 2)}$$

with	N	*native protein*

For a second order (n = 2) aggregation reaction, the refolding yield (Y) is given by:

$$Y(t) = \frac{k_2}{U_0 k_3} ln\left(1 + \frac{U_0 k_3}{k_2}(1 - e^{-k_2 t})\right) \qquad \textit{(Equation 3)}$$

with	Y	*refolding yield*
	U_0	*unfolded protein, initial concentration*

Thus, the final refolding yield (t approaches infinity) is:

$$Y = \frac{k_2}{U_0 k_3} ln\left(1 + \frac{U_0 k_3}{k_2}\right) \qquad \textit{(Equation 4)}$$

Hence, the refolding yield depends on the initial concentration of the unfolded protein and the rate constants for folding and aggregation. The difference of the reaction orders results in a drastic decrease of refolding yields at higher protein concentration, as described previously (see section 3.2.2).

3.2.4 Analysis of folded proteins

In order to evaluate refolding yields, sensitive analytical methods are required which quantify the correctly folded protein. These may either be based on structural or functional features and have to be able to differentiate between folded and misfolded or aggregated protein. Especially for refolding screens, which are designed to evaluate and optimize refolding of a variety of different proteins, the correct quantification is a basic concern (Basu *et al.*, 2011; Middelberg, 2002).

Structure-based methods

Several structure-based methods provide exact data on the folding state, but are not suitable for high-throughput refolding screens. Instead, they are mainly used for stability and folding studies: Intrinsic protein fluorescence is limited to proteins with internal tryptophan residues. Furthermore, quenching effects of buffer components are problematic and fluorescence spectra of the native protein have to be available (Royer, 2006). Circular dichroism spectroscopy (CD) uses the differential absorption of circularly polarized light to investigate the secondary structure of proteins. However, the application is limited to pure protein samples and a high-throughput application in screens is not feasible. Limited proteolysis is based on the higher stability of compact native protein structures against photolytic cleavage. Native protein stability and cumbersome fragment analysis are major drawbacks (Heiring and Muller, 2001). Other methods like nuclear magnetic resonance (NMR) or sophisticated spectroscopy coupled with detailed analysis of spectra might be suitable future methods. However, they are not readily applicable today (Balbach *et al.*, 1995; Middelberg, 2002).

In contrast to above-mentioned methods, a variety of techniques are technologically fully developed and established for large-scale refolding screens. Absorbance, light scattering or turbidity measurements provide information about protein solubility, thus enabling a quantification of protein aggregation (Basu *et al.*, 2011; Middelberg, 2002). Several refolding screens use this as a first analytical step. In a second step, positive refolding conditions are subject to a more detailed analysis with another method to verify correct folding (Dechavanne *et al.*, 2011; Scheich *et al.*, 2004; Willis *et al.*, 2005). Reverse phase high-performance liquid chromatography (HPLC) and hydrophobic interaction HPLC detect the surface hydrophobicity differences of native and misfolded protein by different retention times. Due to the serial mode, data analysis can be time consuming. In addition, native protein has to be available for comparison and the resolution is limited if

many misfolded species occur. Nevertheless, this method is used in several refolding screens (Boyle *et al.*, 2008; Boyle *et al.*, 2009; Cowan *et al.*, 2008).

Function or specific binding based methods

Protein specific assays based on enzymatic activity and immuno- or bioassays provide very reliable information on protein folding (Middelberg, 2002). In comparison to structure-based methods, functional assays are generally rather simple to automate. Furthermore, most assays are either already established in 96-well plate scale or can be easily parallelized. For these reasons, functional assays are the method of choice for most refolding screens (Armstrong *et al.*, 1999; Hofmann *et al.*, 1995; Mannall *et al.*, 2009; Willis *et al.*, 2005). However, each protein of interest requires a suitable enzymatic assay or antibody. Especially for therapeutic proteins, the biological activity is the overall decision criteria and often the exclusive optimization criteria. However, regulations for therapeutic protein often demand a combination of methods, for example bioassays and turbidity measurements, to quantify aggregation which is not accessible by functional assays (Jungbauer and Kaar, 2007).

Concluding remarks

Structure-based methods have the potential advantage of a wide applicability for a high number of proteins. The application for high-throughput refolding screens is usually limited to solubility measurements or HPCL methods. On the other hand, functional assays provide reliable information about protein structure and the final criteria for protein applications (enzymes or therapeutics) is also enzymatic- or biological activity. However, these methods are protein-specific and development time for new target proteins has to be taken into account. Finally, a combination of analytical steps (aggregation measurement and activity) is required for therapeutic proteins.

3.2.5 Model proteins for refolding – overview of the analyzed proteins

Refolding screens from the literature are not standardized regarding the proteins under study. While almost all screens include lysozyme (LYZ) as a well-characterized model protein, other proteins differ from screen to screen (Armstrong *et al.*, 1999; Hofmann *et al.*, 1995; Willis *et al.*, 2005). This chapter briefly outline the six proteins which were optimized within the scope of this thesis.

Green fluorescent protein from *Aequorea victoria* (GFP)

Green fluorescent protein (GFP) constitutes an important reporter and biosensor in molecular biology (Chalfie *et al.*, 1994). GFP is a monomeric protein with a rather small molecular mass of 28 kDa and a pI of 5.7. Its distinctive feature is the intrinsic fluorescence under exposure to blue light. The chromophore (p-hydroxy-benzylideneimidazolidone) is located in the center of an 11-stranded beta barrel which is illustrated in Figure 3.7. Chromophore formation proceeds autocatalytically during folding. Refolding was examined for the engineered enhanced GFP (variant F64L and S65T, Topell *et al.*, 1999) which is more stable than the wild-type.

Green fluorescent protein

from *Aequorea victoria*

PDB 1EMA; UniProt P42212

28 kDa, monomer, pI 5.7

no disulfide-bridges

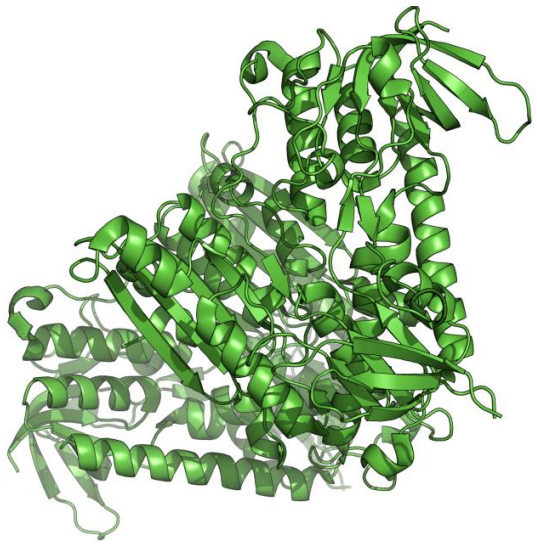
intrinsic fluorescence

Figure 3.7: Structure (Ormö *et al.*, 1996) and key data of GFP. PDB (Protein Data Bank; http://www.rcsb.org; Feb 2012; Berman *et al.,* 2000), UniProt (Universal Protein Resource; http://www.uniprot.org; Feb 2012; The Uniprot Consortium, 2012).

Glutathione reductase from *Saccharomyces cerevisiae* (GLR)

Glutathione reductase (GLR) from *Saccharomyces cerevisiae* exhibits a molecular mass of 53 kDa and a pI of 7.7 (Collinson and Dawes, 1995). The protein contains three distinctive domains and is active as a dimer (Yu and Zhou, 2007). GLR plays an important role in cytoplasmic and mitochondrial redox regulatory systems. The flavo-oxidoreductase (EC 1.8.1.7) reduces oxidized glutathione (GSSG) to the reduced form (GSH) with nicotinamide adenine dinucleotide phosphate (NADPH) as electron donor and flavin adenine dinucleotide (FAD) as coenzyme. The enzyme folds as a 3-layer(bba) sandwich (Yu and Zhou, 2007) (Figure 3.8).

GLR activity is influenced by the redox environment and various metal ions including Zn^{2+}. The active site comprises a redox-active disulfide bond. Hence, GLR activity is quite sensitive to the stated changes to the redox environment. However, ethylenediaminetetraacetic acid (EDTA) was reported to regenerate GLR activity after treatment with Zn^{2+} (Tandoğan and Ulusu, 2007).



Glutathione reductase

from *Saccharomyces cerevisiae*

PDB 2HQM; UniProt P41921

53 kDa, dimer, pI 7.7

no disulfide-bridges

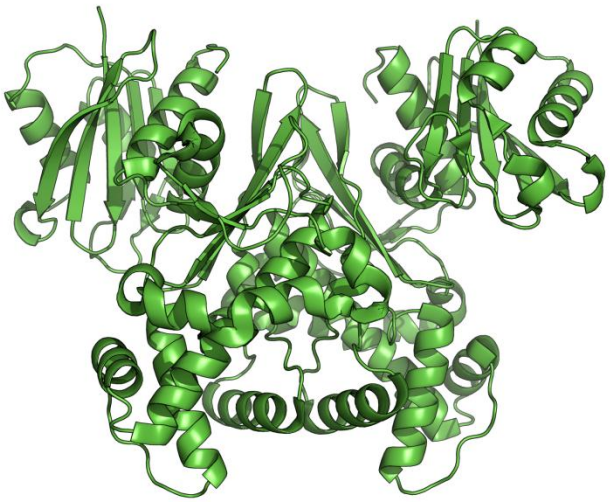
oxidoreductase (EC 1.8.1.7)

Figure 3.8: Structure (Yu and Zhou, 2007) and key data of GLR. PDB (Protein Data Bank; http://www.rcsb.org; Feb 2012; Berman *et al.,* 2000), UniProt (Universal Protein Resource; http://www.uniprot.org; Feb 2012; The Uniprot Consortium, 2012).

Glucokinase from *Escherichia coli* (GLK)

Intracellular glucose in *E. coli* is phosphorylated to glucose-6-phosphate by the enzyme glucokinase (GLK) (EC 2.7.1.2) in the first step of the glycolysis. GLK is not considered essential for the *E. coli* metabolism as glucose is transported into the cell as glucose-6-phosphate (phospho-transferase system). However, GLK plays an important role in the regulation of the carbohydrate metabolism. GLK is a homodimeric protein with a mass of 35 kDa and a pI of 6.1 (Meyer *et al.*, 1997). Each monomer folds into two distinct domains with the active site located in a cleft in between (Lunin *et al.*, 2004) (Figure 3.9).

Compared to closely related hexokinases (EC 2.7.1.1) which phosphorylate various sugars, the substrate specificity of GLK is narrow and limited to glucose. The activity of GLK depends on adenosine-triphosphate (ATP) and Mg^{2+} which is typical for kinases (Meyer *et al.*, 1997).



Glucokinase

from *Escherichia coli*

PDB 1Q18; UniProt P0A6V8

35 kDa, dimer, pI 6.1

no disulfide-bridges

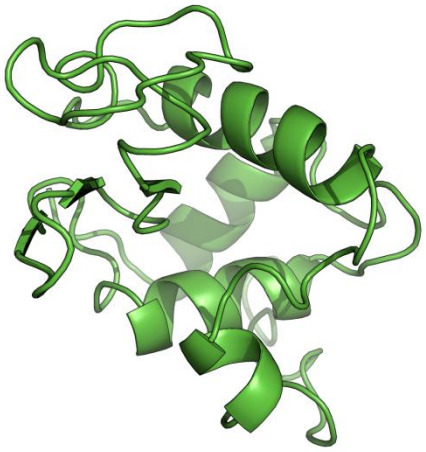
kinase (EC 2.7.1.2)

Figure 3.9: Structure (Lunin *et al.*, 2004) and key data of GLK. PDB (Protein Data Bank; http://www.rcsb.org; Feb 2012; Berman *et al.,* 2000), UniProt (Universal Protein Resource; http://www.uniprot.org; Feb 2012; The Uniprot Consortium, 2012).

Lysozyme from *Gallus gallus* (LYZ)

The well-characterized lysozyme (LYZ) is a disulfide-bridged protein with a small molecular mass of 14 kDa and an exceptionally high pI of 9.3 (Diamond, 1974). Its biological function is primarily bacteriolytic: LYZ exhibits a strong antimicrobial effect as the monomeric enzyme hydrolyses (EC 3.2.1.17) peptidoglycan linkages between N-acetylmuramic acid and N-acetyl-D-glucosamine residues present in microbial cell walls. Thus, the structural integrity of the bacterial cell wall is disturbed. LYZ folds as a compact orthogonal bundle (Rypniewski *et al.*, 1993) (Figure 3.10).

LYZ is an extracellular enzyme which exhibits a high stability. The protein shows activity over a broad pH range with an optimum at slightly acidic conditions (pH 5.5 to pH 6.0) (Xu *et al.*, 2004). Higher ionic strength of the reaction buffer decreases the activity (Davies *et al.*, 1969).



Lysozyme

from *Gallus gallus*

PDB 132L; UniProt P00698

14 kDa, monomer, pI 9.3

4 disulfide-bridges

hydrolase (EC 3.2.1.17)

Figure 3.10: Structure (Rypniewski *et al.*, 1993) and key data of LYZ. PDB (Protein Data Bank; http://www.rcsb.org; Feb 2012; Berman *et al.,* 2000), UniProt (Universal Protein Resource; http://www.uniprot.org; Feb 2012; The Uniprot Consortium, 2012).

Lactate dehydrogenase from *Oryctolagus cuniculus* (LDH)

Lactate dehydrogenase (LDH) (EC 1.1.1.27) catalyzes the conversion of pyruvate to lactate. For eukaryotes the main function of LDH is the recycling of oxidized nicotinamide adenine dinucleotide (NAD^+) in the presence of oxygen limitations (Pineda *et al.*, 2007). The analyzed LDH from *Oryctolagus cuniculus* muscle is a tetrameric protein with a monomer mass of 36 kDa and a pI of 8.2 (Sass *et al.*, 1989). A crystal structure is not available for this subtype of LDH. Hence, another closely related LDH (LDH from human heart) is illustrated in Figure 3.11.

In vivo activity of LDH is influenced by the substrate pyruvate and the product lactate as well as ascorbate (Stambaugh and Post, 1966). The enzyme is most stable at neutral pH and low temperatures (Zheng *et al.*, 2004).

Lactate dehydrogenase

from *Oryctolagus cuniculus*

PDB 1I0Z; UniProt P13491

36 kDa, tetramer, pI 8.2

no disulfide-bridges

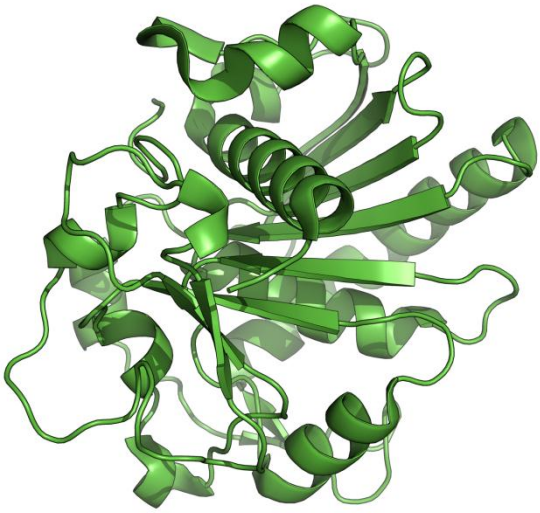
oxidoreductase (EC 1.1.1.27)

Figure 3.11: Structure of the LDH from human heart (Read *et al.*, 2001) and key data of the LDH from rabbit muscle. PDB (Protein Data Bank; http://www.rcsb.org; Feb 2012; Berman *et al.,* 2000), UniProt (Universal Protein Resource; http://www.uniprot.org; Feb 2012; The Uniprot Consortium, 2012).

Lipase from *Thermomyces lanuginosus* (LIP)

The lipase from *Thermomyces lanuginosus* (LIP) is one of the most important industrial enzymes. It is mainly applied in washing agents to remove oils and fats from fabrics (Brzozowski *et al.*, 2000; Jaeger and Reetz, 1998). LIP was one of the first enzymes subjected to intensive protein engineering (Danielsen *et al.*, 2001) resulting in more stable variants which are commercially available (trade names LipolaseUltra and LipoPrime, Novozymes). In this thesis, the wild type LIP (Lipolase, Novozymes) was characterized with regard to refolding.

LIP is a disulfide-bridged monomeric protein with a mass of 29 kDa and a pI of 5.0. The hydrolytic enzyme (EC 3.1.1.3) cleaves triglycerides into glycerol and fatty acids. LIP features an active center which is covered by an α-helical lid (Figure 3.12). This "closed state" of the enzyme is stable in aqueous solution. For catalysis, the lid must be displaced to allow the substrate access to the active center (Ollis et al., 1992). The activation proceeds quickly in the presence of a partially hydrophobic environment: LIP activity increases dramatically at the oil-water interface, a phenomenon known as interfacial activation (Derewenda *et al.*, 1994).



Lipase

from *Thermomyces lanuginosus*

PDB 1TIB; UniProt O59952

29 kDa, monomer, pI 5.0

3 disulfide-bridges

hydrolase (EC 3.1.1.3)

Figure 3.12: Structure (Derewenda *et al.*, 1994) and key data of LIP. PDB (Protein Data Bank; http://www.rcsb.org; Feb 2012; Berman *et al.,* 2000), UniProt (Universal Protein Resource; http://www.uniprot.org; Feb 2012; The Uniprot Consortium, 2012).

3.3 Experimental design strategies

Optimization of experimental problems is a challenging task in both engineering and science. In principle, two different experimental design strategies exist: statistic and stochastic (heuristic) methods. Both aim for an efficient and precise identification of optimal solutions inside the problem specific search space. This subchapter introduces both strategies and details standard designs and algorithms.

3.3.1 Statistical design of experiments (DOE)

Statistic experimental design was established in the 1920s by Ronald Fisher (Fisher, 1971). Next to the three principles of randomization, replication and blocking, he introduced the factorial designs. Response surface methodology (RSM) was the next developmental step in the 1950s. Afterwards, the application of design of experiments (DOE) spread from the agricultural sciences to industry and engineering. Today DOE is widely used, both in the commercial sector and academia (Montgomery, 2009). In addition, DOE constitutes an integral part of quality by design principles, which are applied for product quality control in industrial production processes (Lasky and Boser, 1997; Lionberger *et al.*, 2008).

Statistical DOE is based on a process model (Figure 3.13), which is approximated by more or less complex equations. DOE generally aims to optimize this process with respect to the output, the so called response variables (Y). Examples of Y are yields, product concentrations or costs. While some process variables (or factors) are not controllable (z_i) and thus kept constant, the other variables (x_i) are varied in order to obtain an optimized response.

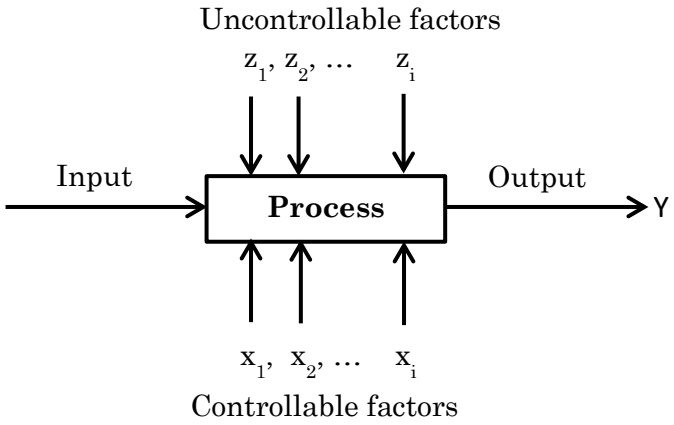


Figure 3.13: General process model with in- and output (Y) and influencing factors. Factors are either controllable (x_i) or uncontrollable (z_i) (Montgomery, 2009).

Depending on the problem, the method for experimental design varies greatly. First, the process or problem of interest is examined and important factors (controllable and significant effect on response) are selected. For simple problems with only one factor influencing the response, univariate methods are pursued. However, most problems show more complexity and many factors need to be considered. Here, the experimenter has the choice between univariate or multivariate DOE approaches (Figure 3.14).

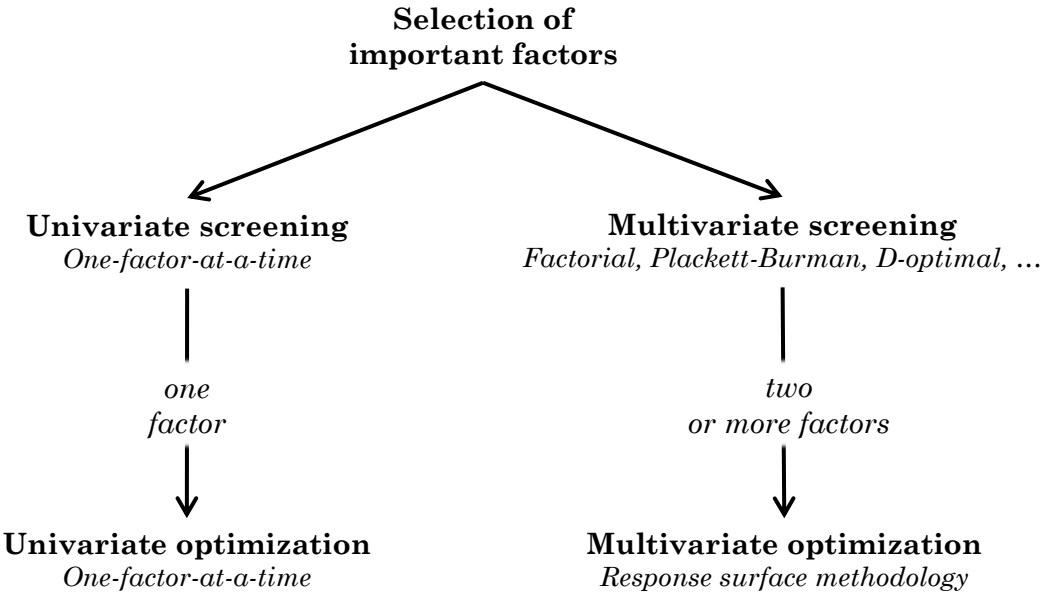


Figure 3.14: Overview of statistical DOE strategies with univariate (left) and multivariate methods (right).

In the classic one-factor-at-a-time approach a series of straightforward univariate optimizations are carried out. One factor is varied, subsequently this factor is fixed at the optimum and the next factor is varied in turn. Consequently, interactions between factors are not considered and only a small part of the experimental space is sampled. Therefore, obtaining a global optimum is not assured and strongly dependent on the initial conditions. Furthermore, the number of required experiments is higher compared to multivariate methods (Montgomery, 2009).

Multivariate strategies vary several factors simultaneously. Thereby, more information can be obtained in less experiments. The advantage of analyzing multiple factors is illustrated in Figure 3.15. On the left side (Figure 3.15, A), the one-factor-at-a-time approach fails to detect the global optimum because of the interaction between both variables. On the right side (Figure 3.15, B), a simultaneous variation of both variables

in a factorial design reveals the direction of the global optimum. This optimum can be further approximated in a subsequent experiment.

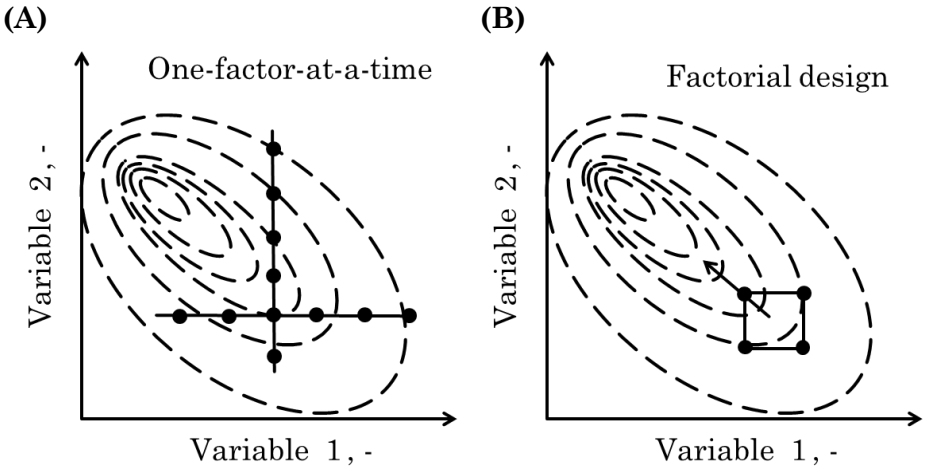


Figure 3.15: Comparison between univariate and multivariate DOE strategies on a problem with two interacting variables. The optimum is illustrated with a contour plot. (A) one-factor-at-a-time, (B) factorial design, (●) planned experiments.

This chapter focuses on simultaneous DOE. Alternatives like the simplex method (Nelder *et al.*, 1965) will not be detailed here. Simultaneous DOE is characterized by an experimental setup with a number of predefined experiments, that are performed in parallel at the same time, the so called experimental design. After the experimental evaluation, the results are statistically evaluated and if necessary, an additional experiment is planned. Statistical DOE is structured into two parts (compare Figure 3.14): In a first screening experiment, a large number of factors are evaluated in relatively few experiments. In the process, variable interactions are usually neglected to minimize experimental effort. Afterwards, a statistical analysis identifies the most significant factors. If only one variable is important, an univariate optimization is applied subsequently. Otherwise, multiple variables and their interactions are analyzed, typically with response surface methodology (RSM).

Screening

Screening methods are mainly based on two-level factorial designs illustrated in Figure 3.16. Factors are varied in two levels (coded −1 and 1) with the shape of the DOE resembling a quadrat (two factors) or cube (three factors).

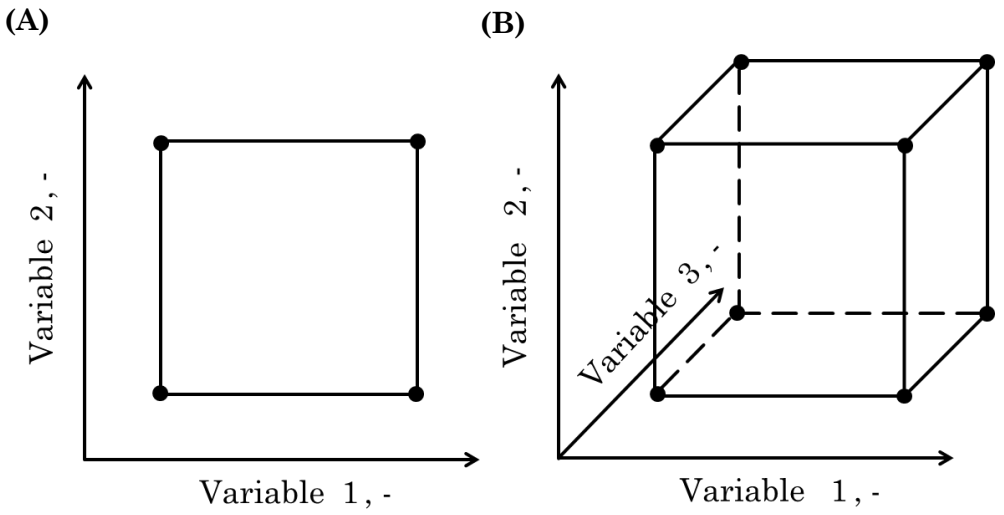


Figure 3.16: Two-level full factorial designs. (A) Two factors 2^2, (B) three factors 2^3, (●) planned experiments.

Full factorial designs

Full factorial designs contain all possible combinations between the factors (f) and their levels (L). All planned experiments (●) of the factorial design are evaluated (see Figure 3.16). This enables an estimation of both main effects (the factors) and their interactions. However, the large experimental effort for processes with many important factors (Equation 5), severely restricts the application for screening purposes. Full factorial designs are, however, the base for most RSM designs (see optimization section). An example for a two-level design is given in the appendix (Table 9.1). In this design, 8 experiments are necessary to examine three factors.

$$N = L^f = 2^f \qquad \textit{(Equation 5)}$$

with	N	*number of experiments*	-
	f	*factors*	-
	L	*level (two is standard for screening methods)*	-

Fractional factorial designs

For screening purposes, only a fraction of the above-mentioned full factorial design is evaluated experimentally. Consequently, some information is lost and not all main and interaction effects can be estimated separately. Therefore, in most cases interactions are neglected for analysis. The experimental effort is much smaller compared to full factorial designs (Equation 6).

$$N = 2^{f-v} \qquad \textit{(Equation 6)}$$

with v *fraction of the full factorial* -

An example for a fractional factorial two-level design is given in the appendix (Table 9.2). Here, 8 experiments are necessary to examine four factors. Thus, half of the experiments of the full factorial design are evaluated.

Plackett-Burman designs

Plackett-Burman designs are a derivate of fractional factorial designs developed by Plackett and Burman (1946). These designs display a very low experimental effort, as k factors can be studied in N = k + 1 experiments. Hence, they are ideal for large screening experiments. An example for a Plackett-Burman design with 7 factors is depicted in the appendix (Table 9.3). Here 8 experiments are necessary to analyze 7 factors. Thus, the experimental effort is very low compared to the $2^7 = 128$ required experiments for the full factorial design (Equation 5). The method to construct Plackett-Burman designs with a differing number of factors is detailed in Montgomery (2009).

The loss of information due to the limited number of performed experiments is an embedded disadvantage of these design. In contrast to normal fractional factorial designs, Plackett-Burman designs cannot be represented as cubes and often depict messy alias structures (Montgomery, 2009). For practical purposes, so called dummy factors are often defined in order to obtain an estimation of experimental error variance. Therefore, the number of necessary experiments will be slightly larger. Typically N = k + 4 experiments are needed (Weuster-Botz, 2000).

Factorial designs with mixed levels

Naturally, real-world problems often demand modifications to the previously detailed theoretical designs. A common issue are processes, in which one factor needs to be varied in more than two levels. This can be accomplished by combining two-level factors into one overall factor (Table 3.4). However, this approach is only straightforward and simple to use for full factorial designs. Mixed fractional factorial or Plackett-Burman designs should be used very carefully, as alias matrices get more complicated and the relative variance of factors can pose problematic (Montgomery, 2009).

Table 3.4: Mixed-level designs. The use of two-level factors to form a three-level factor (Montgomery, 2009).

Two-level factor		**Three-level factor**
−	−	x_1
+	−	x_2
−	+	x_2
+	+	x_3

D-optimal designs

D-optimal designs are model-specific and thus able to address some of the limitations of the previously discussed design types. In essence, knowledge of the experimental domain can be readily integrated into the optimization: It is possible to generate designs with custom models, in which some factors interact and others do not. Furthermore, designs with mixed levels (compare Table 3.4) are easier to realize and statistically more sound, as the relative factor variances of the optimized model are equal in most cases (Montgomery, 2009).

For screening purposes, D-optimal designs are based on a linear regression model with first order terms (main effects) (Equation 7). They examine k factors in $N \geq k + 1$ experiments. Hence, the experimental effort is comparable to Plackett-Burman designs.

$$y = b_0 + \sum b_i x_i \qquad \textit{(Equation 7)}$$

with	y	*response variable*	-
	x_i	*input variables*	-
	b_0, b_i	*zero and first order coefficients*	-

D-optimal designs are generated using a search algorithm and not based on orthogonal matrices. In a first step, an initial design matrix X is generated. Afterwards, an iterative search algorithm minimizes the variance of the model regression coefficients (covariance). This is equivalent to maximizing the determinant $D = |X^TX|$, where X is the design matrix of model terms (columns) evaluated at the different experimental conditions (rows). Most algorithms either exchange entire rows or single elements of X. Both, in the initial design generation and in the incremental change of the search algorithm, random effects are observed. Consequently, parameter estimates may be locally, but not globally, D-optimal. Most publications recommend running the design algorithms multiple times and then selecting the best design. Additionally, unlike the previously discussed designs, D-optimal designs are not based on orthogonal design matrices. Therefore, parameter estimates may be correlated (Dejaegher and Heyden, 2011; Montgomery, 2009).

Supersaturated designs

All above-mentioned experimental designs are saturated, that is k factors are examined in $N > k + 1$ experiments. Another recent class of DOE uses even less experiments, hence they are commonly called supersaturated designs (Sun *et al.*, 2011). Supersaturated designs contain the absolute minimum of necessary experiments. Consequently, even their main effects are confounded and cannot be estimated unconfounded anymore. Supersaturated designs are only sensible with regards to very large screening experiments. In this case the "sparsity of effect principle" often applies: Most of the examined factors have no significant impact on the response, especially in large screening experiments with very many factors. Supersaturated design construction is controversial, in general they are either generated from heuristic local search algorithms or are based on one of the above-mentioned designs (Dejaegher and Heyden, 2011; Sun *et al.*, 2011).

Optimization

The screening experiments (previous section) represent only the first step with regard to process optimization. The most important factors, which were identified in the screening

are subsequently subjected to a more detailed analysis in order to find the optimal conditions for this subset of factors. For this application, response surface methodology (RSM) is predominant (Montgomery, 2009).

RSM describes the response (for example yields or product concentrations) as a function of the analyzed factors, enabling a visualization of the response in the experimental design space. The differences to the previously discussed screening designs are the reduced number of factors and the model complexity (Equation 7). RSM models typically include interaction and second order (quadratic) terms (Equation 8).

$$y = b_0 + \sum b_i x_i + \sum b_{i \cdot i} x_i^2 + \sum\sum b_{i \cdot j} x_i x_j \qquad \textit{(Equation 8)}$$

with	x_i, x_j	*input variables*	-
	b_0, b_i, $b_{i \cdot i}$, $b_{i \cdot j}$	*zero, first, second order and interaction coefficients*	-

Standard experimental designs for RSM are symmetrical and based on full factorial designs (Figure 3.17, A). Central composite designs are generally the method of choice. They contain a two-level full factorial design (the cube), a star design and a centre point. Thus, $N = 2^k + 2k + 1$ experiments are needed for k factors. While the points of the full factorial design describe a cube at factor levels of −1 and 1, the points of the star have a different distance ($-\alpha$ / $+\alpha$) to the centre point, which is situated at zero. Several different designs with varying α exist. The circumscribed design ($\alpha > 1$) is most common. This design type is illustrated for three factors in Figure 3.17. The respective design matrix for this example is depicted in the appendix (Table 9.4).

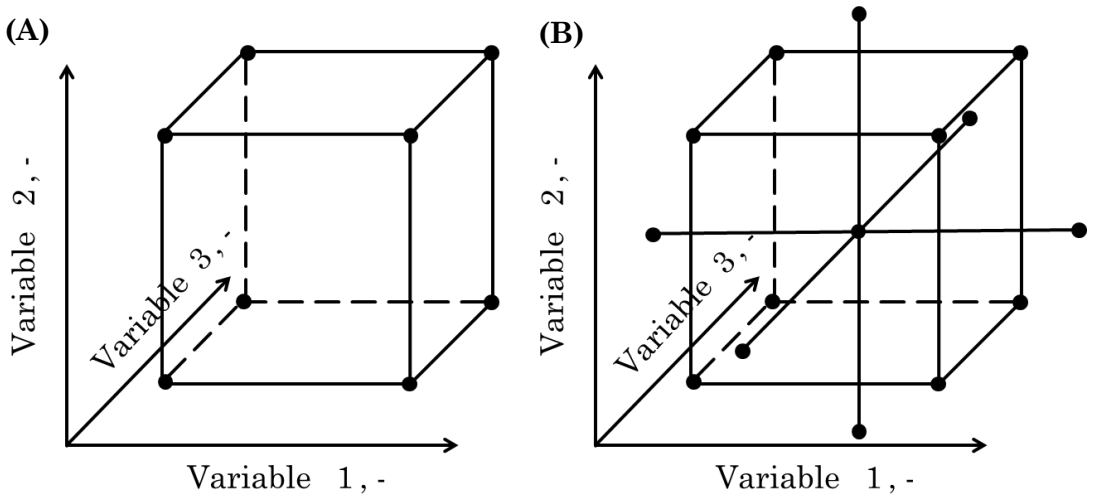


Figure 3.17: (A) Two-level full factorial design. (B) Circumscribed central composite design for response surface methodology with three factors. (●) planned experiments.

Other design types include inscribed ($\alpha < 1$) and faced ($\alpha = 1$) star points. While circumscribed RSMs offer a good accuracy over the entire design space, inscribed designs are better over the central subset. Faced designs are in overall good, but the quadratic coefficients are poorly estimated. Values for α are dependent on the number of analyzed factors (k): $|\alpha| = (2^k)^{0.25}$. Other symmetrical designs include Box-Behnken and Doehlert (uniform shell), which are a less popular choices (Dejaegher and Heyden, 2011; Montgomery, 2009). Asymmetrical designs for specific problems can be generated by the D-optimal method, analogue to the D-optimal screening design. Naturally, a more complex model with interaction and quadratic terms is used for optimization purposes (Equation 8). Therefore, far more experiments are required.

Concluding remarks

Statistical design of experiments (DOE) is based on simplified process models, in which a variable of interest (response) is described by a function of factors. Generally, the aim is to optimize this response by varying the factors in a defined set of experiments. Due to the drastically increasing complexity for problems with many factors, a two-step procedure is typically used: In a first set of experiments, the statistical DOE is confined to linear effects. The aim is to identify the most important factors with the least possible experimental effort. In a second set of experiments, this subset of factors is then optimized. In this step, the process model incorporates both interaction and quadratic effects.

3.3.2 Stochastic optimization strategies for experimental design

Stochastic optimization and global search algorithms are standard methods in informatics, engineering and related sciences (Bianchi *et al.*, 2008). Various heuristic optimization strategies like ant colony optimization, evolutionary algorithms, or particle swarm optimization are applied routinely, especially in multi-objective optimization or for problems with complex search spaces. All these approaches are heuristic, as they try to examine the search space in an "intelligent way": They attempt to find optimal solutions with minimal effort. Marked similarities of all approaches are the stochastic aspect of the optimization (there is no guarantee of reaching the global optimum) and the efficiency of the optimization process compared to classical approaches (Coello Coello, 2006). In this chapter, the experimental application of these algorithms as stochastic

DOE strategies is in the foreground, the focus lies on multi-objective genetic algorithms (GAs).

Principles of genetic algorithms

One subtype of heuristic global search algorithms are GAs, which are inspired by evolutionary principles. GAs are considered robust and powerful search and optimization methods especially for large complex search spaces and multiple objectives (Back *et al.*, 1997a). In principle, GAs simulate the process of natural evolution starting with a set of randomly generated candidate solutions, which iteratively evolve to better solutions during the optimization. Typically, the following nomenclature is used for GAs:

- A candidate solution is termed individual.
- The set of individuals is called population.
- One iteration of the algorithm is called a generation (GEN).

The basic structure of a GA is depicted in Table 3.5. In short, the GA maintains a set of feasible solutions, which change iteratively in each generation. After a number of GENs the GA converges and possibly, but not necessarily finds the global optimum. In order to work correctly, a balance is necessary between selection and evolutionary pressure on the one hand and maintenance of variance on the other hand: Individuals with low scores of the objective functions are removed from the population, while high scoring individuals are retained and hence reproduce. The aim is to narrow the search space to particularly promising areas and to increase the average quality within the population. However, the variance of the population has to be retained at the same time in order to avoid local optima and successfully identify the global optimum. This is achieved through mutation and recombination.

Table 3.5: General structure of a genetic algorithm (GA) (Back *et al.*, 1997a).

Pseudo code	Comment
1. t := 0	Starting point
2. initialize P(t)	Generate (random) first population (P).
3. evaluate P(t)	Evaluate objective function values of all individuals.
4. while not terminate do	Iterate the following steps until a termination criteria is achieved.
5. t := t + 1	Next iteration step, increase iteration numerator.
6. select P(t) from P(t − 1)	Select subset from the previous population.
7. vary P(t)	Apply mating, recombination and mutation operators to generate a new population.
8. evaluate P(t)	Evaluate objective function values of all individuals.
9. end	Terminate if termination criteria is true.

Search, decision and objective space

GAs typically operate in three different spaces (Figure 3.18):

- The decision space (X) constitutes the real-world problem, in this thesis, the refolding buffer conditions with the various experimental parameters. In analogy to evolution, it is also called phenotype space.
- The search space (I) is an encoded representation (often with reduced order) of the decision space in which recombination and mutation takes place. The representation of an individual in the search space is called chromosome. The entire space is also referred to as genotype space.
- The objective space (Y) maps the individuals according to the objective functions and is decisive for fitness assignment and selection, which will be discussed later.

During optimization, high-quality individuals are selected on basis of the objective function (f). For stochastic DOE, the experimental evaluation itself serves as the objective function. Hence, the optimization is not based on a simplified model (compare statistical DOE, section 3.3.1). Fitness assignment and the following selection are

entirely based on the obtained experimental data (y), representing the objective space (Y).

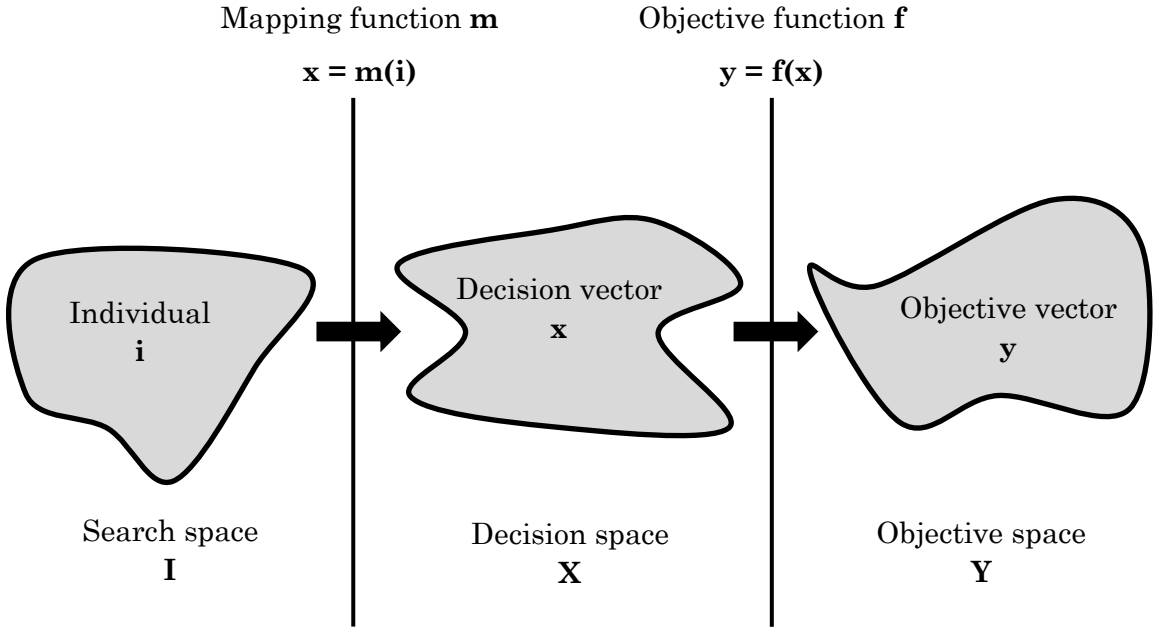


Figure 3.18: Standard spaces used for GAs. The search space (I) with the individual solutions (i), the decision space (X) with the decision vectors (x) and the objective space (Y) with the objective vectors (y). Grey shading illustrates the different structure and dimensional properties of the spaces. Mapping (m) and objective (f) functions connect the spaces (Zitzler, 1999).

As mentioned before, the GA is not working in the decision space (X) itself. Instead, the experimental problem is encoded, typically in form of a bit string (Back *et al.*, 1997a). All operations of the GA are applied to this encoded version of the problem. A decoder or mapping function (m) is necessary to map the search space on to the decision space. Practically, the decoder translates the individual (i), a bit string, into a decision vector (x). This vector describes one experiment, for example in this thesis, one refolding condition (one unique combination of pH and refolding additives).

The encoding of the problem has a strong impact on the results and a well-suited genetic representation (chromosome) is essential for good performance (Back *et al.*, 1997a). A binary representation is often the method of choice, as it is (for most problems) ideal in view of the schema theory (Rudolph, 1994; Schmitt, 2001). The schema theory analyzes the behavior of the chromosome during recombination. It observes how the chromosome changes: which subsets (schema) are retained and what is altered. Evolution is largely attributed to the augmentation and recombination of these schemes, which are also

referred to as building blocks. Mutation happens on the lowest tier (one bit), but the overall optimization is based on the larger subsets (building blocks), that are retained and rearranged during the optimization. This represents a form of dimensional reduction, as the real variables are mapped into a virtual space with reduced order. Hence, GAs are typically considered to be a good choice for complex, multidimensional problem spaces (Back *et al.*, 1997a; Weuster-Botz, 2000).

Although, the binary representation is widespread and often considered standard, various other options exist. These are important if the problem at hand is not well-suited for a binary representation (encoding function to complex) or a problem-related, more natural representation is preferred. In these cases, other evolutionary algorithms (evolutionary programming and evolution strategies) offer suitable alternatives to GAs (Back *et al.*, 1997a; Van Veldhuizen and Lamont, 2000). Another concept are hybrid evolutionary algorithms, which combine the efficiency of a heuristic method with a classic search algorithm for a finer resolution of the optimal region (Grosan and Abraham, 2007).

Selection for multi-objective optimization: the pareto principle

Next to the genetic representation of the problem, several other factors drastically influence the optimization success. Most important are the selected objective functions together with the fitness assignment and the selection procedure. In this thesis a multi-objective GA was used, the *strength pareto evolutionary algorithm* (SPEA 2), which is able to optimize several variables in parallel.

Multi-objective optimization is characterized by a subset of optimal solutions, as it is not possible to select the best candidate if more than one objective is considered. One of the most popular concepts to compare these optimal solutions is the pareto principle, illustrated in Figure 3.19 (Zitzler, 1999).

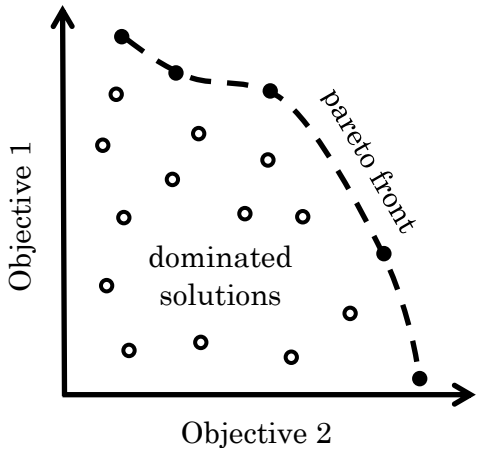


Figure 3.19: Schematic example of pareto optimality with two objective functions. (○) dominated solutions, (●) non-dominated solutions, (– –) pareto front.

According to the pareto principle, a solution (objective vector, y_1) dominates another solution (objective vector, y_2) in the objective space (Y) if no component of y_1 is smaller than the corresponding component of y_2 and at least one component is evaluated better. For two objective functions (compare Figure 3.19), at least one value of either objective 1 or objective 2 has to be higher and the other one has to be at least equal. The sum of all non-dominated solutions in the objective space (Y) is called pareto front.

The same principle can also be applied on the decision space X, but the differences between the two spaces have to be considered. Non-dominated solution vectors (x) may be mapped to different objective vectors (y). Therefore, there may be several non-dominated objective vectors. The set of optimal solutions in the decision space (X) is termed pareto set. A globally optimal solution (the global pareto set) is the non-dominated set in the whole search space (X). The aim of the optimization is to identify this set of optimal solutions. However, due to the heuristic nature, it is not guaranteed that the GA correctly identifies the global pareto set. A local pareto set is defined as a set of solutions (x), for which no objective vector (y) in the neighborhood dominates any member of the set (Zitzler, 1999).

Although the selection process for multi-objective algorithms is typically based on the above-discussed pareto dominance, which is used to assign a fitness value for each individual, the practical implementation is usually modified. One important addition is clustering, which is used to reduce the amount of non-dominated solutions and is applied after evaluation dominance and fitness assignment. A reduction is necessary as too many individuals could reduce the selection pressure and slow down the optimization process (Covas *et al.*, 1999). Specific methods and implementation are dependent on the

algorithm of choice. For details regarding the specific algorithm (SPEA 2) the reader is referred to the original literature (Zitzler *et al.*, 2002).

Recombination

After the selection of the best (highest fitness values) individuals, a new set of candidate solutions is generated by recombination. This procedure is typically divided in three parts. In the first step a mating pool is generated, based on the selected individuals of the current GEN and optionally an external archive with good solutions from previous GENs. Afterwards, it is necessary to determine which individuals recombine with each other. A popular method is binary tournament (Zitzler, 1999). Subsequently, operators like crossing-over are applied, which recombine the individuals in analogy to basic genetic principles. In order to maintain schemes (building blocks) in the optimization, single point crossing-overs are typical (Weuster-Botz, 2000). Finally, the variance of the above-generated candidate solution is increased by mutation. For a representation as a binary string, mutation usually affects each bit individually, this means that each bit has a certain probability to be flipped. In practice, methods and implementation are both dependent on the algorithm and the encoding of the problem (Zitzler, 1999). In addition, other functions may be implemented. A common example is the verification of the new candidate solutions to ensure that only novel solutions are evaluated and no experiment is repeated.

Experimental applications: number of experiments and error

Most GAs focus on pure *in silico* problems or problems in which a simulation is carried out to evaluate the objective function. Thus, the objective function(s) is computationally evaluated and real experiments are limited to a validation of the optimal conditions at the end. In these cases, experimental effort is a matter of computational time and largely neglectable. Therefore, large population sizes and many iterations are the norm. In contrast, following criteria have to be considered for experimental stochastic optimizations. First, the experimental effort (number of experiments) should be minimized, as experiments are the major cost factor. Additionally, a relative high experimental error of up 20 % standard deviation is often observed. Finally, complex problems with a many variables and possible interactions occur on a regular basis (Weuster-Botz, 2000).

Concluding remarks

Genetic algorithms (GAs) are population-based heuristic search methods, that use evolutionary principles to efficiently examine the search space in an intelligent way. GAs are considered robust and powerful, especially for large complex search spaces and multiple objectives (Back *et al.*, 1997a). Typical applications involve *in silico* problems, in which computational time is the only limiting factor. GAs can also be used as a stochastic DOE. In this case, the experimental evaluation itself constitutes the objective function. The distinguishing feature in comparison to the statistical DOE strategies (3.3.1), is the model independence. There is no underlying simplified process model and no unimodality assumed.

3.4 Black-box models for data analysis

Models are typically classified according to the amount of available *a priori* information on the analyzed system. In the best case, the system is well understood and the knowledge about the functional relations between variables can be used to generate a mechanistic model. Thus, the model is only used to estimate unknown parameters. However, many problems offer only limited information about the functional relations. Hence, both function and parameters have to be estimated. These models are typically called black-box models. Two standard approaches will be discussed in this chapter: artificial neural networks (ANN), a biologically-inspired method that mimics neural processing and bagged decision trees (BDT), an example for ensemble models.

3.4.1 Artificial neural networks (ANNs)

Artificial neural networks (ANNs) are applied in virtually every scientific discipline and are widely used in industry as well. Traditionally, ANNs focus on three areas: pattern recognition, data clustering and function fitting (Meireles *et al.*, 2003).

Artificial neurons – the processing units of ANNs

ANNs, like genetic algorithms, belong to the biologically-inspired computing methods. ANNs are based on neurons as the individual processing units of the network (Figure 3.20). Neurons are structured exactly like the biological prototype and characterized by a series of weighted (w) inputs (x), a transfer function (f) and one output (y). Incoming signals are processed by calculating the weighted sum of all inputs (I). This can be done straightforward (Equation 9) or additional weights (bias) can be integrated in this step.

$$I = \sum x_i w_i \qquad \textit{(Equation 9)}$$

with	I	*weighted sum of inputs*	-
	x_i	*inputs*	-
	w_i	*weights*	-

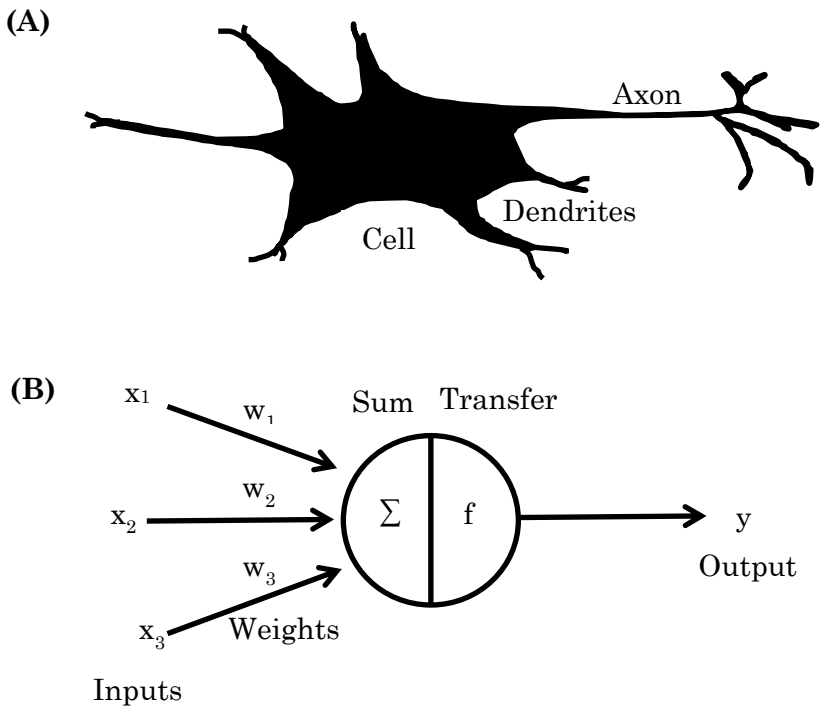


Figure 3.20: Schematic neuron. Comparison between the biological prototype (A) and the artificial model (B) with inputs (dendrites, x_i and w_i), processing (cell, sum and transfer functions) and output (axon, y) (Agatonovic-Kustrin and Beresford, 2000).

After calculating the net signal (weighted sum of inputs, I), the signal is transformed and an output signal is generated. In the biological system, the neuron only responds if a certain threshold value is exceeded. In the model, a transfer function (f) transforms the net signal into an output (y) with an output value ranging from −1 to 1. A variety of transfer functions are illustrated in Figure 3.21. Most common for multilayer networks is the log-sigmoid transfer function.

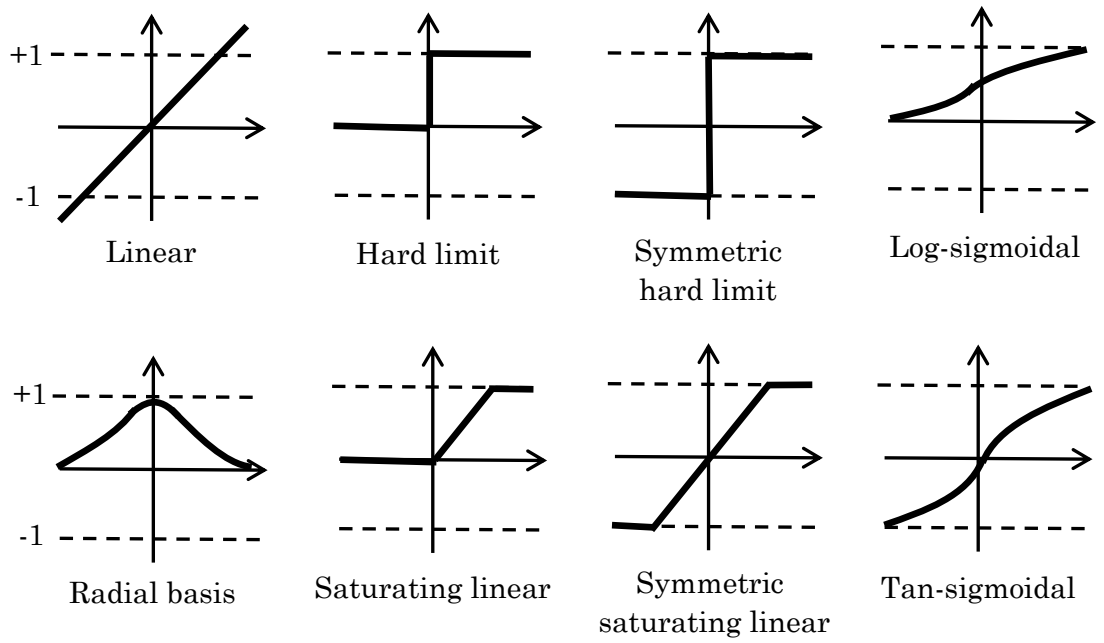


Figure 3.21: Standard transfer functions for neurons (Patnaik, 1998)

Network structure

Neurons are only the individual processing units of the network. An ANN is structured into several layers of neurons interconnected by outputs from neurons of the previous layer and inputs to the next layer, each with their respective weights. Various ANN structures are used which differ in the number of neurons, the connection formula and the training procedure. The general structure of an ANN is the following:

- Inputs, x_i.
- Input layer, with as much neurons as input variables.
- Hidden layer(s), one or more layers with varying number of neurons. Architecture, size and connectivity is strongly dependent on the problem, the type of network used and often subject to iterative changes to optimize the performance. The hidden layers represent the processing part of the network. Because of the complexity, it is usually regarded as a black-box system.
- Output layer, with as much neurons as output variables.
- Outputs, y_i.

A standard ANN network is the feedforward network, also called backpropagation network. Its architecture is depicted in Figure 3.22.

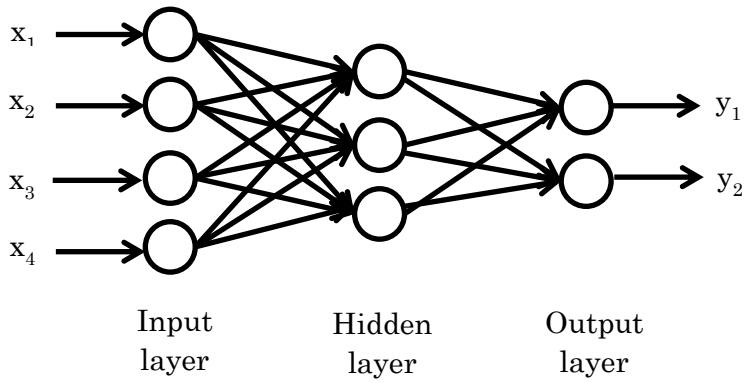


Figure 3.22: Feedforward ANN with three layers and four input (x_i) and two output variables (y_i) (Patnaik, 1998).

Training and validation

For the application, an ANN with a specified structure is first generated and weights (w_i) are initialized. Afterwards, the network is trained using part of the experimental data. In the training process the network weights are adjusted to optimize performance, that is the correct prediction of the functional relationship between inputs (x) and outputs (y). The usual measure of performance is the mean square error (Equation 10) between the network output and the known real output, commonly called target output.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(T_i - A_i) \qquad \textit{(Equation 10)}$$

with	MSE	*mean square error*	-
	T_i	*target outputs*	-
	A_i	*network outputs*	-

A variety of standard numerical algorithms can be used to optimize the network performance. Common choices are: Levenberg-Marquardt, gradient descent, gradient descent with momentum or scaled conjugate gradient. These optimization methods use the gradient of the network performance with respect to the network weights. The gradient is calculated using a technique called backpropagation, which involves performing computations backward through the network according to Rumelhart *et al.* (1986).

After training, the network is validated on the part of the dataset, that was not used for training purposes. Ratios of 75 % (training) and 25 % (validation) are typical if no internal test or cross validation is used. ANN performance in the validation is strongly

dependent on the amount of available training data and the division of training and validation datasets (Meireles *et al.*, 2003; Patnaik, 1998). Due to their complexity ANNs are able to approximate any reasonable function. However, the application of the network on new data (generalization) can pose a serious problem. (Meireles *et al.*, 2003; Razi and Athappilly, 2005).

Concluding remarks

ANNs are biologically-inspired models with wide-spread use in science and industry. ANNs mimic neural processing both with regards to the processing unit (neuron) and the connectivity. Modeling is based on adjusting the network weights in the training and then using the trained network to predict the output for the rest of data (validation). While ANNs are considered powerful tools for data mining and modeling, the generalization error is often problematic (Razi and Athappilly, 2005).

3.4.2 Bagged decision trees (BDT) – random forest

Decision Trees

Decision trees are a common method in data mining which can be used both for classification or regression. In such a tree structure, leaves represent class labels or real numbers and branches are logical conjunctions (variable thresholds). The general structure of a regression tree is illustrated in Figure 3.23.

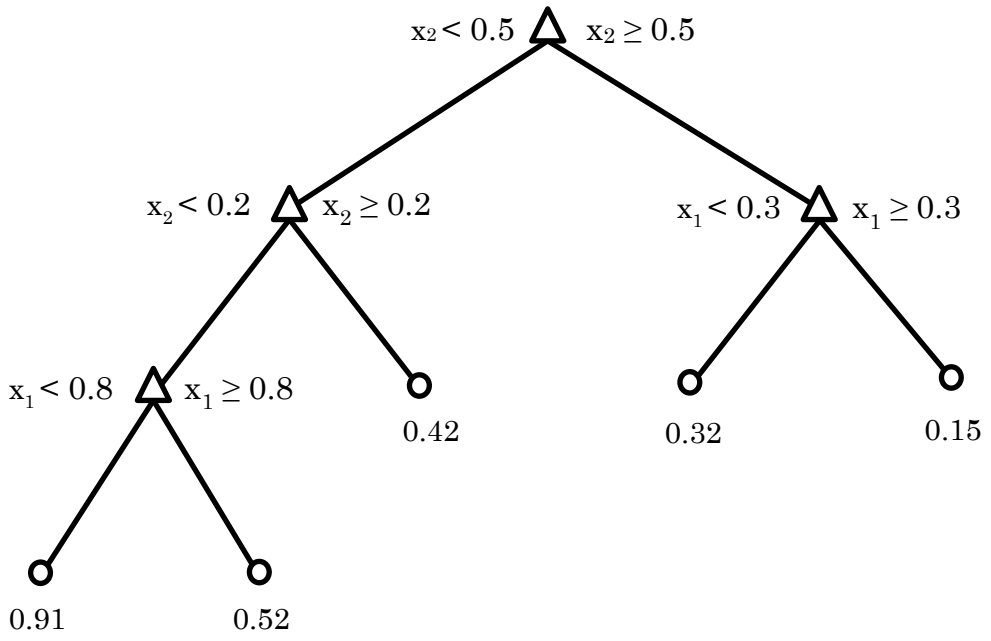


Figure 3.23: Exemplary decision tree for a regression problem with two input variables (x_i), four normal nodes (Δ) and six leaves (○).

The generation of the tree model is generally performed top-down by splitting, that is introducing branch points (recursive partioning). At each branch point, one of the input variables (x_i) is selected and an attribute test (below or under threshold) is performed. If the test is positive, the tree is split and two subsets are generated. This process is repeated until splitting no longer increases the prediction performance (Breiman *et al.*, 1993).

Although decision trees have a variety of advantages (simple to understand, no black-box-model, works with numerical and categorical data) individual trees have serious drawbacks in comparison with other modeling approaches. For a comparison to ANNs the reader is referred to Razi and Athappilly (2005). However, ensemble systems, which are based on many individual trees are far more effective (Breiman, 2001). The concept of ensemble models will be discussed in the following.

Ensemble based systems – bootstrap aggregation

A recent development in modeling is the concept of combining many individual models and using the entire ensemble for prediction. These methods are based on resampling techniques like bootstrap aggregating (bagging). Advantages of ensemble systems include above all a good generalization performance. The model prediction of new data is not as problematic compared to other approaches like ANNs. Individual models show different generalization performance. Thus, averaging over all models reduces the risk of making a poor choice and the overall generalization errors are typically smaller. In addition, ensembles perform better in the absence of adequate training data, that is insufficient experimental data or an unsuitable distribution. Resampling techniques can be used to obtain overlapping subsets of the available dataset. Afterwards each subset is used to train a different individual model. Furthermore, ensemble models are able to approximate complex problems with non-linear interactions. A classification problem in which a complex decision boundary between class 1 (●) and 2 (○) is approximated with an ensemble is exemplified in Figure 3.24. Finally, ensemble systems also perform better on too much data or a fusion between different datasets (Polikar, 2006).

Different methods exist for creating the ensemble. Next to boosting (Freund and Schapire, 1997) bagging is the most popular choice. Bagging is generally considered superior for datasets with high errors (Breiman, 2001; Polikar, 2006).

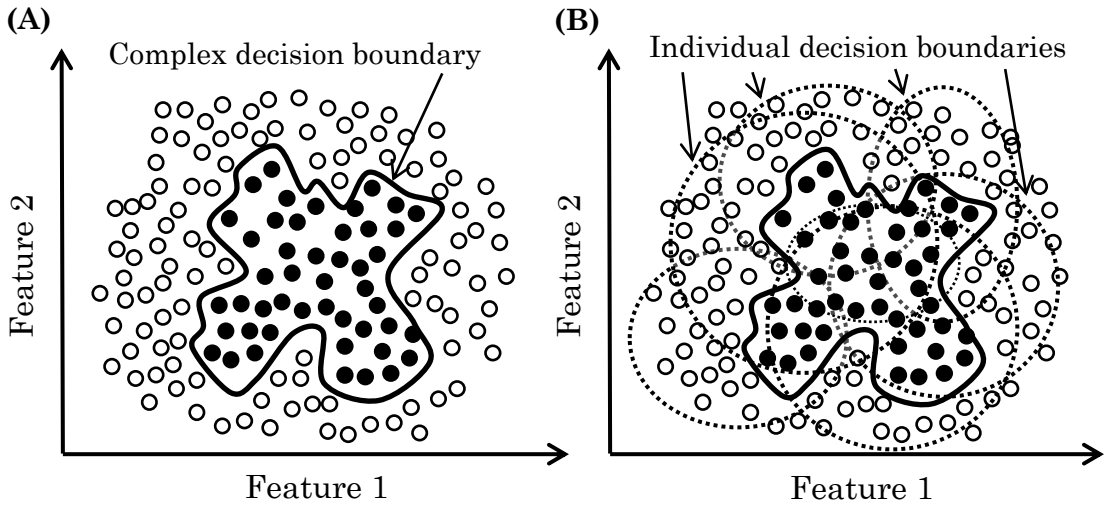


Figure 3.24: Ensemble classification. (A) Classification problem with two classes (● / ○) and a complex decision boundary (———) that cannot be learned by linear or circular classifiers. (B) Ensemble of circular classifiers (·····), which span the decision space and allow an approximation of the decision boundary (Polikar, 2006).

Bagging is based on resampling the dataset by selecting data with replacement. The general structure of the algorithm is as follows (Polikar, 2006):

- Training dataset of size m: (x_1, y_1), ..., (x_m, y_m).
- For i = 1, ..., k
    - Form a bootstrap replicate dataset S_i by selecting m random examples from the training set with replacement. Hence, the same example may appear multiple times in the bootstrap replicate, or not at all.
    - Train one model on S_i and obtain the result h_i.
- Combine the models for prediction and obtain the ensemble model (H). For this, a variety of averaging function (f) can be applied: $H(x) = f(h_1(x), ..., h_k(x))$.

In principle, resampling techniques like bagging can be used with all modeling approaches, including ANNs. However, bagging is especially successful for decision trees. In this case the ensemble of bagged decision trees is also called random forests (Breiman, 2001).

Concluding remarks

Decision trees are an intuitive modeling approach in which the dataset is recursively partitioned in order to obtain a tree like structure for classification or prediction. They are especially powerful in combination with resampling techniques, like bagging, which are used to create an ensemble of individual models (random forest). This ensemble model is then used for prediction by averaging the output of the individual models to generate an overall prediction. In comparison to ANNs, especially the good performance on inadequate data (not enough data or bad distribution) and the good generalization performance (prediction of new data) distinguishes this modeling approach.

4 Material and Methods

This chapter describes the experimental and computational methods used during this project. Abbreviations and material are listed in the appendix. First, the experimental methods are depicted: Section 4.1 describes the protein refolding experiments that were integral for this project. Subsequently, methods for protein analytics and molecular biology are explained in 4.2 and 4.3, respectively. Section 4.4 sums up basic calculations used in this thesis. Finally, the computational methods used for experimental design and modeling are detailed in section 4.5 and 4.6.

4.1 Protein refolding

Protein refolding was optimized in 96-well plate format. After denaturation and refolding, the reaction yield was determined using a functional assay. Proteins were either purchased in purified form, provided by the cooperation partner (Department Chemie, Center for Integrated Protein Science, Technische Universität München) or expressed in *Escherichia coli (E. coli)* (Table 4.1).

Table 4.1: Overview of analyzed proteins (cooperation partner*: Department Chemie, Center for Integrated Protein Science, Technische Universität München).

Abbr.	Protein	Source
GFP	Green fluorescent protein from *Aequorea victoria*	Cooperation partner*
GLK	Glucokinase from *Escherichia coli*	Cooperation partner*
GLR	Glutathione reductase from *Saccharomyces cerevisiae*	Sigma-Aldrich (G3664)
LYZ	Lysozyme from *Gallus gallus*	Sigma-Aldrich (L7651)
LDH	Lactate dehydrogenase from *Oryctolagus cuniculus*	Sigma-Aldrich (61309)
LIP	Lipase from *Thermomyces lanuginosus*	Sigma-Aldrich (L0777) or expressed in *E. coli*

4.1.1 Denaturation

Proteins were denatured in the presence of 6 M guanidine hydrochloride (Gdn·HCl) at room temperature (RT). Specific denaturation conditions were taken from the literature and are listed in Table 4.2. Denaturation was verified via the respective functional assay and circular dichroism spectroscopy (CD) for the lipase from *Thermomyces lanuginosus* (LIP).

Table 4.2: Protein denaturation conditions (**DTT**, dithiothreitol; **EDTA**, ethylenediamine-tetraacetic acid; **PB**, sodium phosphate buffer; **TRIS**, tris-hydroxymethyl-aminomethane).

Protein	Denaturation buffer	Time	c, g L^{-1}	Reference
GFP	50 mM TRIS·HCl, pH 7.5, 6 M Gdn·HCl	Overnight	1.0	(Dashivets *et al.*, 2009)
GLK	50 mM TRIS·HCl, pH 8.0, 6 M Gdn·HCl, 5 mM DTT	Overnight	0.4	(Dashivets *et al.*, 2009)
GLR	100 mM PB, pH 6.9, 6 M Gdn·HCl, 5 mM DTT	3.0 h	0.2	(Hevehan and De Bernardez Clark, 1997)
LYZ	100 mM PB, pH 6.9, 6 M Gdn·HCl, 5 mM DTT	3.0 h	1.0	(Hevehan and De Bernardez Clark, 1997)
LDH	200 mM PB, 1 mM EDTA, 6 M Gdn·HCl, 0.1 mM DTT	0.5 h	0.1	(Rudolph *et al.*, 1977)
LIP	100 mM TRIS·HCl, pH 7.5, 6 M Gdn·HCl (or 10 M urea), 5 mM DTT	2.5 h	0.5	(Rudolph and Lilie, 1996)

4.1.2 Refolding

Protein refolding was optimized regarding the refolding buffer composition. In the conducted experiments protein concentration, temperature, time and stirring were standardized and kept constant. For refolding, the denatured protein was rapidly diluted 15-fold to 200-fold to a final concentration of 1 mg L^{-1} to 33 mg L^{-1} in the respective refolding buffer. Very low protein concentrations between 1 mg L^{-1} and 5 mg L^{-1} were desired to reduce aggregation. However, LIP and GLK required higher concentrations, as the functional assays were not as sensitive. Specific refolding conditions are listed in Table 4.3.

Table 4.3: Protein refolding conditions I – Standardized variables.

Protein	Dilution step	c, mg L^{-1}	Temp, °C	Time	Reference
GFP	100	4.0	10	Overnight	(Dashivets *et al.*, 2009)
GLK	50	20.0	10	Overnight	(Dashivets *et al.*, 2009)
GLR	200	1.0	20	Overnight	(Hevehan and De Bernardez Clark, 1997)
LYZ	200	5.0	20	Overnight	(Hevehan and De Bernardez Clark, 1997)
LDH	68	1.4	20	2.0 h	(Rudolph *et al.*, 1977)
LIP	15	33.0	4	Overnight	(Rudolph and Lilie, 1996)

In contrast to the standardized variables, the refolding buffer composition was optimized with a genetic algorithm (GA). For this purpose experimental parameters were extracted from the refolding literature and combined with the information on approximately 1100 refolding experiments from the REFOLD database (Amin *et al.*, 2006; Buckle *et al.*, 2005) to establish a comprehensive experimental design (summarized in Table 4.4). Functionally related substances and conditions were subgrouped in six different classes.

- Buffer: The first class referred to the pH and the buffering agent. TRIS·HCl and PBs were most prominent in REFOLD (Amin *et al.*, 2006; Buckle *et al.*, 2005). Nevertheless, hydroxylethyl-piperazine-ethanesulfonic acid (HEPES) and morpholino-propanesulfonic acid (MOPS), two other common organic buffers, were also included. Concentrations were varied between 20 mM to 100 mM for phosphate, HEPES and MOPS. For TRIS·HCl up to 1.25 M were examined, because it was previously employed as a refolding additive (Rudolph and Lilie, 1996). With regard to the pH, a range between pH 6.0 and pH 9.5 covered most published refolding experiments (Amin *et al.*, 2006; Buckle *et al.*, 2005). In addition, only conditions within the buffer range were permitted for PB (pH 6.0 to pH 7.5) and TRIS·HCl (pH 7.0 to pH 9.5).
- Salts: NaCl was used as the primary compound to vary the ionic strength of the buffer. Furthermore, the addition of small concentrations (20 mM) of KCl was analyzed.

- Additives: This class was composed of refolding additives, including glycerol and polyethylene glycol (PEG 4000) as well as three commonly used amino acids (arginine, glutamine and glycin). A later version of the stochastic optimization also incorporated glutamate.
- Mineral ions: Divalent metal cations (Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2+} sulfates) utilized in past refolding experiments (Armstrong *et al.*, 1999) and alternatively, EDTA formed the fourth class.
- Detergents: Eight detergents, including different detergent families (zwitterionic, ionic and nonionic) in concentrations between 0 and 4/3 of the critical micellar concentration (CMC) were incorporated in the optimization. CMCs are detailed in the appendix (Table 9.17). In addition, a non-detergent sulfo-betaine that was previously utilized in refolding screens (Qoronfleh *et al.*, 2007; Willis *et al.*, 2005) was included.
- Redox agents: As disulfide bonds play a central role in protein structure, common reducing and oxidizing agents like DTT, tris-carboxyethyl-phosphine (TCEP), reduced L-glutathione (GSH) and oxidized L-glutathione (GSSG) formed the last class.

A refolding condition consisted of at least one pH and one buffer substance, for example TRIS·HCl, pH 8.0. Additionally, substances from other classes could be included, for example TRIS·HCl, pH 8.0 with 100 mM NaCl, 100 mM arginine, 5 mM DTT. Furthermore, combinations within several classes were possible. These were annotated with “and”, not possible combinations with “or”. The specific encoding of the refolding conditions is detailed in 4.4.3.

Table 4.4: Protein refolding conditions II – Parameters of the stochastic optimization. The general setup was modified in the course of this project, changes are highlighted with new values in brackets (*).

Parameter / substance	Min	Max	Unit
pH	6.0	9.5	-
Buffer substances: no combination			
PB	20	100	mM
HEPES	20	100	mM
MOPS	50	100	mM
TRIS·HCl*	20	1250 (1000)	mM
Salts: combination of NaCl and KCl			
NaCl	0	350	mM
KCl*	0	20 (80)	mM
Additives: combination (glycerol or PEG) and (arginine and glutamine and glycine)			
Glycerol	0	15.0	% v/v
PEG 4000	0	0.2	% w/v
Arginine	0	750	mM
Glycine	0	150	mM
Glutamine	0	100	mM
(Glutamate)*	0	- (200)	mM
Cofactors: no combination			
(Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2+})*	0	5 (100)	mM (µM)
EDTA*	0	2 (10)	mM
Detergents: no combination			
zwitterionic			
CHAPS	0	11	mM
ZWITTERGENT 3-12	0	4	mM
NDSB 201	0	1500	mM
nonionic			
TWEEN 20	0	80	µM
TRITON-X 100	0	800	µM

Table 4.4 (continued):

Parameter / substance	Min	Max	Unit
BRIJ 35	0	120	µM
ionic			
SDS	0	12	mM
SDC*	0	8 (-)	mM
*Redox agents: combination DTT or TCEP or (GSH and GSSG)**			
DTT	0	10	mM
TCEP	0	10	mM
GSH	0	5	mM
GSSG	0	5	mM

PB, sodium phosphate buffer; **HEPES**, hydroxylethyl-piperazine-ethanesulfonic acid; **MOPS**, morpholino-propanesulfonic acid; **TRIS**, tris(hydroxymethyl)aminomethane; **PEG**, poly-ethylene glycol**; (Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2+})** sulfates; **EDTA**, ethylenediaminetetraacetic acid; **CHAPS,** cholamidopropyl-dimethylammonium-propanesulfonate; **ZWITTERGENT 3-12,** dodecyl-dimethyl-ammonio-propanesulfonate; **NDSB 201,** non-detergent sulfobetaine 201; **TWEEN 20**, polyethylene glycol sorbitan-monolaurate; **TRITON-X 100,** polyethylene glycol tert-octylphenyl ether; **BRIJ 35**, polyethylene glycol dodecyl ether; **SDS,** sodium dodecyl sulfate; **SDC**, deoxycholic acid sodium salt; **DTT**, dithiothreitol; **TCEP**, tris-carboxyethyl-phosphine; **GSH**, reduced glutathione; **GSSG**, oxidized glutathione.

For refolding, denatured proteins were rapidly diluted using the respective refolding buffer. In addition, control reactions with native, non-denatured protein were carried out. Each generation of the stochastic optimization contained 22 unique, previously untested refolding conditions. Next to these 22 conditions, two controls were measured in all experiments: The buffer of the functional assay and a refolding condition from the literature. The refolding reaction was carried out in 2.2 mL 96-well plates (Sarstedt). Figure 4.1 displays the “one-plate” layout for LDH and LIP. Other proteins were measured on two plates with multiple evaluations of the native activity and an additional control without any protein.

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
	native			refolded 1			refolded 2			refolded 3		
A	1	9	17	1	9	17	1	9	17	1	9	17
B	2	10	18	2	10	18	2	10	18	2	10	18
C	3	11	19	3	11	19	3	11	19	3	11	19
D	4	12	20	4	12	20	4	12	20	4	12	20
E	5	13	21	5	13	21	5	13	21	5	13	21
F	6	14	22	6	14	22	6	14	22	6	14	22
G	7	15	c	7	15	c	7	15	c	7	15	c
H	8	16	c*	8	16	c*	8	16	c*	8	16	c*

Figure 4.1: Experimental setup of the 96-well plate refolding experiments. (1 to 22) refolding conditions of the current generation, (c) native control, the buffer of the functional assay, (c*) refolding control from the literature.

4.1.3 Functional assays

In order to quantify the refolding exactly, a protein-specific functional assay was applied after the refolding screen. All measurements were carried out in 96-well plate scale in a Genios™ or Infinite® M200 plate reader (Tecan), unless otherwise specified. Samples were taken directly from the 2.2 mL 96-deepwell plates with a multichannel pipette and transferred to a 300 µL 96-well plate for the protein-specific assay. GFP, GLK, GLR and most of the LYZ measurements were carried out by the cooperation partner.

GFP: The activity, that is the structural integrity of GFP variant F64L and S65T (Topell *et al.*, 1999) was determined by fluorescence emission of the folded and oxidized protein (Dashivets *et al.*, 2009). Measurements were carried out in a SPEX II fluorescence spectrometer (Jobin Yvon) with a fixed excitation wavelength of 395 nm and an emission scan between 430 nm and 550 nm. The signal intensity of the emission peak (508 nm) was used to calculate refolding yield.

GLK: ATPase activity was measured in an ATP (adenosine-triphosphate) regenerating system (Figure 4.2) coupled to NADH consumption, which was monitored at 340 nm (Nørby, 1988). The measurement was carried out in the presence of 2.5 mM ATP and 800 µM D-glucose. In addition, the assay contained 2 mM phosphoenolpyruvate (PEP), 0.2 mM NADH, 2 U mL^{-1} pyruvate kinase (PK), 10 U mL^{-1} LDH and 15 mM ammonium sulfate.

$$1.\ \mathrm{ATP} + \mathrm{H_2O} \xrightarrow{GLK} \mathrm{ADP} + \mathrm{P_i}$$

$$2.\ \mathrm{Phosphoenolpyruvate} + \mathrm{ADP} \xrightarrow{PK} \mathrm{pyruvate} + \mathrm{ATP}$$

$$3.\ \mathrm{Pyruvate} + \mathrm{NADH} \xrightarrow{LDH} \mathrm{lactate} + \mathrm{NAD^+}$$

$$\mathrm{Sum:\ PEP} + \mathrm{NADH} \rightarrow \mathrm{lactate} + \mathrm{NAD^+} + \mathrm{P_i}$$

Figure 4.2: ATP regenerating system for the glucokinase activity assay (P_i, hydrated inorganic phosphate).

GLR activity was determined by measuring the decrease of the cosubstrate NADPH at 340 nm (Mavis and Stellwagen, 1968). The assay was adjusted to 96-well plates and 250 µL total volume. 25 µL sample were added to 225 µL master mix (PB, NADPH, GSSG, EDTA). The reaction mix contained 75 mM PB, pH 7.6, 2.6 mM EDTA, 1 mM GSSG and 0.09 mM NADPH. After mixing, kinetics were measured at 340 nm at 25 °C for 15 min.

LYZ activity was analyzed using the EnzChek® lysozyme assay kit (Invitrogen) based on fluorescence labeled *Micrococcus lysodeikticus* cell walls. Measurements were carried out in black 96-well plates in a total volume of 100 µL. 50 µL sample were added to 50 µL reaction mix (100 mM PB, pH 7.5, 100 mM NaCl, 25 mg L^{-1} substrate). After mixing, the fluorescence (excitation 485 nm, emission 535 nm, gain 76, flashes 3, integration time 40 µs) was measured for 15 min at 37 °C. The increase of fluorescence was proportional to LYZ activity.

LDH activity was determined by measuring the decrease of the substrate NADH at 340 nm (adapted from Stambaugh and Post, 1966). The assay was performed in 96-well plates and 250 µL total volume. First, a master mix with 160 mM TRIS·HCl, pH 7.3, 6 mM NADH and 50 mM pyruvate was generated with the stock solutions (Table 9.22). 200 µL master mix were pipetted in each well and baseline activity was measured in an El 808 Ultra Microplate Reader (BioTek) at 340 nm for 12 min. Subsequently, 50 µL sample were added and mixed. Finally, reaction kinetics were measured at 340 nm for 12 min at RT.

LIP activity was measured with a 4-nitrophenyl palmitate based assay (Liu *et al.*, 2006) with 0.25 g L^{-1} nitrophenyl palmitate, 0.6 g L^{-1} gum arabic and 2.9 g L^{-1} Triton X-100. The assay was performed in 96-well plates and 270 µL total volume buffered with 1.45 M TRIS·HCl, pH 7.5. Immediately before measurement, solution B (157.5 µL per well) and

solution A (22.5 µL per well) were mixed in a 300 µL 96-well plate (Nunc, Thermo Scientific). Afterwards, baseline absorbance at 410 nm was measured for 5 min at 37 °C. Subsequently, 90 µL sample were added. After mixing, kinetics were measured at 410 nm at 37 °C for 15 min. Stock solutions A and B are listed in the appendix (Table 9.23).

4.1.4 Circular dichroism spectroscopy

Protein denaturation was verified with the respective functional assay (see section 4.1.3). Circular dichroism spectroscopy (CD) was used as an additional, structure-based method. CD measurements were carried out in a J-715 spectropolarimeter (JASCO). First, the protein was diluted into the appropriate buffer (native or denaturizing conditions) and incubated at RT for 2.5 h. Second, the sample was transferred into a 130 µL cuvette cell (106-QS with detachable window, Hellma Analytics). CD spectra between 260 nm and 190 nm were recorded using default parameter settings. For analysis the blank spectra of the buffer was subtracted from the native and denatured spectra. The molar ellipticity θ_{MRW} was calculated as follows:

$$\theta_{MRW}(\lambda) = \frac{\theta_d(\lambda) \cdot 0.1}{n \cdot c \cdot d} \qquad \textit{(Equation 11)}$$

with	θ_{MRW}	*molar ellipticity*	*deg cm² dmol⁻¹*
	MRW	*mean residue weight*	-
	θ_d	*ellipticity*	*deg*
	λ	*wavelength*	-
	n	*number of peptide bonds*	-
	c	*protein concentration*	*mol L⁻¹*
	d	*cuvette width*	*cm*

4.2 Protein analytics

Proteins were characterized regarding purity and concentration using a variety of standard analytical methods.

4.2.1 Protein concentration determination

Protein concentration was determined with three different methods. For the Bradford assay (Bradford, 1976), 30 µL sample and 1.5 mL Bradford solution (Sigma-Aldrich) were mixed and incubated at RT for 5 min. Afterwards, the extinction at 595 nm was

determined in a UV/VIS photometer (spectral photometer BioMate 3, Thermo Scientific). The BCA protein assay (Smith *et al.*, 1985) was performed using the BCA protein assay kit (Thermo Scientific) on a 225 µL-scale. For both assays a BSA standard was used to obtain concentration values. The extinction at 280 nm (Ennis and Layne, 1957) was determined in a UV/VIS photometer with a 1 mL fused glass cuvette. Protein extinction coefficients ε and molecular mass M were calculated with the online tool ProtParam (http://web.expasy.org/protparam; Feb 2012; Wilkins et al., 1999). Subsequently, protein concentrations were calculated according to the Lambert-Beer law (Equation 12).

$$E = c \cdot d \cdot \varepsilon \qquad \textit{(Equation 12)}$$

with	E	*extinction*	-
	c	*concentration*	$mol\ L^{-1}$
	ε	*extinction coefficient*	$mol\ L^{-1}\ cm^{-1}$
	d	*cuvette width*	*cm*

4.2.2 Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)

Protein purity was analyzed via SDS-PAGE (Fling and Gregerson, 1986; Laemmli, 1970) with a 3 % stacking gel and a 12.5 % separating gel. After adding 5 x Laemmli buffer the samples were incubated at 95 °C for 5 min. Subsequently, the gel was loaded with 30 µL sample. Roti®-Mark standard (14 kDa to 212 kDa, Carl Roth) was used for size determination. Electrophoresis was carried out in running buffer (0.25 M TRIS·HCl, 2 M glycine pH 8.8, 1 % SDS) at 30 mA and a maximal power of 300 W. Afterwards, the gels were stained with Coomassie (Fairbanks *et al.*, 1971) and digitized. Staining solutions and buffers are listed in the appendix (Table 9.18 and Table 9.19).

4.2.3 Protein dialysis

Purchased lipase from *Thermomyces lanuginosus* (LIP, Lipolase™, Sigma-Aldrich) was dialyzed in order to remove stabilizing agents (propylene glycol, $CaCl_2$). 10 mL protein solution were dialyzed (1 to 500 dilution) overnight using a 14 kDa molecular weight cut off in dialysis buffer (0.1 M TRIS·HCl, pH 7.5) at 4 °C.

4.3 Molecular biology

LIP was purchased from Sigma-Aldrich in soluble form. In addition, the protein was also expressed in *E. coli*. After oligomer assembly, the two constructs (with and without His-tag) were expressed using standard procedures in *E. coli* BL21 (DE3). Subsequently, both proteins were purified and evaluated in terms of refolding.

4.3.1 Design of DNA oligomers

DNA oligomers were designed using DNAWorks (http://helixweb.nih.gov/dnaworks; Feb 2012; Hoover and Lubkowski, 2002;) with standard parameter settings and the *E. coli* class II codon frequency table. The first and last oligomers included restriction sites and an extension for enzymatic digestion. In addition, the terminal codon 36 was modified manually: variant A (containing a GSG linker and C-terminal His-tag) and variant B (wild type sequence, including a stop codon). All oligomers are listed in the appendix (Table 9.9).

4.3.2 Assembly of DNA oligomers

DNA oligomers were purchased from Eurofins MWG Operon (grade: salt free, scale: 0.01 µmol.) The oligomers were dissolved in ddH_2O and diluted to a final concentration of 150 µM. Afterwards, oligomers were pooled (oligomer 1 to 35 and either oligomer 36 a or b) and the mix diluted to a final concentration of 25 µM. In a first polymerase chain reaction (PCR) using the Phusion polymerase (New England Biolabs) the oligomers were assembled (Table 9.9) Afterwards the products were amplified in a second PCR (Table 9.11). The products were subsequently purified with the GeneElute™ PCR clean-up kit (Sigma-Aldrich). Yield and length were evaluated by agarose gel electrophoresis using 1 % v/v agarose and 0.4 mg L^{-1} ethidium bromide staining. Electrophoresis was carried out in a TAE buffer system (Table 9.20) at 120 V.

4.3.3 Ligation and transformation

The PCR product and vector (pET21-a +, Novagen) were digested using 10 U restriction enzymes (NdeI and XhoI) according to manufactures instruction in 20 µL or 50 µL total volume. Subsequently, the vector was dephosphorylated in 70 µL total volume with 10 U antarctic phosphatase. All enzymes were purchased from New England Biolabs (Table 9.12).

Prior to ligation, the digested PCR product and vector were purified using the GeneElute™ PCR clean-up kit (Sigma-Aldrich). Subsequently, the ligation was carried

out with the Quick Ligation™ kit (New England Biolabs). Thereafter, *E. coli* DH5*α* (Invitrogen) was transformed using the ligated vectors: 200 µL of competent cells were thawed on ice. 10 µL ligation product were added and the cells were incubated on ice for 30 min. After a heat shock for 30 s at 42 °C 600 µL NZY medium were added. The samples were incubated for 1 h at 37 °C. Subsequently, the cells were gently pelleted (1000 g for 5 min) and spread on Luria broth (LB)-plates containing 50 mg L^{-1} ampicillin for selection. Cloning success was evaluated by colony PCR (Table 9.14). For this purpose a swap of the colony was suspended in 10 µL ddH_2O and used as template in a three step PCR using Taq polymerase (New England Biolabs). Positive clones were cultivated in a 5 mL preculture with LB medium containing 50 mg L^{-1} ampicillin at 37 °C and 250 rpm overnight. Afterwards, plasmid DNA was extracted using the QIAprep Spin Miniprep kit (Quiagen). DNA sequencing was carried out by Eurofins MWG Operon. Finally, positive samples were transformed in *E. coli* BL21 (DE3) for expression.

4.3.4 Expression and purification

Proteins expression was performed in 1 L standard shaking flasks with 200 mL terrific growth medium containing 50 mg L^{-1} ampicillin for selection. After inoculation with 5 mL overnight preculture, the cells were incubated at 37 °C and 250 rpm up to an OD_{600} of 1. Protein expression was induced with 1 mM isopropyl-thiogalactopyranoside. After 4 h cultivation at 30 °C, the cells were pelleted at 3250 g, 20 min at 4 °C. After resuspension in 10 mL 100 mM TRIS·HCl, pH 7.5, the cells were aliquoted and stored at −80 °C.

Purification of His-tagged lipase: After thawing and centrifugation (3250 g, 5 min at 4 °C) the pellet was resuspended in 8 mL binding buffer (20 mM phosphate, pH 7.4, 30 mM NaCl, 500 mM NaCl, 8 M urea, 5 mM DTT). Cells were disrupted with a W450E Branson Sonifier® (Branson) using 30 % of the maximal amplitude and 6 pulses of 15 s. After each pulse, the sample was cooled on ice for 30 s. Subsequent to rigorous mixing for 15 min and centrifugation (3250 g, 5 min at 4 °C), the supernatant containing the soluble protein was purified with immobilized metal ion affinity chromatography (Hochuli *et al.*, 1987) using His-Trap columns with 1 mL column volume (His-trap FF crude, GE Healthcare). Subsequent to column equilibration (binding buffer), the sample was applied using a flow rate of 1 column volume min^{-1}. The protein of interest was eluted with (20 mM phosphate, pH 7.4, 500 mM NaCl, 500 mM NaCl, 8 M urea, 5 mM DTT) and fractions containing protein were collected. All buffers contained denaturizing agents (urea, DTT), to ensure inclusion body solubilization and protein denaturation.

Purification of non-tagged lipase: After thawing and centrifugation (3250 g, 5 min at 4 °C), the pellet was resuspended in 8 mL phosphate buffer (PB) with 20 mM phosphate at pH 7.4 and 5 mM DTT. The cells were disrupted with a sonicator (see above). Possible contaminants were removed by successive washing with 8 mL buffer and subsequent centrifugation. Step 1: PB. Step 2: PB with 2 % v/v TRITON-X 100. Step 3: PB with 2 % v/v TWEEN 20. Finally, the inclusion bodies were resuspended and solubilized in 8 mL binding buffer under rigorous mixing (15 min).

4.4 Basic calculations

4.4.1 Ionic strength computation

Ionic strength of the refolding buffer was calculated as the sum of all ionic components at the experimentally determined pH. First, the Henderson-Hasselbach equation (Equation 13) was applied on all relevant compounds (buffer substances, amino acids). Thereby, small (µm to mM) concentrations of detergent, redox substances and protein were neglected.

$$pH = pK_a + log\frac{[A^-]}{[HA]} \qquad \textit{(Equation 13)}$$

with	pH	*pH*	-
	pK_a	*acid dissociation constant*	-
	A^-	*acid, deprotonated*	-
	HA	*acid, protonated*	-

Input for the Henderson-Hasselbach equation was the experimentally derived pH and pK_a values from the literature. For amino acids with multiple acidic / alkaline groups all charged species were considered. However, zwitterions without net charge were omitted as they are not contributing to the ionic strength (Stellwagen *et al.*, 2008). The result was a list of all ionic species. This list was supplemented with the amount of added titration substance (HCl, NaOH). Finally, the ionic strength was calculated as the sum of all charged species (Equation 14).

$$I = \frac{1}{2}\sum_{i=1}^{n} c_i \cdot z_i^2 \qquad \textit{(Equation 14)}$$

with			
	I	*ionic strength*	*mol L⁻¹*
	c_i	*concentration of compound i*	*mol L⁻¹*
	z_i	*charge number of compound i*	-

4.4.2 Refolding yields

The activities of native and refolded proteins were determined in the respective functional assays (4.1.3). Linear kinetics were applied and the slope was used to calculate the enzymatic activity. Afterwards, the refolding yield was determined as the quotient of the refolded activity and the activity of the native enzyme diluted in the same refolding buffer.

$$refolding\ yield = \frac{refolded\ activity}{native\ activity} \qquad \textit{(Equation 15)}$$

4.4.3 Experimental costs

As certain refolding additives, for example arginine and redox agents, were expensive compounds, the overall costs of the refolding buffer was considered. Using the pricing of the provider, individual costs of the respective compounds were summarized and indicated as overall costs of the respective refolding buffer (Equation 16).

$$costs = \sum_{i=1}^{n} c_i \cdot M_i \cdot p_i \qquad \textit{(Equation 16)}$$

with			
	costs	*experimental costs*	*€ mL⁻¹*
	c_i	*concentration of compound i*	*mol L⁻¹*
	M_i	*molar mass of compound i*	*g mol⁻¹*
	p_i	*price of compound i*	*€ g⁻¹*

4.5 Design of experiments (DOE)

Two different design of experiments (DOE) strategies were used in this work. A stochastic, heuristic design based on a genetic algorithm (GA) and a classic statistical design incorporating a D-optimal screening step and a subsequent optimization with response surface methodology (RSM). Both strategies are depicted in this subchapter.

4.5.1 Genetic algorithm

Protein refolding was optimized iteratively with a multi-objective GA. This subchapter depicts implementation and encoding. The experimental procedure is detailed in 4.1.

Optimization algorithm

For this project the *strength pareto evolutionary algorithm* (SPEA 2) was used (Zitzler, 1999). SPEA 2 is a multi-objective algorithm, allowing a simultaneous optimization of refolding yield, activities or cost. SPEA 2 was already applied for a variety of experimental problems (Gobin *et al.*, 2007; Gobin and Schüth, 2008; Havel *et al.*, 2006). The limiting factor for experimental optimizations are the number of experiments and the experimental error. Consequently, an algorithm that was previously used on similar problems with known parameters and settings was considered favorably. Furthermore, both the source code and a user friendly program with a graphical user interface (GUI) was available.

SPEA 2 was implemented in Matlab (Mathworks, R2009a) and an Excel (Microsoft, 2003) based file exchange was established. SPEA 2 was used with the following optimization parameters: population size 22, crossover points two, mutation rate one percent per bit, other parameters were left default (Zitzler *et al.*, 2002). The fitness evaluation was done experimentally. An alternative implementation with a GUI was based on the program GAME.opt (Link and Weuster-Botz, 2006). This method included an Excel file with the coded variables and substitution functions for “wrong” combinations of pH and buffering agent. However, it deviated from the Matlab based method as no repair function could be implemented in Excel. Therefore the variable encoding was slightly divergent.

Encoding of the experimental problem

Critical for the optimization success with a GA is the encoding, which is the representation of the problem specific decision space by a data structure. In this thesis, a

classical binary vector was used. As the experimental problem at hand consisted of two parts, this choice had to be evaluated for both.

First, the variable combination. The binary representation is generally used for encoding combinatorial problems (Back *et al.*, 1997b; Michalewicz, 1999). While the most direct way of binary encoding, the usage of one bit for each element was investigated (Wolf *et al.*, 2000), it is more efficient to use constraints (see below) and limit possible combinations. Hence, experimental knowledge of the problem was used to limit the search space and problem complexity. Second, the encoding of the variable values and concentrations. This part constituted a continuous problem. Hence, floating point vectors were an obvious choice as data structures. However, binary vectors have distinct advantages: Above all, a uniform and consistent representation of the problem should be the aim. It was considered suboptimal to split the problem in a combinatorial (binary) and a continuous (floating points) part as described by Wolf *et al.* (2000), since it would be difficult to generate a recombination operator for the whole vector. In addition, a continuous representation is only sensible if the experimental error of the design variable is close to zero. This is the case for most *in silico* optimizations but generally not for experimental problems. For experimental problems, the search space is discretized using a step size smaller than the experimental error. A discretization also reduces the accuracy and thus the search space enabling a faster convergence of the GA (Link and Weuster-Botz, 2006).

A total of 26 variables (see Table 4.4 in the experimental section) constituted the refolding problem. These were classified in 6 groups and boundary conditions were introduced to incorporate biochemical knowledge.

- Functionally related variables were subgrouped in the following categories: buffers, salts, additives, mineral ions, detergents, redox agents (see 4.1.2).
- Minimal requirement for a refolding condition was a buffer substance and the associated pH, for example TRIS·HCl, pH 8.0.
- Additions of components from all other groups were allowed, resulting in complex refolding conditions like TRIS·HCl, pH 8.0 with 100 mM NaCl, 100 mM arginine, 5 mM DTT.
- In order to screen for synergistic interactions, combinations inside one functional class were allowed in several cases for example both glutamine and arginine (Dashivets *et al.*, 2009). A repair function was introduced to remove infeasible variable combinations. The latter were replaced with new (random) solutions.

- Finally, the discretization of the 26 variables was critical. Refolding literature and the REFOLD database (Amin *et al.*, 2006; Buckle *et al.*, 2005) was evaluated in terms of variable concentrations, variance and influence. For some variables, for example detergents, only the presence or absence in the refolding buffer was important. Consequently, only three detergent concentrations around the CMC were examined. Other variables, like NaCl or arginine, were subjected to a detailed analysis with a far higher resolution.

The entire optimization problem was coded in bit form with a length *(L)* of the binary string of 32. Considering the number of refolding conditions (M = 22) analyzed in every generation, it was possible to calculate the probability to reach each point in the search space via crossover (Equation 17) (Reeves, 1993). p indicates if the chosen population size was adequate for the complexity of the search space. With the given setup it would even be possible to expand the design space, for example testing more substances or concentrations in future optimizations.

$$p = (1 - 0.5^{M-1})^{L} \geq 99.99\ \% \qquad \textit{(Equation 17)}$$

with	p	*probability to reach each point in the search space*	-
	L	*length of the binary string*	-
	M	*population size*	-

4.5.2 Statistical design of experiments

In addition, to the stochastic optimization (see above), a classic statistical DOE was utilized. This design consisted of two steps: a D-optimal screening step and a subsequent RSM based optimization.

D-optimal screening

First, the parameter space of the stochastic optimization was translated into a D-optimal statistical design with 27 variables and a linear model (see appendix Table 9.5). The Matlabs (Mathworks) coordinate exchange algorithm *cordexch* was used to generate the experimental design matrix. This algorithm constructed an initial design matrix X that was optimized iteratively to increase the determinant $D = |X^T X|$, thereby minimizing the covariance. As the initial design was random, design solutions might be locally, but not globally D-optimal. Therefore, the method was repeated 25 times and the best try was selected and subsequently verified experimentally. Refolding experiments were carried out in the corresponding design buffers and native and refolded activities were

measured threefold. Afterwards, the linear regression model (Equation 7) was applied for both activities and non-significant terms were iteratively removed. The remaining predicators with the highest impact (regression coefficients) were subject to the following RSM optimization. DOE data are summarized in the appendix: the problem encoding (Table 9.5), the design matrix with the experimental results (Table 9.6) and the regression matrix (Table 9.7).

Response surface methodology

A circumscribed central composite design type (compare Table 9.4) with a second order polynomial model was used to optimize the remaining predicators. The Matlabs (Mathworks) *ccdesign* function was used to generate the design. Predicator variables were coded appropriately (−2, −1, 0, 1, 2). Subsequently, refolding was evaluated in the designed refolding conditions with the standard setup (see 4.1). Instead of only repeating the center point, all data points were measured threefold. The measured native and refolded enzyme activities were used to fit second order polynomial models (Equation 8). Afterwards, non-significant terms were iteratively removed. RSM data and the measured activities are detailed in the appendix (Table 9.8).

4.6 Black-box models for data analysis

Data from refolding experiments were stored in a relational database (MySQL 5.1). Data analysis and modeling was carried out in Matlab (Mathworks), while the *Database Toolbox* enabled direct access to the database. A variety of standard methods from data mining and modeling were applied to this dataset. These are explained in the following sections.

4.6.1 Artificial Neural networks (ANNs)

ANNs are a standard modeling technique that is applied for both classification and regression. In this thesis, a variety of network sizes, architectures and function were examined. Creation, testing and validation of the ANNs were done with the Matlabs *Neural Network Toolbox*.

Input / Output

Input for the ANNs was the normalized refolding data from one protein optimization, that is the composition of the refolding buffer (see Table 5.12). In addition, parameters like the ionic strength of the buffer and protein-specific variables were examined. The models predicted the normalized refolding success: either the activities or the refolding yield (see Table 5.11).

Creation

Modeling started with a simple network architecture: feedforward, two layers, 10 to 25 neurons in the hidden layer and a sigmoid transfer function. If the performance was unsatisfactory, changes were done according to the following priorities: Network size (more neurons), architecture (more layers), training function, network type and data division (random vs. fixed). Weights were initialized before training network.

Training

Network training (adjustment of the weight matrix) generally used the Levenberg-Marquardt algorithm and backpropagation. 70 % of the dataset was used for training, the rest was omitted for validation and testing. Training performance was measured as the mean squared error (Equation 10) of the validation data.

Validation

All models were subject to an independent validation. Therefore, parts of the original dataset (typically 30 %) were retained.

4.6.2 Bagged decision trees (BDT)

The Matlabs *TreeBagger* function was used to generate an ensemble of bagged decision trees (BDT). Individual trees were grown on independently-generated bootstrap replicas of the data. The part of the data that was not included in a replica was “out-of-bag” regarding the respective tree. Ensemble predictions were the averaged prediction from all individual trees.

Input

Equivalent to the ANNs (see 4.6.1).

Creation and Training

First, a small tree model (50 trees) was generated to identify the optimal *leaf* size. Afterwards, ensemble models with 100 to 250 trees were generated. In comparison to the ANNs, the ensemble offers an additional quality parameter. The out-of-bag mean squared error enables an estimation of the true ensemble error and was used as a measure of the model performance. Nevertheless, an experimental validation was also performed for the final model (see below).

Validation

Preliminary models were not validated. Instead the models were trained on the entire dataset and the out-of-bag-error was the sole performance criteria. For the refined model, an independent validation with 88 refolding experiments from the final LIP optimizations was carried out.

5 Results and Discussion

Within the scope of this thesis a standardized optimization strategy for protein refolding conditions was developed. This chapter contains the experimental and computational results and is structured as follows: In section 5.1, the experimental optimization of the refolding conditions of a variety of model proteins is depicted successively. Afterwards, the optimization strategy is further evaluated and compared to standard statistical approaches (section 5.2). Finally, section 5.3 details data analysis and modeling.

5.1 Experimental optimization of protein refolding

Protein refolding from the denatured state constitutes a complex problem with a great variety of variables and is strongly dependent on the examined protein. Therefore, optimal refolding conditions are typically determined experimentally (Clark, 2001; Lilie *et al.*, 1998; Middelberg, 2002). In this thesis, a genetic algorithm (GA) was used to iteratively optimize protein refolding using a standardized experimental design in mL-scale and 96-well plate format. For this purpose experimental parameters were extracted from the refolding literature and combined with the information on approximately 1100 refolding experiments from the REFOLD database (Amin *et al.*, 2006; Buckle *et al.*, 2005) to establish a comprehensive experimental design, which is presented in detail in the previous material and methods chapter.

Figure 5.1 illustrates the iterative character of the design strategy. At the beginning of each optimization, 22 refolding conditions were randomly generated and subsequently evaluated experimentally by diluting the denatured protein into the respective refolding condition. Depending on the results, fitness values were assigned to each condition. Similar to evolution, experimental conditions with high fitness were subsequently selected to calculate a new set of refolding conditions. These new refolding conditions, which are based on the most efficient solutions of the previous set, were evaluated experimentally again. The optimization was terminated if no increase in performance was determined in several iterations or a fixed limit of experiments was reached.

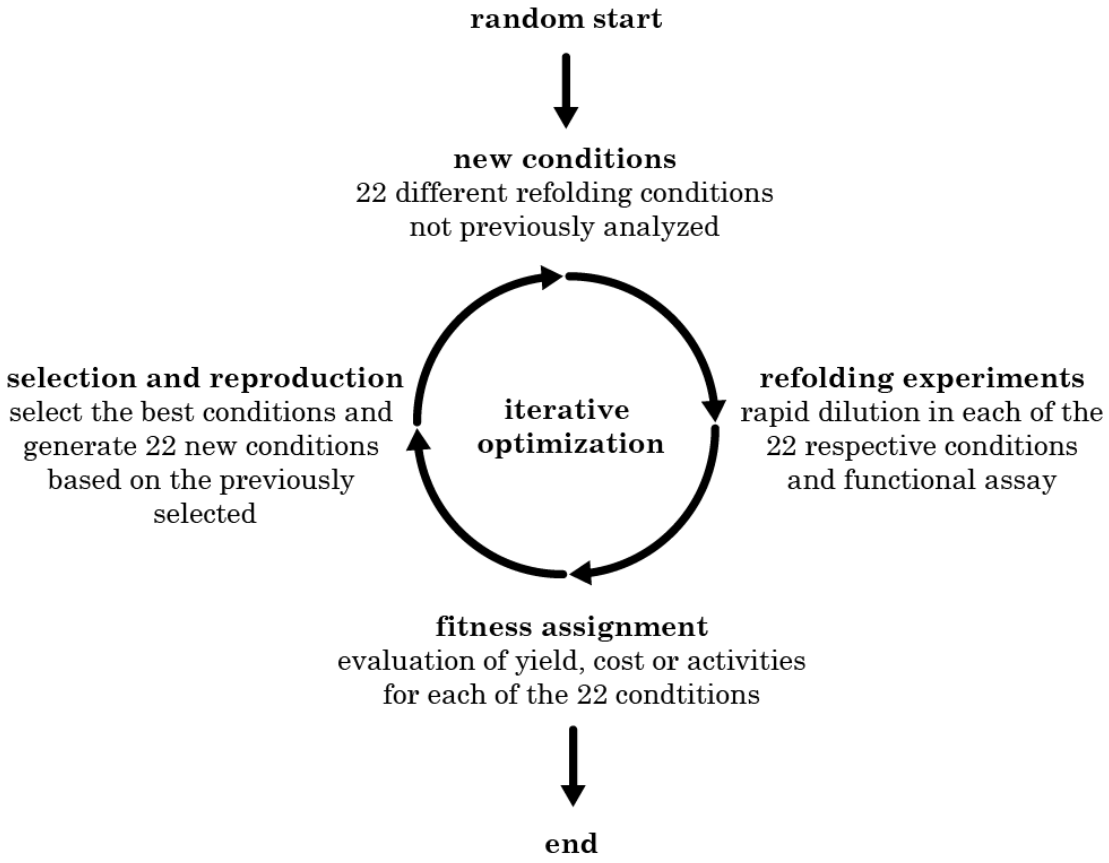


Figure 5.1: Scheme of the novel stochastic optimization strategy for protein refolding in mL-scale and 96-well plate format.

The proposed optimization strategy was evaluated with six functionally and structurally different model proteins (Table 5.1). In the following subsections results of the individual proteins will be presented successively. In this regard, several aspects have to be pointed out.

Protein-specific functional assays were used to quantify the refolding success (see 4.1.3). For the experiments, both denatured and native proteins were diluted into the respective refolding conditions. Afterwards, refolding yields were calculated as the ratio of the activity of the refolded protein and the native protein in the respective refolding condition (Equation 15). Measurements of the first four proteins (compare Table 5.1) were carried out by the cooperation partner (Department Chemie, Center for Integrated Protein Science, Technische Universität München).

An experimental reference was deemed necessary as refolding method, temperature and protein concentration influence the refolding success. Hence, a direct comparison to literature values is problematic. The reference was a known literature refolding condition, which was examined in each iteration of the optimization under the same

experimental conditions. It was not used as a starting point for the optimization. Instead it served as a comparison and as a measure for the experimental error.

A varying number of experiments was performed for each protein, as the termination criteria was a lack of progress. In GA terminology, an iteration is generally referred to as a generation (GEN), which will be consecutively numbered in roman numerals (I to X).

Table 5.1: Overview of analyzed proteins. GFP, GLK GLR and LYZ measurements were carried out by the cooperation partner. (qs*) quaternary structure, (ds*) disulfide bonds.

Abbr.	Protein	M, kDa	pI	qs*	ds*	Activity	Organism
GFP	Green fluorescent protein	28	5.7	monomer	-	intrinsic fluorescence	*Aequorea victoria*
GLR	Glutathione reductase	53	7.7	dimer	-	reduction of glutathione disulfide	*Saccharomyces cerevisiae*
GLK	Glucokinase	35	6.1	dimer	-	phosphorylation of glucose	*Escherichia coli*
LYZ	Lysozyme	14	9.3	monomer	4	hydrolysis of peptidoglycan linkages	*Gallus gallus*
LDH	Lactate dehydrogenase	36	8.2	tetramer	-	reduction of pyruvate	*Oryctolagus cuniculus*
LIP	Lipase	29	5.0	monomer	3	hydrolysis of triacylglycerol	*Thermomyces lanuginosus*

5.1.1 Green fluorescent protein from *Aequorea victoria* (GFP)

Green fluorescent protein (GFP), that exhibits fluorescence under exposure to blue light, constitutes an important reporter and biosensor in molecular biology (Chalfie *et al.*, 1994). In contrast to the other analyzed proteins (Table 5.1), GFP has no enzymatic function. Refolding is quantified by measuring the intrinsic fluorescence. During the optimization process, both refolding yield (Equation 14) and the experimental costs of the refolding buffer (Equation 16) were optimized simultaneously.

Experimental data of the stochastic optimization were plotted according to the two objectives and the respective GENs to provide an overview of the optimization progress (Figure 5.2). In each GEN a standard refolding condition from the literature (Dashivets et al., 2009) was evaluated as an experimental control. This reference was inside the search space of the stochastic optimization. Its mean is depicted as a star. In addition, the optimization progress is highlighted by black dashed lines. The general aim of the optimization was to identify refolding conditions with high yields and low costs, that are conditions in the upper right corner of the graph. GFP showed a steady and fast increase in both objectives, achieving 100 % refolding yield in GEN_{IV}. Further GENs led only to an improvement of experimental costs. The optimization was terminated after GEN_{VI}.

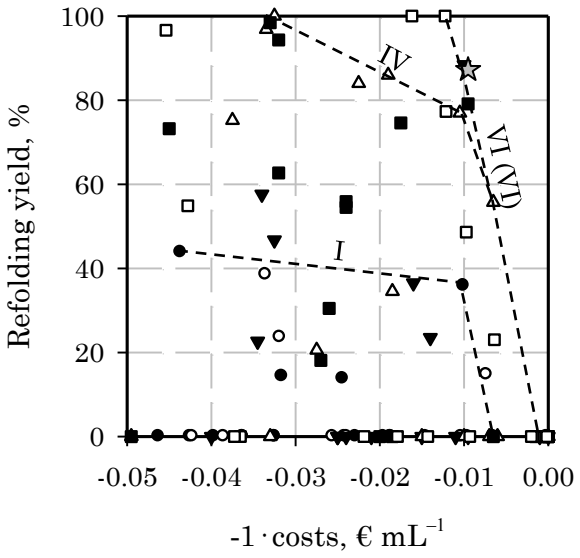


Figure 5.2: Overview of the GFP optimization. Experimental data of the individual GENs (I ●, II ○, III ▼, IV Δ, V ■, VI □) were plotted according to the two objectives, only conditions with costs smaller than 0.05 € mL are displayed. The star (☆) represents an experimentally verified standard refolding condition (Dashivets *et al.*, 2009). In addition, the optimization progress from the start (I) to the end (VI, last improvement in VI) is highlighted for several GENs by black dashed lines.

GAs are based on a set of feasible solutions (population) and exhibit a competition between selection and evolutionary pressure on the one hand and maintenance of variance on the other hand (see 3.3.2). Hence, a trend towards the optimization objectives is observable if the optimization progresses.

During the optimization, the median costs of GFP decreased steadily in each GEN, from 0.041 € mL^{-1} in the first GEN to 0.015 € mL^{-1} in GEN_{VI}. The refolding buffer of the reference condition (Dashivets *et al.*, 2009) amounted to 0.010 € mL^{-1}. For the refolding yields, a trend to higher yields was observed until GEN_{IV} (Figure 5.3).

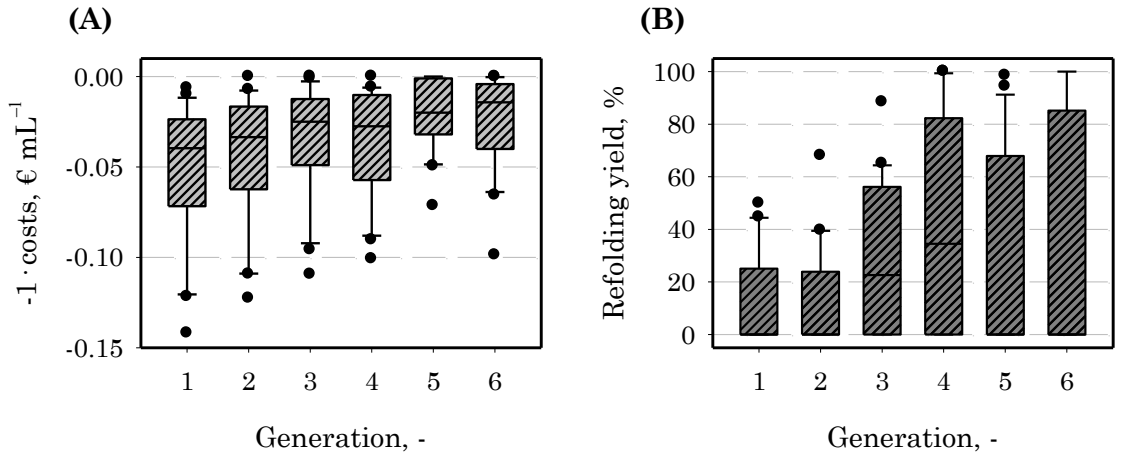


Figure 5.3: Development of the two objectives during the GFP optimization illustrated as box plots: (A) experimental costs, (B) refolding yield. The boxes contain the middle of 50 % of the data, whiskers denote the 10th and 90th percentiles. (●) outliers, (—) median.

Fluorescence of the native and refolded protein was evaluated in triplets for each refolding condition. Due to the large errors of both measurements, the calculated relative refolding yield featured high standard deviations of 30 % to 50 % (data not shown). The reference condition (Dashivets *et al.*, 2009) was measured along in each GEN. Here, the experimental error of each measurement was smaller (about 10 %) and generally comparable over the GENs. But large deviations could be observed for the refolded activity in GEN_I and GEN_{IV} (Figure 5.4).

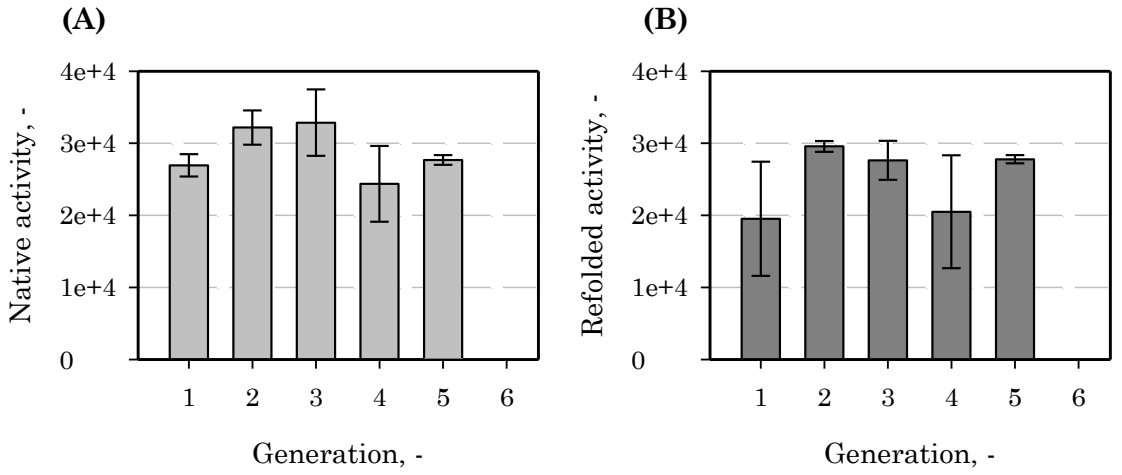


Figure 5.4: Error of the fluorescence measurements of GFP in the reference refolding condition (Dashivets *et al.*, 2009) during the optimization. (A) native activity, (B) refolded activity.

5.1.2 Glutathione reductase from *Saccharomyces cerevisiae* (GLR)

Glutathione reductase (GLR) maintains high levels of reduced glutathione in the cytosol of eukaryotes. In contrast to GFP, GLR is a dimer with a mass of 53 kDa. Analog to the previous optimization, experimental costs and refolding yield were optimized in parallel (Figure 5.5). A very fast optimization was observed, as the first GEN already contained a refolding condition with 100 % yield. However, the respective refolding buffer was expensive (0.075 € mL^{-1}). In the following GENs experimental costs were optimized, leading to improved refolding conditions with 100 % yield and reduced costs of 0.006 € mL^{-1}.

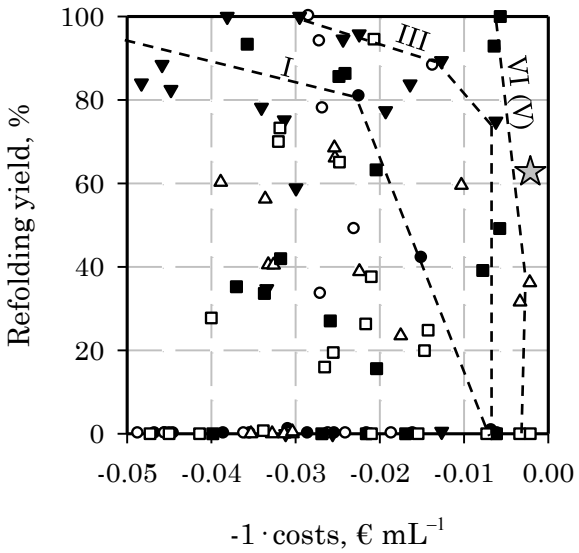


Figure 5.5: Overview of the first GLR optimization. Experimental data of the individual GENs (I ●, II ○, III ▼, IV Δ, V ■, VI □) were plotted according to the two objectives, only conditions with costs smaller than 0.05 € mL are displayed. The star (☆) represents an experimentally verified standard refolding condition (Nordhoff *et al.*, 1997). In addition, the optimization progress from the start (I) to the end (VI, last improvement in V) is highlighted for several GENs by black dashed lines.

The development of the two objective functions during the optimization is illustrated in Figure 5.6. On the one hand, a trend towards lower average costs was observed in each GEN. Median costs were reduced from 0.045 € mL^{-1} in the first GEN to 0.025 € mL^{-1} at the end. The reference refolding condition featured costs of 0.002 € mL^{-1}. On the other hand, maximal refolding (100 %) was detected in nearly all generations, but there was no increase in average performance, as a lot of refolding conditions showed no refolded activity.

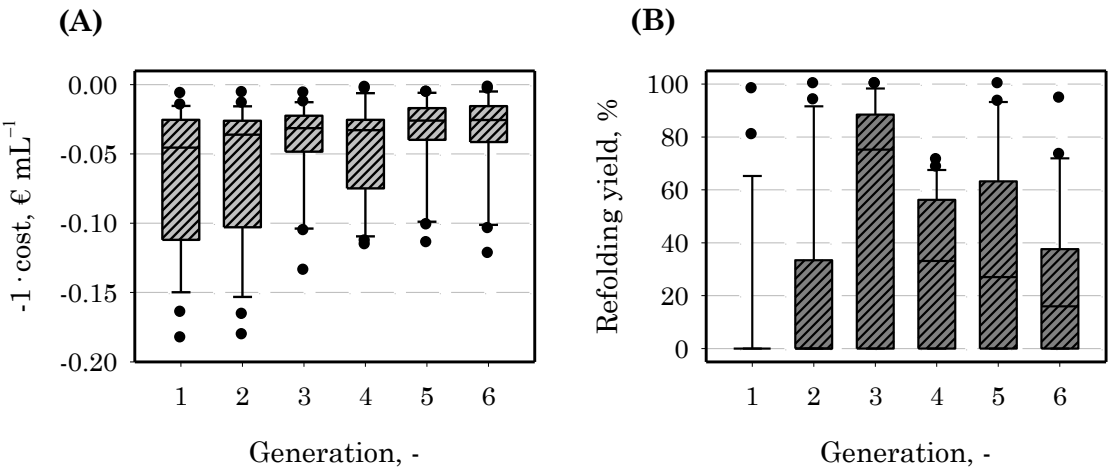


Figure 5.6: Development of the two objectives during the first GLR optimization illustrated as box plots: (A) experimental costs , (B) refolding yield. The boxes contain the middle of 50 % of the data, whiskers denote the 10th and 90th percentiles. (●) outliers, (—) median.

During the stochastic optimization, GLR exhibited many refolding conditions with 100 % refolding yield. However, the respective activities itself varied between 40 U mg^{-1} and 100 U mg^{-1} depending on the respective buffer condition (exemplified in Figure 5.7).

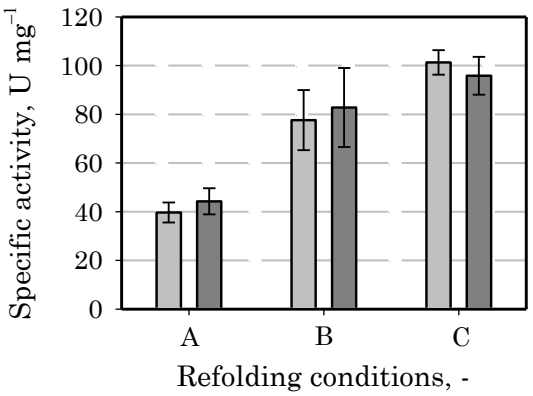


Figure 5.7: Specific activities of the native (grey) and refolded (dark grey) GLR in three different refolding conditions (A, B, C). (A) 1 M TRIS·HCl, pH 8.5, 150 mM NaCl, 10 % v/v glycerol, 500 mM arginine, 100 mM glutamine, 2 mM EDTA; (B) 100 mM PB, pH 7.5, 250 mM NaCl, 20 mM KCl, 500 mM arginine, 100 mM glutamine, 2 mM EDTA, 5 mM GSH; (C) 100 mM MOPS, pH 8.5, 150 mM NaCl, 20 mM KCl, 500 mM arginine, 50 mM glutamine, 5 mM EDTA.

All illustrated refolding conditions (Figure 5.7) showed 100 % yield but very different specific activities. Consequently, it was not possible to differentiate conditions with 100 % yield solely based on the relative refolding yield. Instead, the specific activity of the refolded protein under the refolding conditions offered more information. Therefore,

a second independent optimization with modified objectives (native and refolded activity instead of refolding yield and cost) was performed for GLR in order to obtain buffers with highly active protein and maximum refolding yield (Figure 5.8).

In contrast to the previous optimizations of GFP and GLR, the aim was to identify refolding conditions with high native and refolded activities. Ideal conditions would be in the upper right corner of the graphs and close to the bisecting line which indicates 100 % refolding yield. Figure 5.8 shows an increase in the best native and refolded activity until GEN_{VI}. The activities of the optimal refolding conditions (refolded activities of 100 U mg^{-1} to 120 U mg^{-1} and roughly 100 % yield) were comparable to those obtained in the first optimization. As no further improvement could be detected in the last two GENs, the optimization was terminated in GEN VIII.

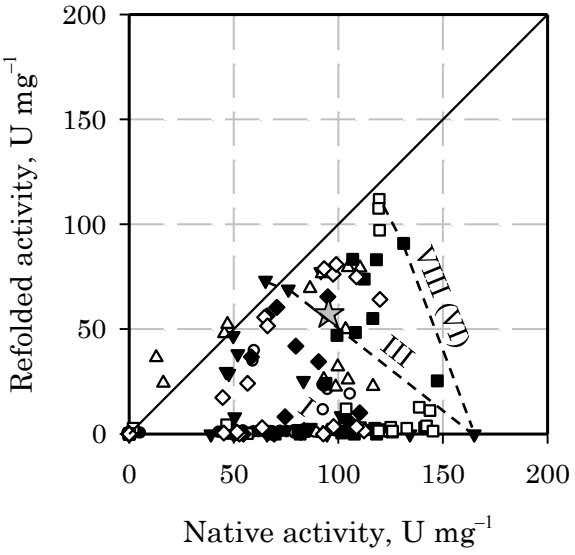


Figure 5.8: Overview of the second GLR optimization. Experimental data of the individual GENs (I ●, II ○, III ▼, IV Δ, V ■, VI □, VII ♦, VIII ◊) were plotted according to the two objectives. The star (☆) represents an experimentally verified standard refolding condition (Nordhoff *et al.*, 1997). The bisecting line denotes 100 % refolding yield and therefore the best refolding buffers at different activities. In addition, the optimization progress (last improvement in VI) is highlighted for several GENs by black dashed lines.

With regard to the native activity an unsteady trend towards higher activities was observed until GEN_{VI}, afterwards the values stagnated or decreased. The median of the native activity increased from 75 U mg^{-1} to 120 U mg^{-1}. Progress of the refolded activity was erratic, as only the outliers exhibited increased activities (Figure 5.9).

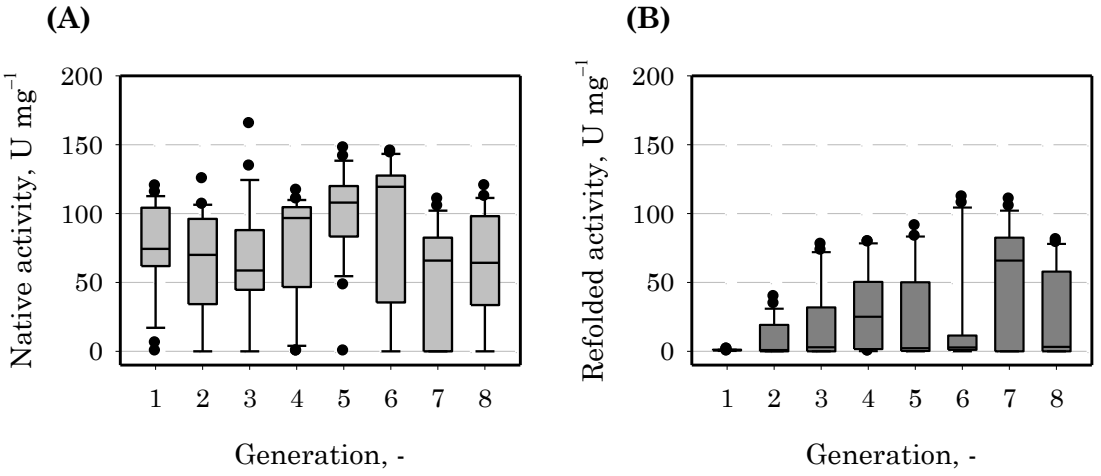


Figure 5.9: Development of the two objectives during the second GLR optimization illustrated as box plots: (A) native activity , (B) refolded activity. The boxes contain the middle of 50 % of the data, whiskers denote the 10th and 90th percentiles. (●) outliers, (—) median.

The experimental error of the GLR activity measurements was generally less than 15 % (data not shown). However, sequential reproducibility was problematic, as indicated by the error of the reference condition, which was measured in each GEN (Figure 5.10).

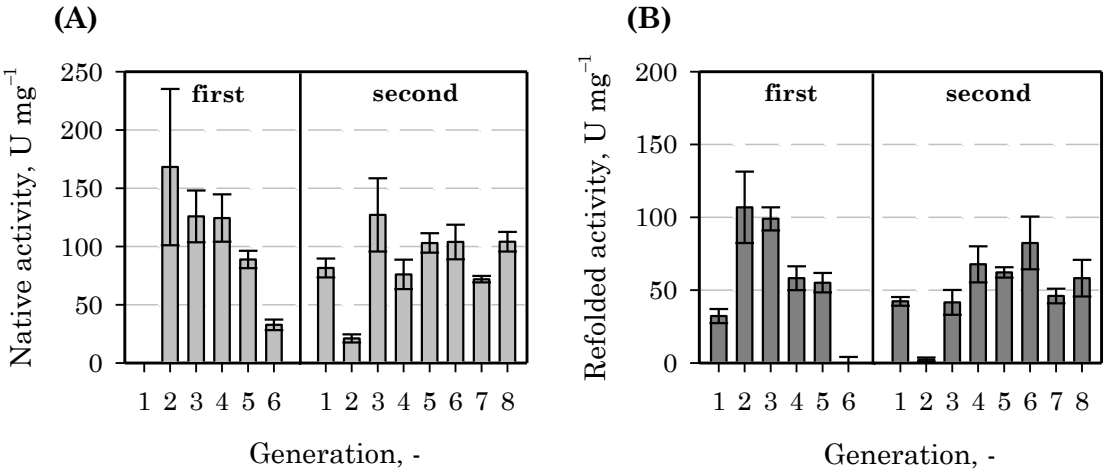


Figure 5.10: Error of the activity measurements of the reference refolding condition (Nordhoff *et al.*, 1997) during the first and second optimization of GLR. (A) native activity, (B) refolded activity, white bars were classified as outliers.

5.1.3 Glucokinase from *Escherichia coli* (GLK)

Glucokinase (GLK), a dimeric protein with a mass of 35 kDa, phosphorylates glucose in the first step of the glycolysis and plays a critical role in the regulation of the carbohydrate metabolism. Refolding of GLK was optimized with native and refolded activity as objectives (Figure 5.11). Both native and refolded activity increased during the optimization from maximal values of 100 U mg^{-1} in GEN$_I$ to 310 U mg^{-1} at the end. Furthermore, 100 % refolding yield could be achieved in 40 % of the examined conditions (points near the bisecting line).

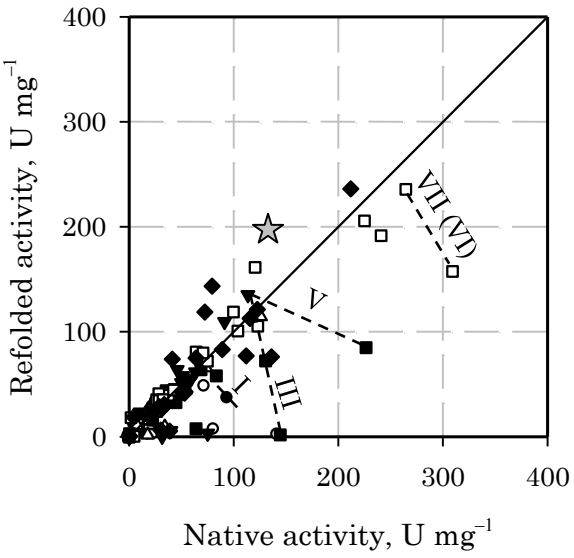


Figure 5.11: Overview of the GLK optimization. Experimental data of the individual GENs (I ●, II ○, III ▼, IV Δ, V ■, VI □, VII ♦) were plotted according to the two objectives. The star (☆) represents the buffer of the GLK functional assay and the bisecting line denotes 100 % refolding yield. In addition, the optimization progress (last improvement in VI) is highlighted by black dashed lines.

During the optimization a trend towards higher activities was observed until GEN$_{VI}$, with a very similar development for both activities. The median activities increased from 25 U mg^{-1} to 50 U mg^{-1}. Progress was unsteady and many outliers with high activity data (up to 310 U mg^{-1}) occurred (Figure 5.12).

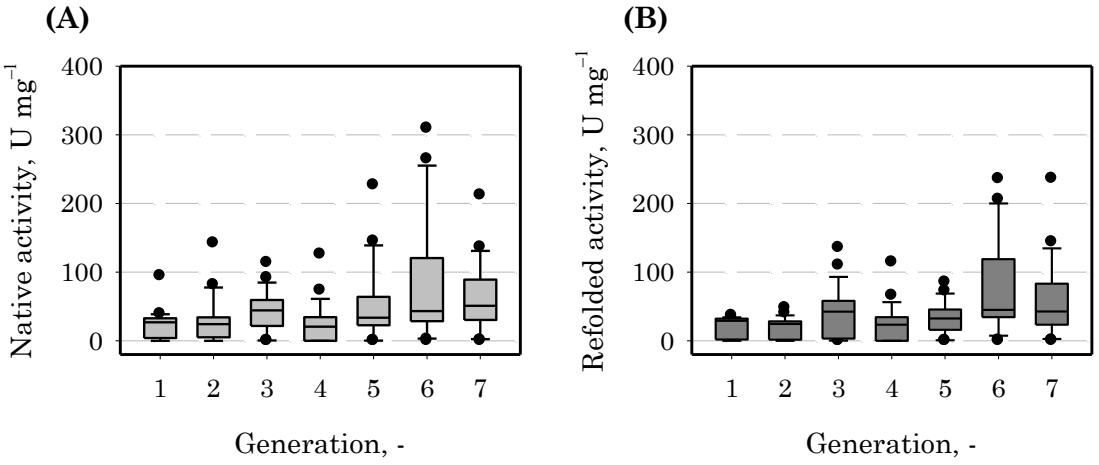


Figure 5.12: Development of the two objectives during the GLK optimization illustrated as box plots: (A) native activity , (B) refolded activity. The boxes contain the middle of 50 % of the data, whiskers denote the 10th and 90th percentiles. (●) outliers, (—) median.

As mentioned above, GLK refolded in many conditions with maximum yield. This was also the case for the buffer of the functional assay (50 mM HEPES, pH 7.5, 150 mM KCl, 10 mM $MgCl_2$), that was included as a reference. In addition, the activities of the reference notably exceeded the above-mentioned median in the optimization. Only the outliers exhibited comparable or higher activities (Figure 5.12). The measurement data of the reference with the respective standard deviations are shown in Figure 5.13. Overall experimental errors of the GLK activity measurements were in general smaller than 20 % (data not shown).

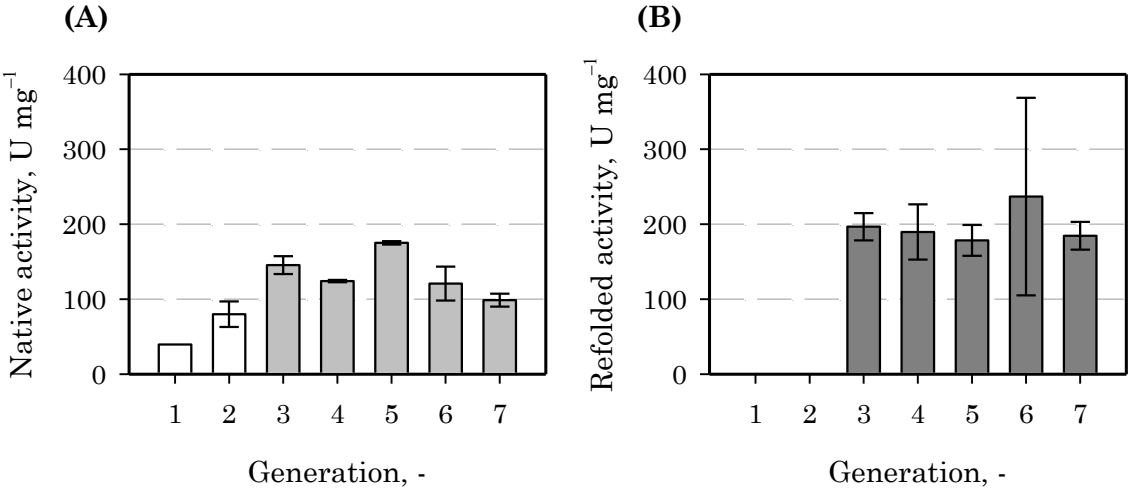


Figure 5.13: Error of the activity measurements of the reference refolding condition (buffer of the GLK functional assay) during the optimization of GLK. (A) native activity, (B) refolded activity, white bars were classified as outliers.

5.1.4 Lysozyme from *Gallus gallus* (LYZ)

Lysozyme (LYZ) is a well-characterized model protein. Compared to the majority of proteins, it is very small (14 kDa) and it features four disulfide bonds. LYZ was first optimized with the standard configuration of the stochastic optimization approach and native and refolded activities as objectives (Figure 5.14, A). However, positive refolding was very sparse, as only two out of 88 conditions showed refolding. These two refolding conditions contained oxidative redox conditions (reduced GSH and oxidized glutathione GSSG). Therefore, a second independent optimization approach was started with a modified configuration (Figure 5.14, B). In this approach constrained redox conditions were utilized. Specifically, purely reductive conditions with tris-carboxyethyl-phosphine (TCEP) or dithiothreitol (DTT) were removed from the setup.

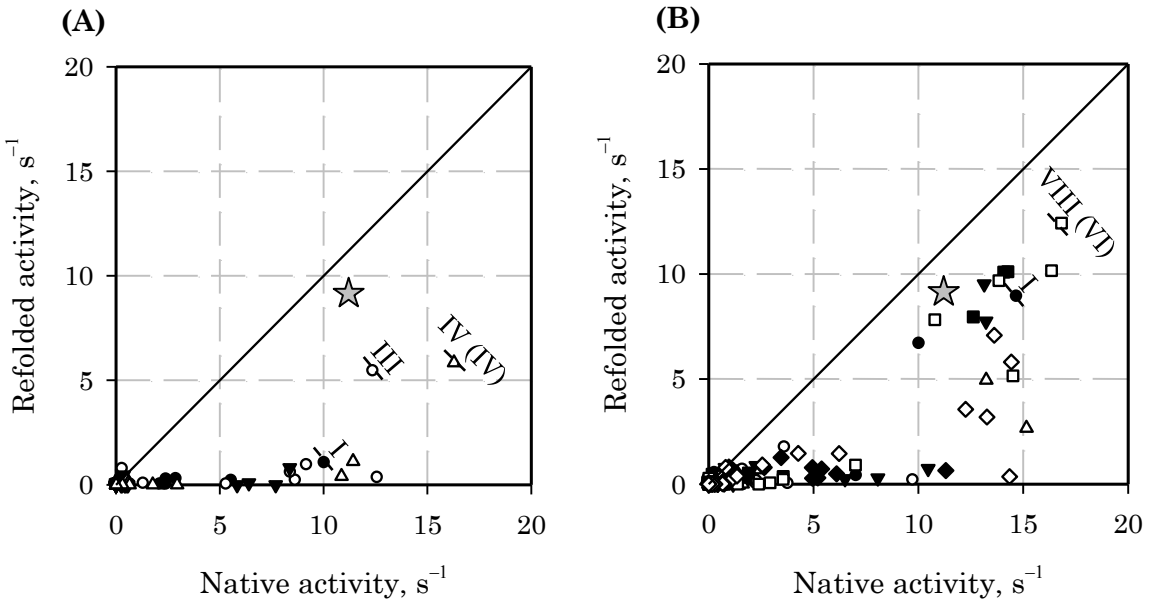


Figure 5.14: Overview of the independent first (A) and second (B, with modified redox conditions) optimizations of LYZ refolding. Experimental data of the individual GENs (I ●, II ○, III ▼, IV Δ, V ■, VI □, VII ♦, VIII ◊) were plotted according to the two objectives. The star (☆) represents an experimentally verified standard refolding condition (Hevehan and De Bernardez Clark, 1997) and the bisecting line denotes 100 % refolding yield. In addition, the optimization progress (last improvement in IV / VI) is highlighted for several GENs by black dashed lines.

Positive data were sparse in both optimizations (Figure 5.14). Most refolding conditions showed no or only native activity. Nevertheless, the second optimization slightly progressed and refolding conditions exhibiting 40 % higher activities than the reference could be detected. Statistics of the optimization objectives are not shown as most data were close to zero and featured no refolding.

LYZ activity was correlated with the ionic strength of the refolding buffer, which was calculated as the sum of all charged species at the experimentally determined pH (Equation 14). Above 0.7 M ionic strength, both native and refolded activity were close to zero (Figure 5.15). Consequently, analog to the reductive conditions in the second optimization, a modified third optimization was performed which incorporated a threshold for the ionic strength.

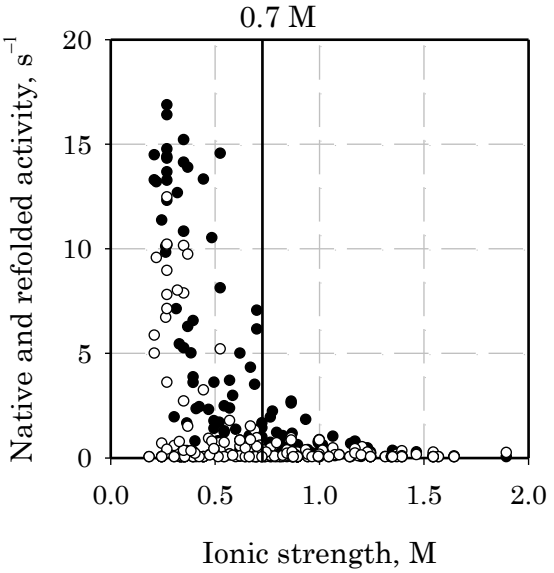


Figure 5.15: Impact of the ionic strength of the refolding buffer (264 experiments) on the LYZ activity. (●) native activity, (○) refolded activity.

In the independent third optimization, far more positive results, that is conditions with both native and refolded activity could be observed (Figure 5.16). The problematic sparsity of positive data in the previous optimizations, in which the majority of the refolding conditions showed no or only native activity (Figure 5.14), did not occur for this modified approach. However, the overall activities were smaller (< 10 s^{-1}) compared to the previous experiments. In addition, the native activity of reference was smaller than previously measured. Therefore, the optimization was terminated after GEN_{III}.

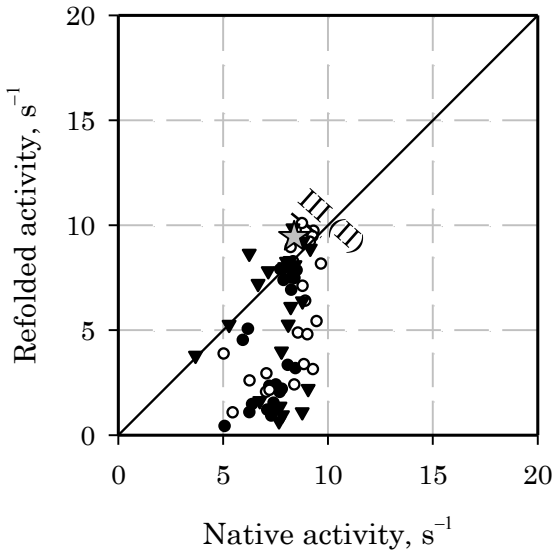


Figure 5.16: Overview of the third LYZ optimization with constrained ionic strength and redox conditions. Experimental data of the individual GENs (I ●, II ○, III ▼) were plotted according to the two objectives. The star (☆) represents an experimentally verified standard refolding condition (Hevehan and De Bernardez Clark, 1997) and the bisecting line denotes 100 % refolding yield. In addition, the optimization progress (last improvement in II) is indicated.

The measurement data of the reference condition (Hevehan and De Bernardez Clark, 1997), which was evaluated in all three optimizations in each GEN, is shown in Figure 5.17. In the third optimization, the mean native activity (9.15 ± 1.3 s^{-1}) was smaller than in optimization one and two (11.2 ± 0.9 s^{-1}), while the refolded activity (9.45 ± 1.0 s^{-1}) was comparable (8.39 ± 0.6 s^{-1}). The overall error of the LYZ activity measurements was 10 % to 15 % (data not shown).

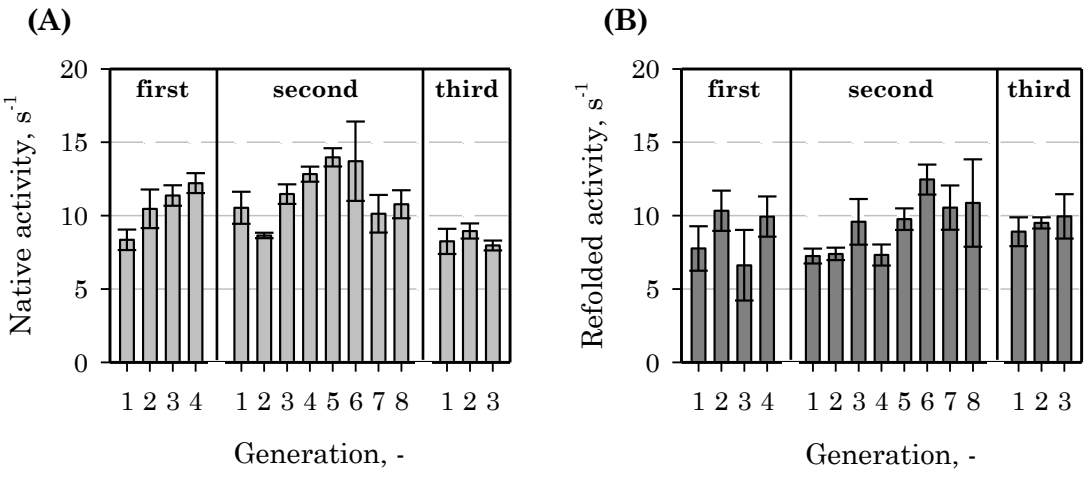


Figure 5.17: Error of the activity measurements of the reference refolding condition (Hevehan and De Bernardez Clark, 1997) during the first, second and third optimization of LYZ. (A) native activity, (B) refolded activity.

5.1.5 Lactate dehydrogenase from *Oryctolagus cuniculus* (LDH)

The tetrameric enzyme lactate dehydrogenase (LDH) catalyzes the conversion of pyruvate to lactate. In higher eukaryotes different tissues may exhibit different LDH subtypes. The analyzed LDH is from rabbit (*Oryctolagus cuniculus)* muscle. In contrast to the previously discussed proteins, refolding experiments with LDH were not performed by the cooperation partner. The LDH assay was adapted from Stambaugh and Post (1966). While the assay proved robust and sensitive, the stability of the protein was low (Figure 5.18). In a buffered solution (pH 7.3), the activity of LDH dropped from 68 U mg^{-1} to 35 U mg^{-1} in 4 h and decreased further. Therefore refolding times were limited to 2 h (compare Table 4.3).

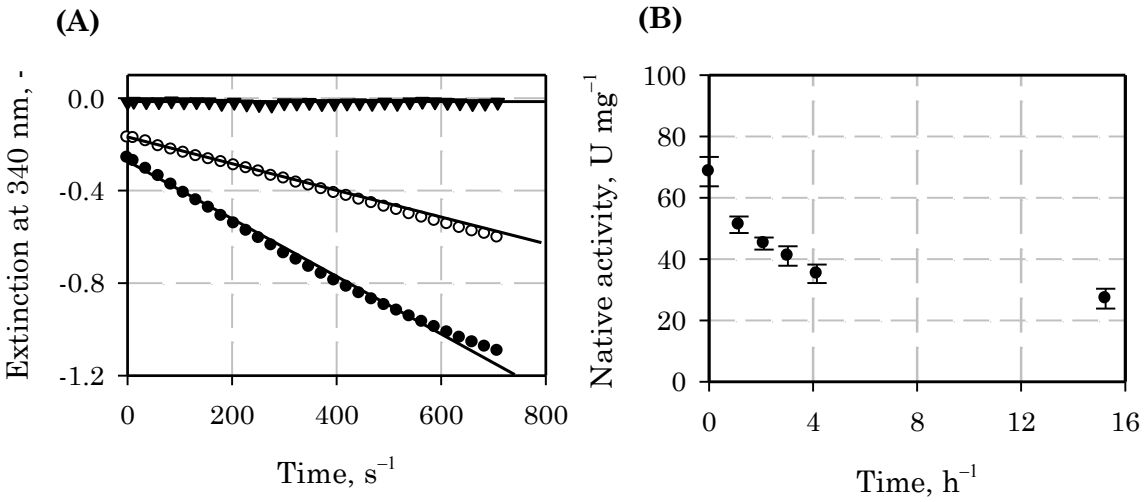


Figure 5.18: LDH activity and stability. (A) Exemplary activity assay of native (●, 0.2 M TRIS·HCl, pH 7.3), refolded (○, 0.2 M TRIS·HCl, 50 mM NaCl, 100 mM arginine and 10 mM DTT, pH 8.75) and denatured (▲) protein, linear regression is indicated. (B) Stability of LDH in solution (0.2 M TRIS·HCl, pH 7.3) at room temperature.

LDH refolding was optimized using a new standard configuration of the stochastic optimization with minor modifications (see 5.1.8) and native and refolded activities as objectives (Figure 5.19). In the stochastic optimization, most conditions exhibited only native activity. Nevertheless, the optimization progressed. At the end (GEN_V) refolding conditions with 40 % higher activities than the reference (Rudolph *et al.*, 1977) could be obtained.

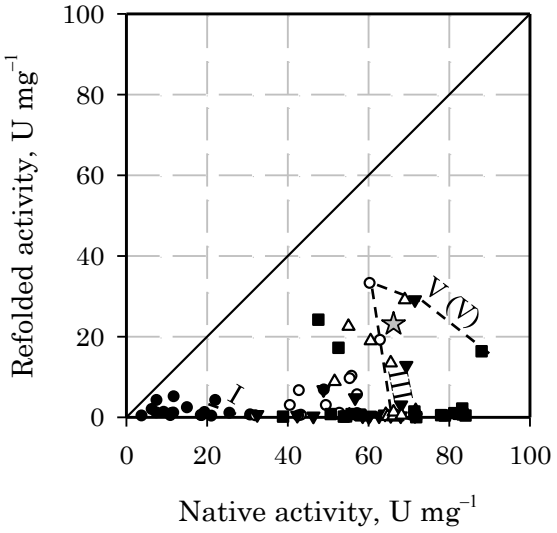


Figure 5.19: Overview of the LDH optimization. Experimental data of the individual GENs (I ●, II ○, III ▼, IV Δ, V ■) were plotted according to the two objectives. The star (☆) represents an experimentally verified standard refolding condition (Rudolph *et al.*, 1977) and the bisecting line denotes 100 % refolding yield. In addition, the optimization progress (last improvement in V) is highlighted for several GENs by black dashed lines.

Figure 5.20 illustrates the development of the two optimization objectives during the optimization. While the median of the native activity increased from 12 U mg^{-1} to 72 U mg^{-1}, only stagnating refolded activities could be observed. The overall experimental errors of the refolding experiments were comparable to the previously detailed proteins (data not shown).

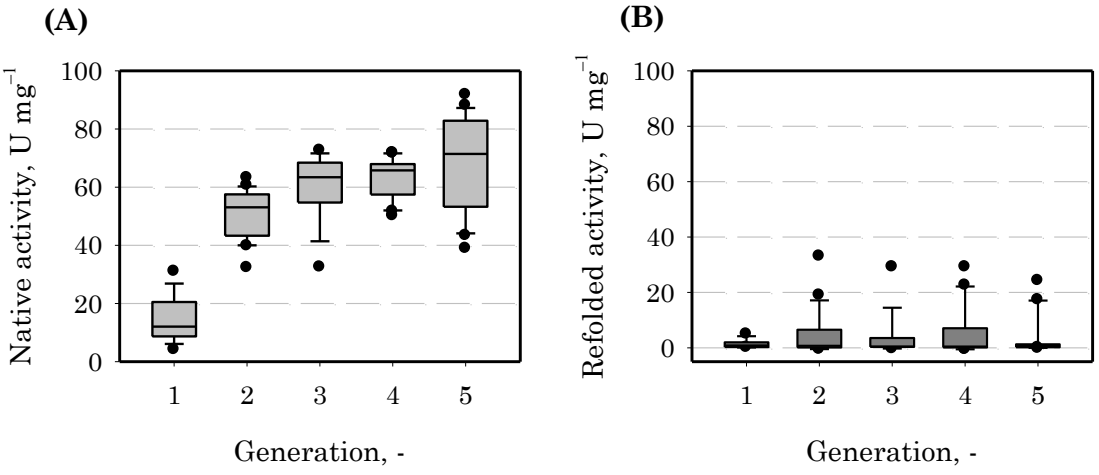


Figure 5.20: Development of the two objectives during the LDH optimization illustrated as box plots: (A) native activity, (B) refolded activity. The boxes contain the middle of 50 % of the data, whiskers denote the 10th and 90th percentiles. (●) outliers, (—) median.

5.1.6 Lipase from *Thermomyces lanuginosus* (LIP)

Like LYZ (section 5.1.4), the lipase from *Thermomyces lanuginosus* (LIP) is an extracellular enzyme and disulfide-bridged. The hydrolytic enzyme is mainly used in washing agents to remove oils and fats from fabrics. LIP refolding was evaluated with a nitrophenyl palmitate based activity assay (Liu *et al.*, 2006), as exemplary illustrated in Figure 5.21.

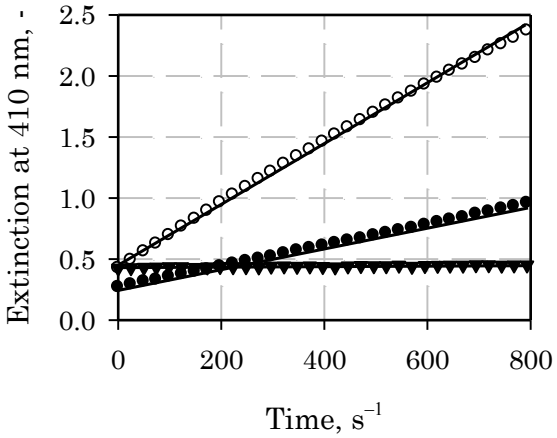


Figure 5.21: LIP activity. Exemplary activity assay of native (●, 0.1 M TRIS·HCl, pH 7.5), refolded (○, 0.1 M TRIS·HCl, 50 mM NaCl, 100 mM arginine and 2.5 mM GSSG, pH 8.75) and denatured (▲) protein, linear regression is indicated.

The extinction coefficient (ε) of the product 4-nitrophenol was dependent on the pH. Therefore, the influence of the refolding buffers (pH 6.0 to pH 9.5) was examined in a preliminary experiment (Table 5.2).

Table 5.2: Influence of the pH of the refolding buffer on the pH in the LIP activity assay (buffered with 0.1 or 2.5 M TRIS·HCl, pH 7.5) and the extinction coefficient (ε) of the product 4-nitrophenol (PB, sodium phosphate buffer).

	Addition of 1 M TRIS·HCl, pH 7.5		**Addition of 0.1 M PB, pH 6.0**		**Addition of 1 M TRIS·HCl, pH 9.5**	
	pH, -	ε, M^{-1} cm^{-1}	pH, -	ε, M^{-1} cm^{-1}	pH, -	ε, M^{-1} cm^{-1}
0.1 M TRIS·HCl	7.50	9627	6.77	4160	8.90	16 800
2.5 M TRIS·HCl	7.50	14 547	7.52	13 924	7.76	14 535

The addition of acid or alkaline refolding buffers resulted in large deviations for both pH and ε in the original assay, which was buffered with 0.1 M TRIS·HCl at pH 7.5. An increased buffer capacity (2.5 M) exhibited far better performance with a maximal deviation of 435 M^{-1} cm^{-1} for ε. Hence, refolding experiments were carried out with the modified assay and the mean value for ε in the 2.5 M buffer (14 350 M^{-1} cm^{-1}) was used for LIP activity calculations.

LIP refolding was optimized analog to LDH with native and refolded activities as objectives (Figure 5.22, A). Native activities of up to 680 U g^{-1} and refolded activities of up to 380 U g^{-1} could be obtained within six GENs. These activities remarkably exceeded the reference (Ahn *et al.*, 1997), which exhibited 76 U g^{-1} refolded activity.

In order to estimate the robustness of the stochastic optimization, LIP was subjected to a further, independent optimization approach using the same setup (random start). In the second approach, the best refolding condition from the first optimization was chosen as a new reference (Figure 5.22, B). Native LIP activities of up to 1400 U g^{-1} were measured in the second approach. Data analysis revealed sodium dodecyl sulfate (SDS) in all refolding conditions with activities higher than 700 U g^{-1}.

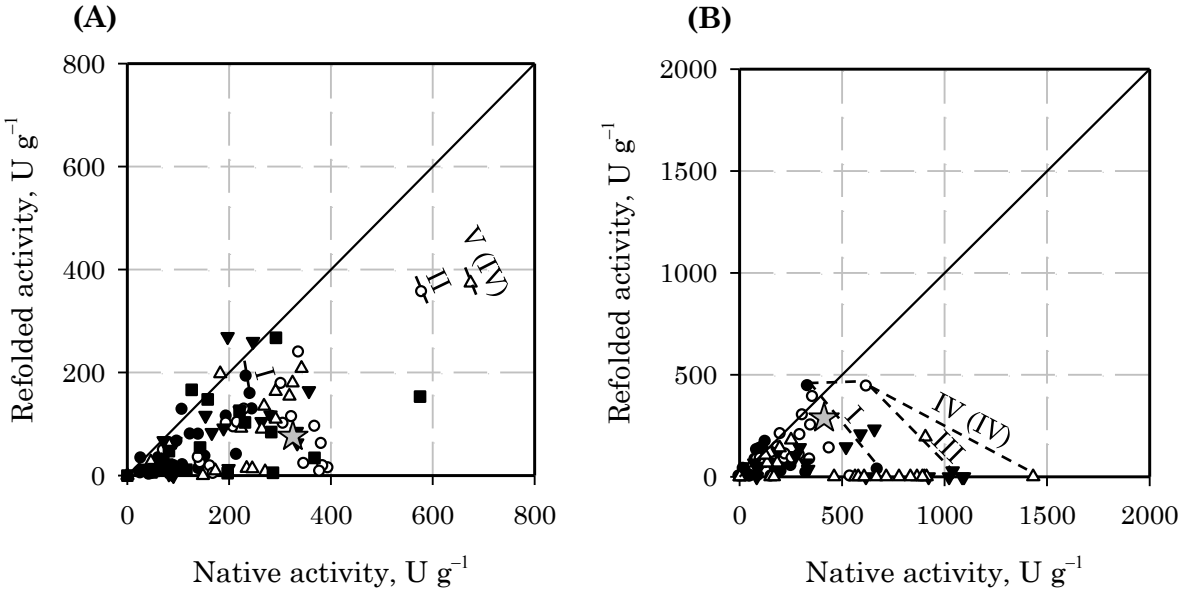


Figure 5.22: Overviews of the first (A) and second (B) independent optimization approaches for LIP. Experimental data of the individual GENs (I ●, II ○, III ▼, IV Δ, V ■) were plotted according to the two objectives. The star (☆) represents a standard refolding condition (A: Ahn *et al.* (1997), B: best result of the first optimization) and the bisecting line denotes 100 % refolding yield. In addition, the optimization progress (last improvement in IV / IV) is highlighted for several GENs by black dashed lines.

A negative side effect of the presence of SDS was a low reproducibility, since remains from the diluted denaturant guanidine hydrochloride (Gdn·HCl) caused precipitation in presence of SDS (results not shown) Therefore, the second optimization was terminated prematurely after four GENs and alternative protein denaturation with urea was tested and verified by circular dichroism spectroscopy (CD) (Figure 5.23).

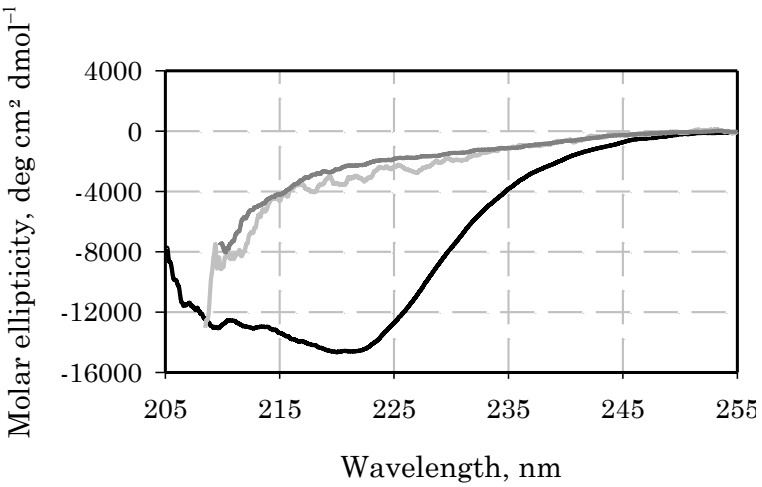


Figure 5.23: CD spectroscopy of native (black) and denatured LIP incubated in a denaturation buffer with 10 mM DTT and 6 M Gdn·HCl (dark grey) or 10 M urea (grey).

In the following third stochastic optimization (Figure 5.24), Gdn·HCl was replaced with urea for denaturation. Otherwise, there were no changes and the third approach was performed independently of the previous optimizations (random start). In the third optimization, up to 1400 U g^{-1} refolded activity were determined. This was 2.8-fold higher compared to the second optimization. In the latter, high native activities were measured, but maximum refolding was smaller than 500 U g^{-1} and thus comparable to the first optimization. Furthermore, native activities were improved in the third approach as well, up to 1750 U g^{-1} were measured in refolding conditions with SDS.

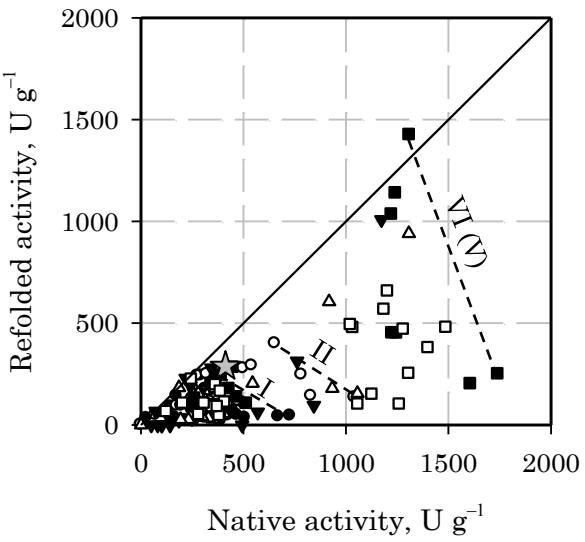


Figure 5.24: Overview of the third LIP optimization, LIP was denatured with urea. Experimental data of the individual GENs (I ●, II ○, III ▼, IV Δ, V ■, VI □) were plotted according to the two objectives. The star (☆) represents a standard refolding condition (best result of the first optimization) and the bisecting line denotes 100 % refolding yield. In addition, the optimization progress (last improvement in V) is highlighted for several GENs by black dashed lines.

5.1.7 Comparing refolding from soluble proteins and inclusion bodies

One of the major assumptions of the refolding experiments was the comparability between soluble, native protein and inclusion bodies (IBs). Despite the amount of literature on refolding, a search revealed no information about this specific topic. Therefore, an experimental verification for at least one protein was considered important. LIP, which was purchased in soluble form from Sigma-Aldrich was chosen as an example protein. LIP was expressed in *Escherichia coli* (*E. coli)* and subsequently refolding was evaluated.

Expression and purification

Protein- and DNA sequences were derived from the UniProt database (Jain *et al.*, 2009). After gene synthesis and transformation into *E. coli* BL21 (DE3), protein expression was validated (Figure 5.25). In the SDS polyacrylamide gel electrophoresis (SDS-PAGE) protein expression was evident in the cell pellet, indicating IB formation. Whereas, no or only very faint bands were visibly in the soluble fractions. This was the case for both the His-tagged and the non-tagged construct. In addition, protein sizes were comparable to calculations (29.3 kDa and 30.3 kDa).

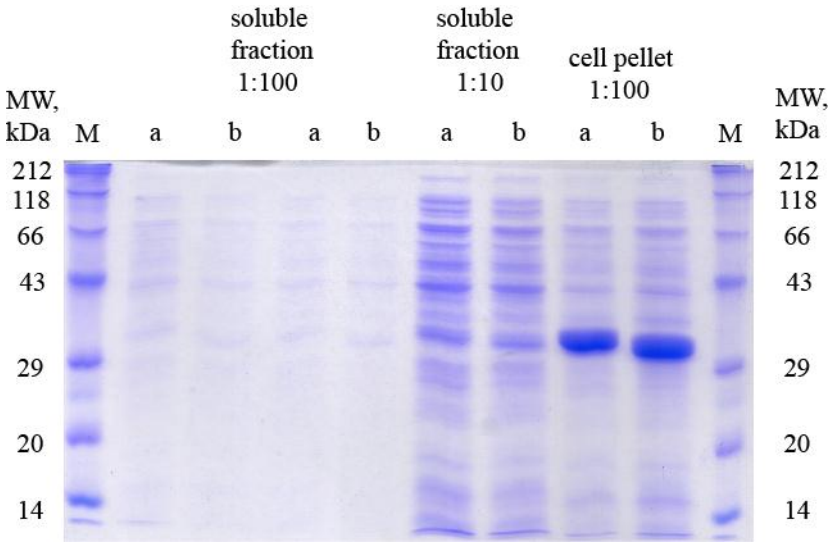


Figure 5.25: Soluble fractions and cell pellets of expressed LIP. (a) LIP with His-Tag, (b) non-tagged LIP, (M) marker with molecular weights (MW) between 14 kDa and 212 kDa (Carl Roth).

Purified protein is generally preferred for refolding, as aggregating side reactions can reduce yields (see 3.2). Hence, a purification step was carried out prior to refolding. IBs of His-tagged LIP were purified using immobilized metal ion affinity chromatography (IMAC). IBs of non-tagged LIP were subjected to several washing steps with detergents to remove possible membrane proteins decontaminants. The SDS-PAGE from the purification is depicted in the appendix (Figure 9.1). In case of the His-tagged protein, no other bands indicating contaminant proteins were visible. The sample contained even less impurities than the purchased control. In contrast, the non-tagged protein was slightly purified after washing with detergent containing buffers. According to densitometry analysis the non-tagged LIP was approximately 90 % pure, while the purchased protein was 95 % pure.

Refolding experiment

After purification, refolding was examined in five different conditions from the previous stochastic optimizations (section 5.1.6). As both constructs formed IBs and almost no soluble protein (Figure 5.25), a comparison of the native activities was not possible. The results of the refolding experiment are detailed in Figure 5.26. While the LIP with His-tag exhibited 49 % to 57 % of the control activity, refolded activities of the non-tagged protein and the control were in good agreement for all five experimental conditions.

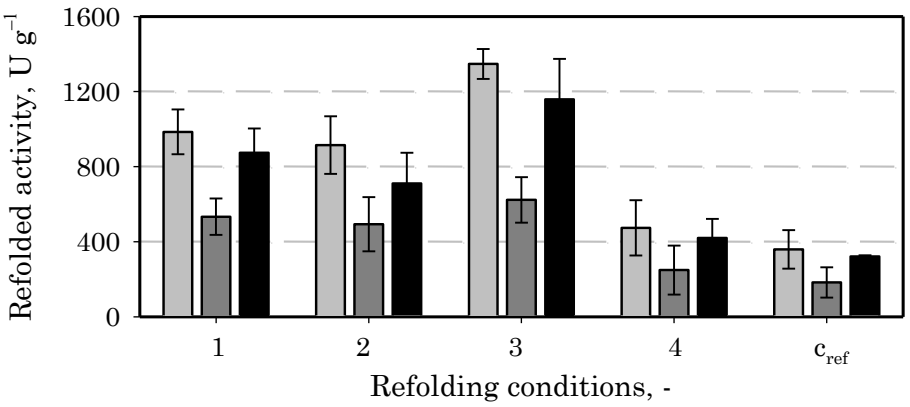


Figure 5.26: Refolding with different LIP forms. Comparison of soluble protein (grey) purchased from Sigma-Aldrich and protein from IBs: (dark grey) His-tagged protein, (black) non-tagged protein, (1 to 4, c_{ref}) refolding conditions and the reference from the previous optimizations.

5.1.8 Overall comparison of the proteins under study

This subchapter gives an overview of the performance of the stochastic optimization strategy regarding all six proteins. Hence, the reference conditions for each protein and the optima obtained in the optimization are compared.

Performance of the stochastic optimization strategy

For all six proteins the stochastic optimization strategy identified comparable or better refolding conditions than the reference gained from the literature. For GFP and GLR refolding yields and experimental costs were optimized, achieving refolding buffers with 100 % yield in conjunction with costs of 0.006 € $mL^{-1}$and 0.012 € mL^{-1}. GLR, GLK, LYZ, LDH and LIP were optimized regarding the native and refolded activities. Thereby, the optimization of the enzymatic activities was successful for all five proteins, obtaining a 1.3-fold to 30.6-fold increase compared to the activity of the reference (Table 5.3). In addition, 100 % refolding yield could be observed for all proteins except LDH.

Table 5.3: Overall performance of the stochastic optimization for protein refolding. Maximum measured native and refolded activities relative to the experimentally verified standard refolding conditions.

	GFP	GLR	GLK	LYZ	LDH	LIP
Improvement of native activity	n/a	130 %	130 %	40 %	40 %	5.3-fold
Improvement of refolded activity	n/a	250 %	30 %	40 %	40 %	30.6-fold
100 % yield achieved	yes	yes	yes	yes	no	yes

Composition of the best refolding conditions

Although both native and refolded activity were optimized in parallel, refolding was considered to be of primary importance for the analysis. The experimental conditions with the highest refolded activities and yields, which can be stated as the "optima" of the stochastic optimizations for each protein, are detailed in Table 5.4.

All listed refolding buffers contained 6 to 10 compounds and were therefore rather complex compared to the reference conditions, which contained 3 to 5 substances. The similarity of the two best conditions (optima A and B in Table 5.4), varied significantly. LYZ optima were similar and differed only in one variable: the pH (pH 7.5 and pH 7.0). For the other proteins, comparable refolded activities could be determined in two different buffers. For example GLK differed in the ionic strength, the buffer substances, the refolding additives and the redox agent, but both conditions exhibited 236 U mg^{-1} refolded activity in the assay. One overall trend was visible: GFP, GLR, GLK and LDH optima all featured a reductive component (DTT or TCEP). In contrast, LYZ and LIP optima contained a combination of oxidative (GSSG) and reductive (GSH) glutathione.

Modifications in the general configuration of the stochastic optimization

After the evaluation of the first four proteins, the acquired data were analyzed with regards to negative trends for all proteins. High concentrations of divalent metal ions (Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2+}) and TRIS·HCl and the presence of SDC (deoxycholic acid sodium salt) frequently resulted in refolding buffers that exceeded the maximal solubility. Hence, refolding buffers could not be prepared or precipitation occurred at lower temperatures (data not shown). In subsequent optimizations (LDH and LIP), variables

were adjusted accordingly. In the modified GA parameter space (Table 4.4), mineral ions were reduced to 0.1 mM, TRIS·HCl to 1 M and SDC was removed from the list of tested detergents. At the same time, glutamate was added as an additional refolding additive and other minor modifications were performed (see Table 4.4).

Table 5.4: The best identified refolding conditions of the stochastic optimization (A and B) and the experimentally evaluated reference refolding conditions (R) for GFP, GLR, GLK, LYZ, LDH and LIP. Listed are the composition, the individual activities of the native and refolded protein and the yield achieved in the respective refolding conditions.

Best refolding conditions (A and B) and reference (R) for each protein	**Native activity***	**Refolded activity***	**Relative yield, %**
GFP A 40 mM PB, pH 7.0, 100 mM NaCl, 10 % v/v glycerol, 50 mM arginine, 50 mM glutamine, 5 mM EDTA, 7.5 mM DTT	18100 ± 5110	19862 ± 1743	100 ± 37
GFP B 50 mM TRIS·HCl, pH 7.0, 250 mM NaCl, 15 % v/v glycerol, 100 mM arginine, 50 mM glutamine, 2.5 mM TCEP	25700 ± 7500	24885 ± 5717	97 ± 51
GFP R (Dashivets *et al.*, 2009) 40 mM PB, pH 7.5, 300 mM NaCl, 50 mM arginine, 50 mM glutamine, 5 mM DTT	28480 ± 3330	24900 ± 4152	87 ± 9
GLR A 100 mM MOPS, pH 8.5, 150 mM NaCl, 20 mM KCl, 500 mM arginine, 50 mM glutamine, 5 mM EDTA, 0.06 mM TWEEN 20, 2.5 mM DTT	101 ± 5	96 ± 8	95 ± 12
GLR B 50 mM MOPS, pH 8.5, 300 mM NaCl, 0.1 % w/v PEG 4000, 100 mM arginine, 100 mM glycine, 2 mM EDTA,	120 ± 8	112 ± 18	94 ± 20
GLR R (Nordhoff *et al.*, 1997) 20 mM PB, pH 6.9, 0.5 mM EDTA, 2 mM DTT	95 ± 20	60 ± 15	61 ± 18
GLK A 20 mM HEPES, pH 9.5, 350 mM NaCl, 0.05 % w/v PEG 4000, 5 mM EDTA, 5 mM DTT	213 ± 23	236 ± 8	100 ± 14
GLK B 20 mM TRIS·HCl, pH 9.5, 50 mM NaCl, 0.15 % w/v PEG 4000, 50 mM arginine, 50 mM glutamine, 5 mM EDTA, 7.5 mM DTT	266 ± 13	236 ± 8	89 ± 7
GLK R (assay buffer) 50 mM HEPES, pH 7.5, 150 mM KCl, 10 mM $MgCl_2$	137 ± 29	198 ± 24	100 ± 28

Table 5.4 (continued):

Best refolding conditions (A and B) and reference (R) for each protein	Native activity*	Refolded activity*	Relative yield, %
LYZ A 100 mM TRIS·HCl, pH 7.5, 0.05 % w/v PEG 4000, 50 mM arginine, 100 mM glutamine, 25 mM glycine, 1 mM GSSG	16.8 ± 1.3	12.4 ± 0.9	74 ± 11
LYZ B 100 mM TRIS·HCl, pH 7.0, 0.05 % w/v PEG 4000, 50 mM arginine, 100 mM glutamine, 25 mM glycine, 1 mM GSSG	16.3 ± 1.5	10.2 ± 1.2	62 ± n/a
LYZ R (Hevehan and De Bernardez Clark, 1997) 50 mM TRIS·HCl, pH 8.0, 1 mM EDTA, 0.5 mM GSH, 5 mM GSSG	11.2 ± 1.8	9.2 ± 1.8	82 ± 16
LDH A 20 mM TRIS HCl, pH 7.75, 100 mM NaCl, 50 mM arginine, 100 mM glutamine, 0.12 mM BRIJ, 3.75 mM TCEP	60.6 ± n/a	33.0 ± 1.3	54 ± 2
LDH B 40 mM MOPS, pH 6.75, 25 mM NaCl, 33 mM KCl, 0.05 % w/v PEG 4000, 2 mM EDTA, 2.5 mM TCEP	71.5 ± n/a	29.1 ± 1.0	41 ± 1
LDH R (Rudolph *et al.*, 1977) 200 mM PB, pH 7.6, 1 mM EDTA, 0.1 mM DTT	66.2 ± 5.5	23.1 ± 5.5	36 ± 11
LIP A 500 mM TRIS·HCl, pH 8.5, 175 mM NaCl, 50 mM KCl, 0.05 % w/v PEG 4000, 250 mM arginine, 200 mM glutamate, 12 mM SDS, 0.5 mM GSH, 5 mM GSSG	1306 ± n/a	1430 ± 175	100 ± 12
LIP B 750 mM TRIS·HCl, pH 7.5, 50 mM KCl, 25 mM arginine, 50 mM glutamine, 12 mM SDS, 5 mM GSH, 5 mM GSSG	1451 ± 286	1335 ± 172	92 ± 32
LIP R (Ahn *et al.*, 1997) 50 mM TRIS HCl, 10 mM $CaCl_2$, 5 mM DTT	325 ± 76	45 ± 10	14 ± 5

***** GFP fluorescence intensity at 408 nm, GLR, GLK, LDH specific activity in U mg^{-1} and LIP in U g^{-1}, LYZ activity according to the EnzChek® assay in s^{-1}; **PB**, sodium phosphate buffer; **TRIS**, tris(hydroxymethyl)aminomethane; **DTT**, dithiothreitol; **TCEP**, tris-carboxyethyl-phosphine; **MOPS**, morpholino-propanesulfonic acid; **EDTA**, ethylenediaminetetraacetic acid; **TWEEN 20**, polyethylene glycol sorbitan-monolaurate; **PEG**, polyethylene glycol; **GSH**, reduced glutathione; **GSSG**, oxidized glutathione; **HEPES**, hydroxylethyl-piperazine-ethanesulfonic acid; **BRIJ 35**, polyethylene glycol dodecyl ether

5.1.9 Discussion

Six functionally and structurally different model proteins were successfully optimized in terms of the refolding buffer, thereby proving the applicability of the proposed stochastic optimization strategy. Both yields and the underlying activities could be significantly improved compared to the experimentally validated literature references.

Optimization results of the individual proteins

GFP and GLR were optimized focusing on costs and refolding yields. The optimal refolding conditions for GFP identified in this work were quite similar to the reference condition described by Dashivets *et al.* (2009) and exhibited roughly identical yields considering the high measurement errors. Refolding yields and activities of GLR could be significantly increased compared to the reference. The experimentally determined refolding yield of the GLR reference (61 ± 18 %) corresponded well to Nordhoff *et al.* (1997) who reported up to 70 % yield depending on the protein concentration. While maximum refolding yields could be achieved for both proteins, the cost optimization was not saturated after six GENs. This was indicated by the lower costs of the reference and the steady progress of the cost criteria until the end of the optimization. Costs are a critical process parameter as 60 % to 75 % of the operating costs of industrial-scale inclusion body processes relate to the refolding step. For batch dilution, roughly 85 % of these costs comprise the raw materials, that is mainly the refolding buffer (Lee *et al.*, 2006). Typical costs for refolding buffers are not detailed, but redox agents are the most costly compounds (Freydell *et al.*, 2011).

GLK and LDH were probably suboptimal targets for a refolding screen. GLK exhibited "too" easy refolding, as 40 % of the examined conditions featured 100 % refolding yield. Furthermore, even dilution of the denatured protein in the buffer of the functional assay resulted in maximum refolding yields. Nevertheless, higher specific activities could be obtained during the optimization of GLK. The optimization of LDH failed to achieve 100 % refolding yield. Overall, the activities of the refolded LDH were low and only the native protein exhibited a trend towards higher activities during the optimization. The observed instability in solution was probably related. Rudolph *et al.* (1977) reported no loss of activity, but the LDH under study was a different subtype.

Reductive conditions were predominant for all four above-detailed proteins. Probably because DTT and TCEP prevented the oxidation of free SH-groups by air oxygen and thus positively affected the stability and activity. In contrast, LYZ and LIP required oxidative conditions for successful refolding and exhibited no or reduced enzymatic

activity in the presence of DTT or TCEP. This can be easily explained, LYZ and LIP contain disulfide bonds (Table 5.1).

LYZ constituted a difficult target protein, as two factors (oxidative conditions and ionic strength) drastically influenced the activity. By putting constraints on both factors and performing modified optimizations with a reduced search space, the proportion of positive conditions could be increased to enable progression of the GA. LYZ is probably the best analyzed protein and well-characterized. Hence, both the necessity of glutathione (GSH / GSSG) and the impact of ionic strength were known (Davies *et al.*, 1969; Hevehan and De Bernardez Clark, 1997). However, data analysis was carried out independently without further information. Both trends were clearly observed in the datasets (Figure 5.14 and Figure 5.15). Hence, it should be possible to apply the same approach to new, unknown proteins. Unfortunately, a different batch of LYZ was used in the third experiment. Therefore, the activities were different compared to the previous optimizations.

LIP activity was strongly influenced by SDS. Furthermore, the protein required (like LYZ) oxidative conditions for refolding. In this regard, the reference with 5 mM DTT was rather unsuitable as it was reductive. This reference was chosen because refolding conditions for the Lipase from *Thermomyces lanuginosus* (LIP) were not available. Ahn *et al.*, (1997) studied the lipase from another organism (*Pseudomonas fluorescence).* Hence a comparison to LIP is not valid. For subsequent optimizations the best refolding condition from the first optimization was selected as a new reference. In these optimizations high LIP activities could be obtained in refolding conditions with SDS. However, the reproducibility was low in presence of SDS and Gdn·HCl impeding the comparison of the three stochastic optimizations.

Behavior of the GA during the optimization

All optimal refolding conditions identified with the stochastic optimization were complex mixtures of 6 to 10 substances (Table 5.4). This bias towards complicated refolding buffers was probably a side effect of the encoding (see 4.5.1). Although a “soft” limit for complexity was introduced by classification in functional groups, there was no actual function to reduce complexity. Therefore, non-significant substances may remain during the optimization as the evolutionary pressure does not apply. This point is backed by the first two optimizations, which included experimental costs as objective. In these optimizations the general complexity was lower (compare GFP Table 5.4), as the costs acted as an indirect pressure and thereby penalized non-significant variables. Hence,

either the encoding strategy has to be adapted by defining a maximum for the number of compounds or an additional objective analog to the costs has to be introduced. The introduction of a third objective should be possible without changes to the experimental setup (22 experiments in each GEN).

During the optimization, the GA "stuttered": Although the mean of the objective functions increased during most optimizations, a clear pareto front with many individuals did not occur. As the GA parameters for population size, mutation and crossover were default and already used in other experimental optimizations (Gobin *et al.*, 2007; Havel *et al.*, 2006), this probably indicated a non-steady search or decision space. The high experimental error of up to 50 % for the refolding yield of GFP probably was one reason. Comparable stochastic optimizations of fermentation media performed well with standard deviations of up to 20 % (Weuster-Botz, 2000).

Suitability of the functional assays

Detection of refolded proteins by functional assays constitutes a standard approach for refolding screens, which offers reliable information about folding and can be easily parallelized in 96-well plate scale (Armstrong *et al.*, 1999; Middelberg, 2002; Willis *et al.*, 2005). Using this method, six functionally and structurally different proteins (compare Table 4.1) were successfully optimized within the scope of this thesis. The proteins under study differed notably in their monomer mass (14 kDa to 53 kDa), their quaternary structure (monomer to tetramer) and the pI (pH 5.0 to pH 9.3). Next to GFP, five different enzymes were examined: two oxidoreductases (LDH, GLR), two hydrolases (LYZ, LIP) and one transferase (GLK). This variability is sufficient for a proof of concept of the stochastic optimization strategy. However, in the light of the natural diversity of proteins a further generalization (prediction for unknown proteins) is problematic.

Refolding screens from the literature are limited to the analysis of the refolded protein. In this project, the activity of the native protein in the refolding buffer was measured as well. Thereby, effects of refolding additives on the activity itself could be excluded and refolding yields were quantified individually for each refolding condition. The major advantage was the ability to differentiate between effects on refolding and the activity. This made it possible to gain insights in the enzymatic activity, which is especially interesting for industrial biotransformations. Within this context, the switch from optimizing yields to optimizing the underlying activities itself was important (Figure 5.7). For enzymatic applications the overall activity of the protein is decisive, not the refolding yields in the production process.

Measurements of the native activity are only possible if soluble protein is available. Further, the application in the refolding optimization is based on the assumption, that refolding from the denatured soluble protein is analog to refolding from the denatured IB. The comparability was proven for LIP (Figure 5.26), verifying the application of soluble model proteins in refolding screens. Tagged LIP was far less active, probably because the C-terminal His-tag was too close to the active center. Functional tags may influence protein activity and refolding, both positive and negative effects were observed for other proteins (Ishibashi *et al.*, 2011).

5.2 Analysis of design of experiments (DOE) strategies

Design of experiments (DOE) strategies aim to efficiently and precisely identify optimal solutions inside the problem specific search space (see 3.3). In this thesis, optimal protein refolding conditions were obtained in a small number of experiments using a stochastic optimization strategy based on a genetic algorithm (GA). In total, six different proteins were optimized, which are detailed in the previous chapter (5.1). In this section, the optimization performance of the proposed optimization strategy is first further characterized (5.2.1) and then compared to a classic statistical DOE (5.2.2). Experiments were performed with the lipase from *Thermomyces lanuginosus* (LIP).

5.2.1 Robustness of the stochastic optimization

Experimental design strategies based on GAs are heuristic and stochastic. Hence, it is not guaranteed to reach the global optimum. Depending on the random start and the subsequent experiments, optimization approaches may perform differently. In order to analyze the robustness and the stochastic nature of the proposed optimization strategy, LIP refolding was optimized consecutively several times.

Sodium dodecyl sulfate (SDS) was observed to facilitate high native and refolded activities of up to 1750 U g^{-1} but also caused precipitation in the presence of guanidine hydrochloride (Gdn·HCl) (see section 5.1.6). However, optimal refolding conditions with SDS were not identified in all optimizations. While the first optimization (OPT_I) failed to identify SDS and exhibited only moderate (up to 700 U g^{-1}) enzymatic activities, the second, independent optimization (OPT_{II}) contained conditions with highly active (up to 1400 U g^{-1}) native protein. Finally, very high refolded activities (up to 1400 U g^{-1}) were measured in a third optimization (OPT_{III}). Hence, the performance was different, even though all optimizations were performed independently (random start) using the same parameters.

Optimization of LIP without SDS

In order to examine the above-detailed differing performance, first the necessity of SDS was examined in detail. SDS was removed from the list of included detergents and two additional optimizations were performed with the GA. Both approaches were independent with a random start (Figure 5.27). The maximum identified refolded activities were at the level of the reference (about 300 U g^{-1}), which was the best refolding condition from OPT_I. However, the native activity could be slightly improved compared to OPT_I: 800 U g^{-1} were measured in a complex alkaline refolding buffer

without SDS. In all experiments (264 unique refolding conditions), LIP activity was significantly smaller compared to the previously determined conditions with SDS.

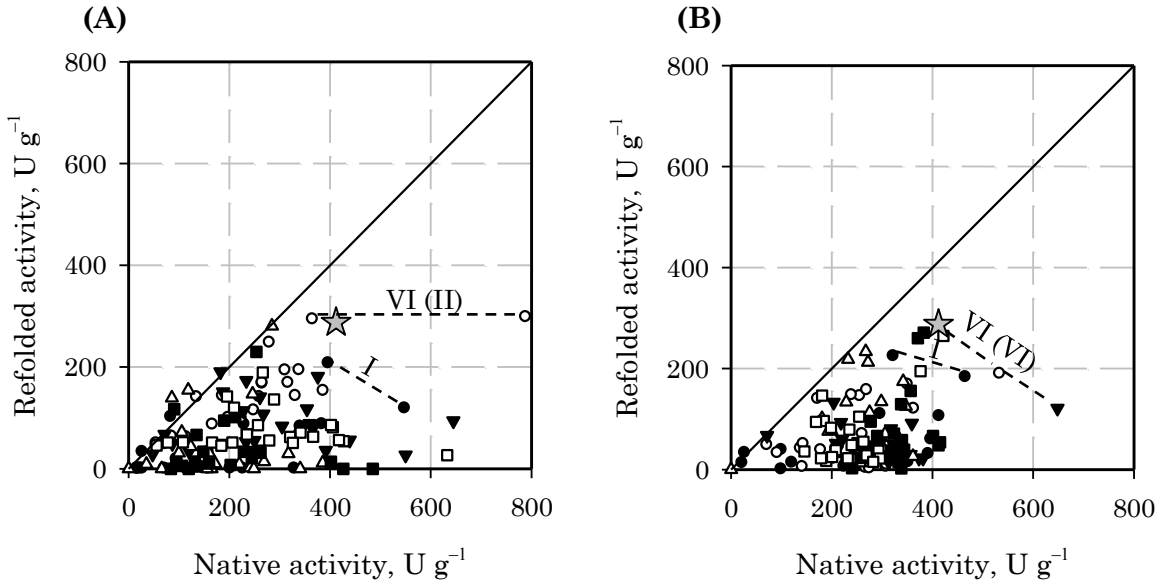


Figure 5.27: Overviews of the fourth (A) and fifth (B) LIP optimization approaches, SDS was excluded from the search. Experimental data of the individual GENs (I ●, II ○, III ▼, IV Δ, V ■, VI □) were plotted according to the two objectives. The star (☆) represents a standard refolding condition (best result of the first optimization) and the bisecting line denotes 100 % refolding yield. In addition, the optimization progress (last improvement in IV / IV) is highlighted for several GENs by black dashed lines.

Continuation of the first and second optimization

While the first two optimizations failed to identify experimental conditions which exhibited high refolded activities, both were terminated rather quickly. Only four to five generations (GENs) were evaluated (compare Figure 5.22). Therefore, OPT_I and OPT_{II} were continued by performing additional GENs, in which Gdn·HCl was replaced with urea for protein denaturation (Figure 5.28). OPT_I was previously discontinued after five GENs. In the next few GENs there was little progress, but in GEN_{VIII} 1300 U mg^{-1} native activity could be measured in a buffer containing SDS. Finally, GEN_X yielded a refolding buffer with 100 % yield and about 1000 U g^{-1} activity. In addition, native activities could be further increased to 1450 U g^{-1} (Figure 5.28, A). OPT_{II} was previously aborted after four GENs. This optimization already contained a lot of experiments with SDS and high native activities. In the new experiments the next GEN (GEN_V) exhibited refolded activities of up to 1300 U g^{-1}. These could be slightly improved in the following GENs. The optimization was terminated after GEN_{VIII} (Figure 5.28, B).

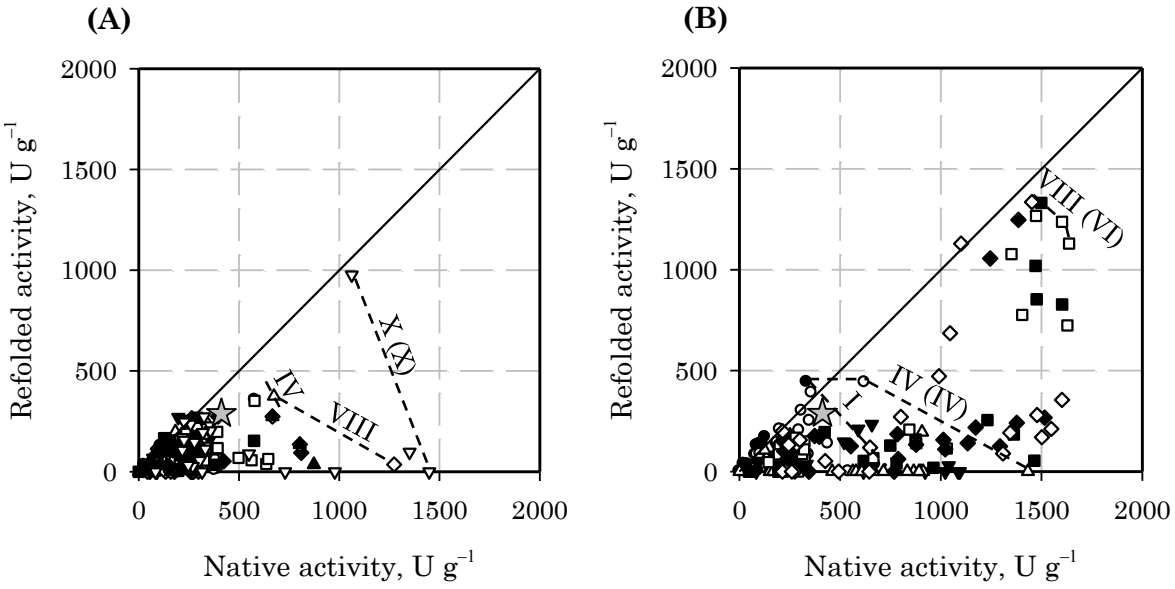


Figure 5.28: Overviews of the continued first (A) and second (B) optimization approaches. Experimental data of the individual GENs (I ●, II ○, III ▼, IV △, V ■, VI □, VII ♦, VIII ◊, IX ▲, X ▽) were plotted according to the two objectives. The star (☆) represents a standard refolding condition (best result of the first optimization, Figure 5.22) and the bisecting line denotes 100 % refolding yield. In addition, the optimization progress (last improvement in X / VI) is highlighted for several GENs by black dashed lines. LIP was denatured with urea for the additional experiments.

Figure 5.29 compares the progress of the three independent stochastic optimizations for LIP, which incorporated SDS (continued OPT_{I+II} and OPT_{III}). While OPT_{II} and OPT_{III} identified comparable maximum activities, the best refolded activities in OPT_{I} were slightly lower. Nevertheless, all optimizations revealed both conditions with 1450 U g^{-1} or more native activity and refolded activities of 1000 U g^{-1} or more, which all contained SDS. The occurrence of high activities in the three optimizations differed notably. OPT_{I} exhibited only a few activities higher than 750 U g^{-1}. In contrast, OPT_{II} and OPT_{III} featured more conditions with high activities. This was closely related to the presence of SDS during the optimization (Table 5.5). OPT_{I} contained only 18 refolding conditions with SDS, equivalent to 8 % of the total experiments. In the other two optimizations, up to 43 % of the refolding conditions featured SDS. All optimizations enriched SDS during the optimization, indicated by the high percentage of SDS in the last GEN.

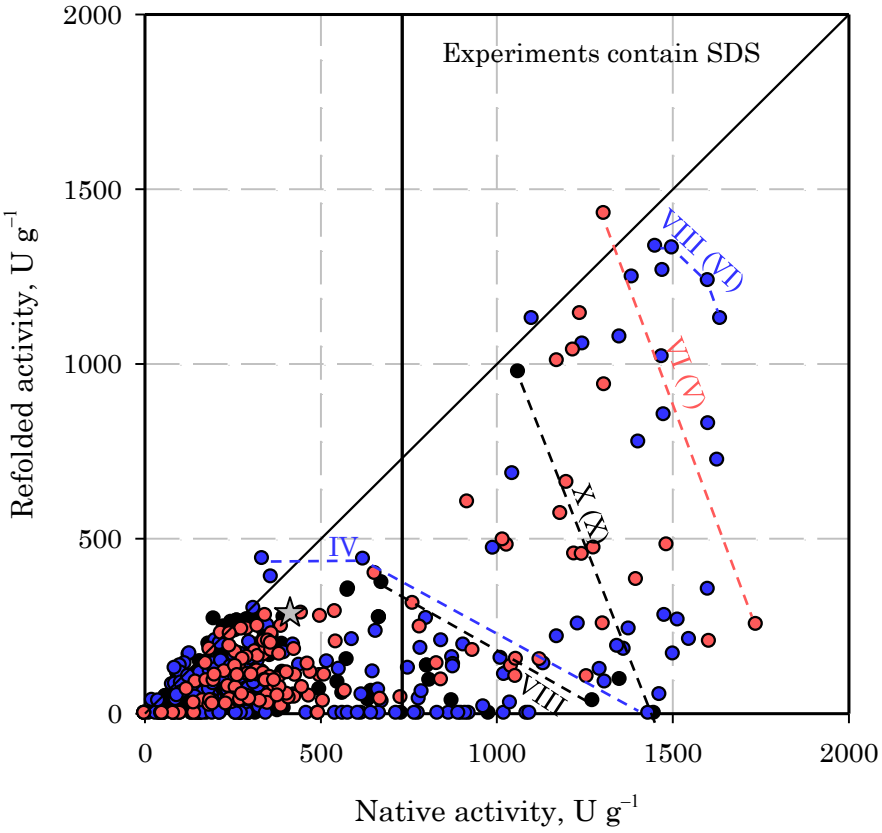


Figure 5.29: Overview of the first (●), second (●) and third (●) independent stochastic optimizations of LIP refolding. Experimental data were plotted according to the two objectives. The star (☆) represents a standard refolding condition and the bisecting line denotes 100 % refolding yield. In addition, the optimization progress (last improvement in X, V, VI) is highlighted for several GENs by dashed lines.

Table 5.5: Occurrence of SDS in LIP optimization one, two and three.

Optimization	GENs and total experiments	Conditions with SDS	Conditions with SDS in the last GEN
One	10 / 220	18 (8 %)	7 (32 %)
Two	8 / 176	76 (43 %)	12 (55 %)
Three	6 / 132	38 (29 %)	15 (68 %)

Although LIP refolding was subjected to an extensive analysis with 968 refolding experiments, a claim for the identification of a global optimum for LIP refolding was not justified. However, it was possible to compare the (local) optima identified by each optimization. Table 5.6 lists the refolding conditions with the highest refolded activities (optima). For all three optima, the specific activities were approximately 1000 U g^{-1} or higher with a refolding yield between 92 % and 100 %. Several compounds were ubiquitous: an alkaline buffer, SDS and a combination of reduced (GSH) and oxidized glutathione (GSSG). Furthermore, the composition of all refolding buffers was rather complex with 7 to 9 substances. A more detailed view revealed additional similarities between the optima of OPT_{II} and OPT_{III} (Figure 5.30). Both contained high concentrations of TRIS·HCl (500 mM to 750 mM), KCl, arginine and the maximum concentrations of SDS (12 mM) and GSSG (5 mM). On the other hand, the optimum of OPT_{I} exhibited a slightly lower activity. In this optima another buffering agent was applied and the concentrations of SDS (3 mM) and GSSG (0.5 mM) were much smaller.

Table 5.6: Highest refolded activities in optimization one, two and three. Listed are the composition, the individual activities of the native and refolded protein (* U g^{-1}) and the yield achieved in the respective refolding conditions.

Best LIP refolding condition (highest refolded activity) in each optimization	**Native activity***	**Refolded activity***	**Relative yield, %**
Optimization one (OPT_{II}) 100 mM MOPS, pH 9.25, 350 mM NaCl, 25 mM glutamate, 7.5 mM EDTA, 3 mM SDS, 3.75 mM GSH, 0.5 mM GSSG	1062 ± 296	977 ± 33	92 ± 29
Optimization two (OPT_{II}) 750 mM TRIS·HCl, pH 7.5, 50 mM KCl, 25 mM arginine, 50 mM glutamine, 12 mM SDS, 5 mM GSH, 5 mM GSSG	1451 ± 286	1335 ± 172	92 ± 32
Optimization three (OPT_{II}) 500 mM TRIS·HCl, pH 8.5, 175 mM NaCl, 50 mM KCl, 0.05 % w/v PEG 4000, 250 mM arginine, 200 mM glutamate, 12 mM SDS, 0.5 mM GSH, 5 mM GSSG	1306 ± Na	1430 ± 175	100 ± 12

MOPS, morpholino-propanesulfonic acid; **TRIS**, tris(hydroxymethyl)aminomethane; **EDTA**, ethylenediaminetetraacetic acid; **GSH**, reduced glutathione; **GSSG**, oxidized glutathione **PEG**, polyethylene glycol; **TRIS**, tris(hydroxymethyl)aminomethane.

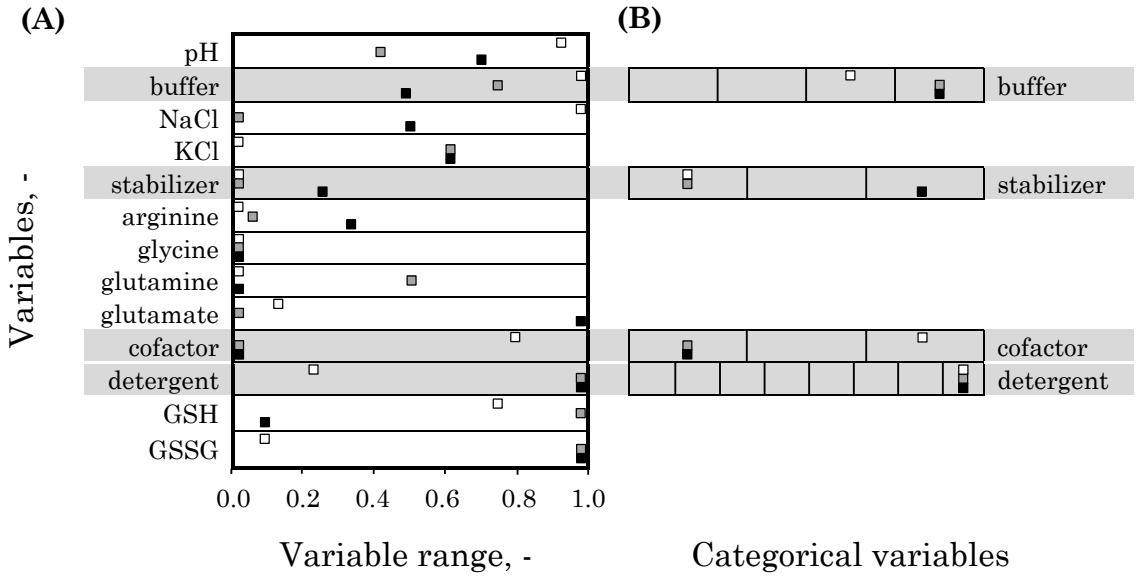


Figure 5.30: Composition of the best refolding conditions in each optimization. (▫) OPT_I, (▪) OPT_{II} and (■) OPT_{III}. (A) Variables were normalized using the maximal values of the search space (Table 4.4). (B) Complex variables with more than one substance were grouped in the respective categories sorted from left to right: buffer type (PB, HEPES, MOPS, TRIS·HCl), stabilizer type (none, Glycerol, PEG 4000), cofactor type (none, Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2+}, EDTA), detergent type (none, CHAPS, ZWITTERGENT 3-12, NDSB 201, TWEEN 20, TRITON-X 100, BRIJ 35, SDS).

5.2.2 Comparison to a standard, two-step statistical design of experiments (DOE)

In contrast to the previously discussed stochastic optimization, standard DOE strategies are statistic and centered on a simplified process model. This model is described by more (including quadratic and interactions terms) or less (only linear terms) complex equations. In this subsection the results of the stochastic optimization of LIP (see 5.1 and 5.2.1) are compared to a standard two-step DOE, which includes a D-optimal screening experiment and the subsequent optimization by response surface methodology (RSM).

Screening

First, a D-optimal screening design was generated. This DOE incorporated the variables of the stochastic optimization (compare Table 4.4), but as a screening experiment the resolution was drastically lower. Only two levels were examined for each variable, for example pH 6.00 and pH 9.75 or no NaCl and 350 mM NaCl. Apart from the pH, the addition of 25 refolding substances was analyzed. Concentrations represented the upper limit of the stochastic optimization (Table 5.7).

Table 5.7: Screened variables for LIP refolding. The D-optimal screening contained 27 variables (V_i), that were analyzed in two levels.

Nr.	Variable and abbreviation	Experimental values
V_1	pH	pH 6.00, pH 9.75
V_2	sodium phosphate buffer, PB	0 mM, 100 mM
V_3	hydroxylethyl-piperazine-ethanesulfonic acid, HEPES	0 mM, 100 mM
V_4	morpholino-propanesulfonic acid, MOPS	0 mM, 100 mM
V_5	tris-carboxyethyl-phosphine, TRIS·HCl	0 mM, 1000 mM
V_6	NaCl	0 mM, 350 mM
V_7	KCl	0 mM, 80 mM
V_8	glycerol	0 % v/v , 15 % v/v
V_9	polyethylene glycol 4000, PEG 4000	0 % w/v, 0.25 % w/v

Table 5.7 (continued):

Nr.	Variable and abbreviation	Experimental values
V_{10}	arginine	0 mM, 750 mM
V_{11}	glycine	0 mM, 350 mM
V_{12}	glutamine	0 mM, 350 mM
V_{13}	glutamate	0 mM, 350 mM
V_{14}	Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2+}, mineral ions (sulfates)	0 µM, 100 µM
V_{15}	ethylenediaminetetraacetic acid, EDTA	0 mM, 10 mM
V_{16}	cholamidopropyl-dimethylammonium-propanesulfonate, CHAPS	0 mM, 11 mM
V_{17}	non-detergent sulfobetaine 201, NDSB 201	0 mM, 1500 mM
V_{18}	dodecyldimethyl-ammonio-propanesulfonate, ZWITTERGENT 3-12	0 mM, 4 mM
V_{19}	polyethylene glycol sorbitan-monolaurate, TWEEN 20	0 µM, 80 µM
V_{20}	polyethylene glycol tert-octylphenyl ether, TRITON-X 100	0 µM, 800 µM
V_{21}	sodium dodecyl sulfate, SDS	0 mM, 12 mM
V_{22}	polyethylene glycol dodecyl ether, BRIJ 35	0 µM, 120 µM
V_{23}	dithiothreitol, DTT	0 mM, 10 mM
V_{24}	tris-carboxyethyl-phosphine, TCEP	0 mM, 10 mM
V_{25}	reduced L-glutathione, GSH	0 mM, 5 mM
V_{26}	oxidized L-glutathione, GSSG	0 mM, 5 mM
V_{27}	combination of GSH and GSSG	0 mM, 5 mM each

In order to ensure the comparability of the D-optimal design and the stochastic optimization, knowledge was incorporated in form of variable groups (factors). Most factors were simple numerical two-level factors, like the previously mentioned pH or NaCl. However, three factors grouped substances of the same class in categorical factors with more levels, as it made little sense to include more than one buffer substance or detergent into a refolding buffer. Furthermore, this process translated the class constraints of the GA (see section 4.5.1) into the statistical DOE and saved experimental

effort as fewer experiments were required. Details of the design are listed in the appendix (Table 9.5).

LIP refolding was carried out by diluting the denatured protein (urea) in the respective refolding buffer using the same experimental procedure as in the stochastic optimization (see 4.1). Refolded activities and the activity of the native protein were determined three-fold for each experimental condition. Refolding conditions and the measured activities are summarized in the appendix (Table 9.6). After performing the experiments, multi-linear regression with the design matrices (Table 9.7) gave a regression model with one constant and 28 linear coefficients. Subsequently, the model was refined by removing non-significant coefficients (Table 5.8).

Regression was carried out twice: first for the native and then for the refolded activities. Both could be modeled with high coefficients of determination (R^2) of 0.9995 and 0.9487. Even after refinement, the models were quite complex, since 18 and 22 variables had a significant effect on LIP activities. However, roughly 50 % had a negative effect and were consequently not of interest for a subsequent optimization. KCl, glycerol and mineral ions exhibited a moderate negative influence. Additionally, several detergents and the reductive redox agents DTT and TCEP severely decreased the activity.

The native activity of LIP was influenced positively by an alkaline pH and the addition of TRIS·HCl, arginine, glutamate, EDTA, GSH and three detergents. Many of these variables had a positive impact on refolding as well (alkaline pH, TRIS·HCl, arginine, EDTA and GSH). However, the refolded activity was also strongly influenced by redox agents (GSSG and the combination of GSH and GSSG). With regard to the detergents, SDS had a positive effect in both cases, but was outperformed by other detergents (TRITON-X 100 and TWEEN 20).

After the evaluation, the most important variables were selected for the subsequent RSM optimization. Efficient refolding conditions were the overall aim. Hence, the refolded activity (highest regression coefficient, b_i) was selected as decision criteria (Table 5.8). Consequently, the pH, TRIS·HCl, EDTA, TWEEN 20 and GSSG were chosen as variables for the next optimization. SDS exhibited a high coefficient but TWEEN 20, another detergent, performed even better.

Table 5.8: Coefficients (b_i) of the first order regression models for the native LIP activity (b^{nat}) and the refolded activity (b^{ref}) in the D-optimal screening. (*) variables selected for the RSM.

b	Variable	b^{nat}	b^{ref}
b_0	(constant)	260.2	78.07
b_1	**pH ***	**43.2**	**46.83**
b_2	PB	18.3	-
b_5	**TRIS·HCL ***	**20.4**	**59.36**
b_7	KCl	−75.4	−27.15
b_8	glycerol	−22.9	−24.49
b_{10}	arginine	25.3	27.37
b_{11}	glycine	−15.2	-
b_{12}	glutamine	−17.7	18.16
b_{13}	glutamate	67.3	-
b_{14}	Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2}	−51.8	−32.97
b_{15}	**EDTA ***	**35.6**	**35.89**
b_{16}	CHAPS	−92.1	−80.70
b_{17}	NDSB 201	−178.7	−59.34
b_{18}	ZWITTERGENT 3-12	50.5	-
b_{19}	**TWEEN 20 ***	**-**	**90.35**
b_{20}	TRITON-X 100	94.3	−39.87
b_{21}	SDS	75.8	69.24
b_{22}	Brij	−94.8	−59.55
b_{23}	DTT	−135.8	-
b_{24}	TCEP	−143.8	-
b_{25}	GSH	36.4	68.27
b_{26}	**GSSG ***	**-**	**149.96**
b_{27}	GSH/GSSG	13.3	83.02

RSM optimization

Based on the results of the previous screening, LIP refolding was optimized using response surface methodology (RSM). Five variables were examined: pH, TRIS·HCl, EDTA, TWEEN 20 and GSSG (Table 5.9). In contrast to the screening, the optimization was based on a second order polynomial, so interaction and quadratic terms were included (Equation 8). Hence, variables had to be analyzed in more detail (five levels, coded −2, −1, 0, 1, 2) and more experiments were necessary for each variable. A circumscribed central composite design with the mentioned five levels and 27 experiments was used (see appendix Table 9.8). Refolding was evaluated experimentally in each condition using the standard setup. Mean values of the activity measurements and the standard deviation are listed in the appendix (Table 9.8).

Table 5.9: RSM model for LIP refolding. Encoded levels (−2,−1, 0, 1, 2) variables (V_1 to V_5) and the respective experimental values.

Variable	V_1	V_2	V_3	V_4	V_5
Coding	**pH, -**	**TRIS·HCl, M**	**EDTA, mM**	**TWEEN 20, mM**	**GSSG, mM**
−2.00	8.00	0.00	0.00	0.00	0.00
−1.00	8.50	0.25	5.00	0.04	2.50
0.00	9.00	0.50	10.00	0.08	5.00
1.00	9.50	0.75	15.00	0.12	7.50
2.00	10.00	1.00	20.00	0.16	10.00

Analogous to the screening, both native and refolded activities were fitted to the model of the DOE (Equation 8) by multi-linear regression. Subsequently, the models were refined by removing non-significant coefficients. While the RSM model of the refolded activity featured an R^2 of 0.9796, the agreement of experimental and estimated native activities was lower (R^2 0.8894). Table 5.10 lists the significant constant, linear, interaction and quadratic terms for both models. Overall, the RSM confirmed the results of the screening experiment, as GSSG was not important for the native activity and TWEEN 20 had a very limited effect. In contrast, all five variables affected the refolded activity significantly.

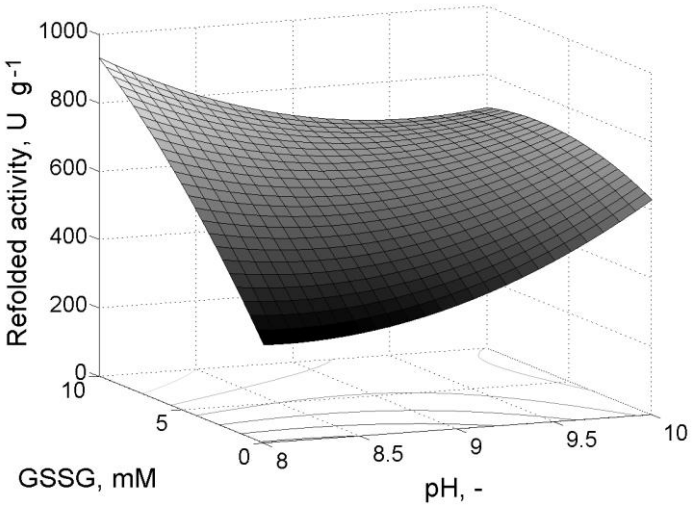
Table 5.10: Coefficients (b) of the second order polynomial regression models for the native LIP activity (b^{nat}) and the refolded activity (b^{ref}) in the RSM optimization. Constant (b_0), linear (b_i), interaction ($b_{i\cdot j}$) and quadratic ($b_{i\cdot i}$).

b	Variable	b^{nat}	b^{ref}
b_0	(constant)	364.3	359.3
b_1	pH	−31.0	14.0
b_2	TRIS·HCl	−32.7	7.5
b_3	EDTA	18.4	26.0
b_4	TWEEN 20	10.0	27.1
b_5	GSSG	-	45.1
$b_{1\cdot2}$	pH · TRIS·HCl	21.4	-
$b_{1\cdot4}$	pH · TWEEN 20	20.9	-
$b_{1\cdot5}$	pH · GSSG	-	−35.6
$b_{2\cdot5}$	TRIS·HCl · GSSG	-	22.3
$b_{4\cdot5}$	TWEEN 20 · GSSG	12.5	-
$b_{1\cdot1}$	$(pH)^2$	44.1	23.9
$b_{2\cdot2}$	$(TRIS\ HCl)^2$	26.1	12.2
$b_{3\cdot2}$	$(EDTA)^2$	36.2	-
$b_{4\cdot2}$	$(TWEEN\ 20)^2$	-	15.6
$b_{5\cdot2}$	$(GSSG)^2$	-	−13.0

In the RSM model, the native activity of LIP was mostly dependent on the pH and the concentrations of TRIS·HCl and EDTA. The maximum of the model estimation was outside the borders set by the design space. An activity of 1171 U g^{-1} were estimated for the point at a corner (−2, 2, 2, 2, 2). However, this computational maximum was not verified experimentally. The highest experimentally determined activity was at 663 U g^{-1} (appendix Table 9.8). With regards to the model coefficients (b^{nat}), the pH was most important, it interacted with almost all other variables and exhibited a large quadratic term. In the RSM model, the optimal pH was the lower limit of the design, pH 8.

With respect to the refolded activity, the RSM model estimated the highest activities (929 U g^{-1}) at the same corner point (−2, 2, 2, 2, 2). In comparison, the highest experimentally determined refolded activity, 508 U g^{-1}, was low. Considering the model coefficients (b^{ref}), GSSG was of exceptional importance (highest linear regression coefficient) and various interactions with other variables could be identified (Figure 5.31). While the highest activities were estimated for pH 8 and maximal GSSG concentration (Figure 5.31, A), an extreme point existed for a higher pH with maximal refolded activities at 6.5 mM GSSG. The interactions with EDTA and TWEEN 20 at this extreme point are illustrated in Figure 5.31 B and C.

In conclusion, the RSM approach improved both native and refolded activities compared to the activities in the previous screening experiment. Maximum experimental values increased 10 % to 55 % for native (426 U g^{-1} to 663 U g^{-1}) and refolded (453 U g^{-1} to 508 U g^{-1}) activities, respectively. The polynomials estimated even higher activities for the corner points of the design space. Experimentally measured LIP activities were low compared to the 1750 U g^{-1} identified by the stochastic DOE approach (see 5.2.1).

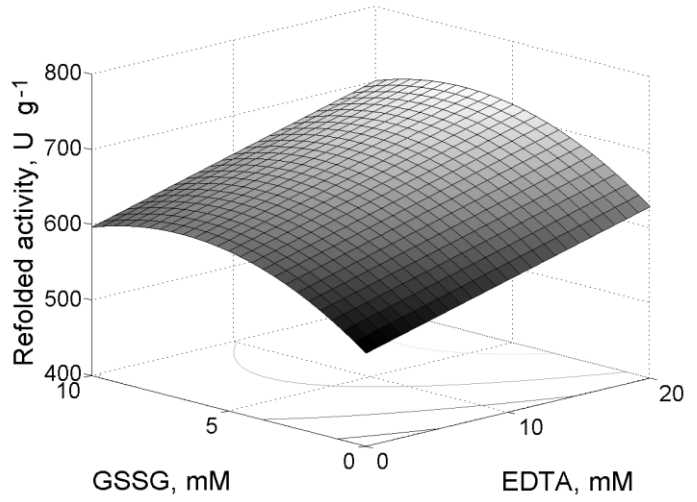


A (pH – GSSG)

1 M TRIS·HCl

20 mM EDTA

0.16 mM TWEEN 20

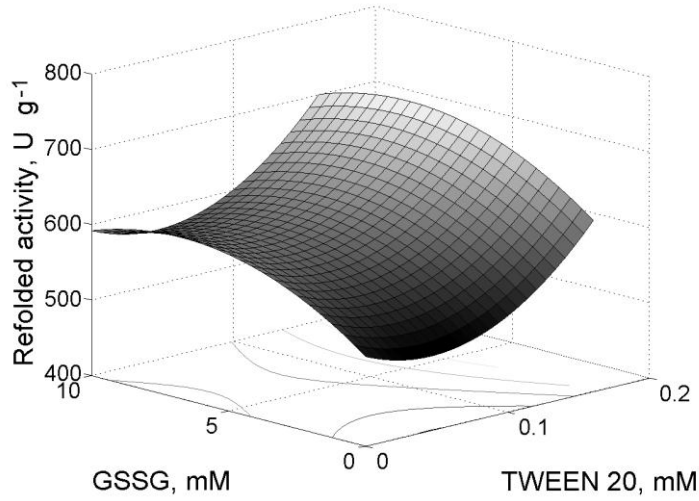


B (GSSG – EDTA)

pH 10

1 M TRIS·HCl

0.16 mM TWEEN 20



C (GSSG – TWEEN 20)

pH 10

1 M TRIS·HCl

20 mM EDTA

Figure 5.31: Visualization of the estimated interactions of GSSG in the RSM model for the refolded activity of LIP. (A) pH, (B) EDTA and (C) TWEEN 20.

Analysis of the complexity of LIP refolding

Standard statistical DOE with preliminary screening and subsequent optimization assumes that a linear approach is sufficient to identify essential variables in the screening. In order to evaluate, why the screening experiments failed to gain high refolded activities, the linear approach (Equation 7) was applied on the entire dataset of the stochastic optimizations (Figure 5.32, A). Using this simple process model with one constant and 26 linear terms it was only possible to achieve small correlation coefficients (R^2) of 0.6756 (native) and 0.5218 (refolded). Elevated refolded activities were estimated incorrectly and too low, as all normalized estimated refolded activities were smaller than 0.4. Hence, the linear approach of the screening did not capture all necessary information to correctly estimate LIP activity in the dataset from the stochastic optimizations.

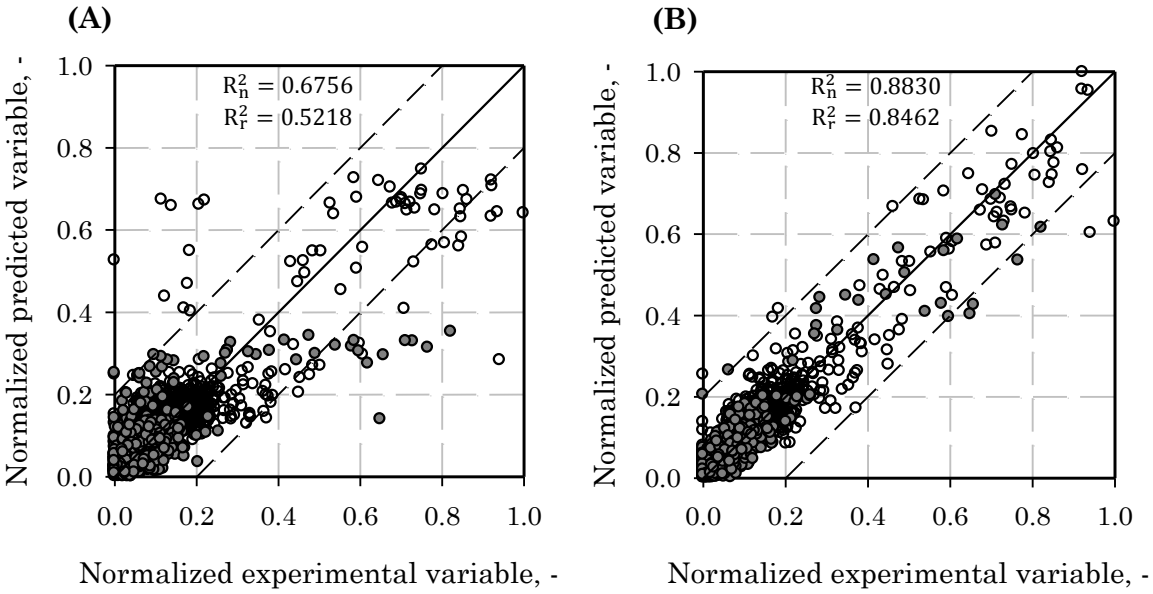


Figure 5.32: Regression models for the stochastic optimization of LIP refolding with 679 experiments. (A) linear model, (B) second order polynomial with interaction terms. (○) native activity, (●) refolded activity, (—) perfect fitting line, (– –) 20 % deviations limits, (R_n^2, R_r^2) correlation coefficients for the native and refolded activities.

Applying the more complex second order polynomial (Equation 8) of the RSM approach on the dataset was difficult, as the entire model included 378 terms: 1 constant, 26 linear, 325 interaction, 26 quadratic terms. A straightforward regression was not possible as the dataset (679 experiments) contained insufficient information. Hence, a regularized least-squares regression using *lasso* (Tibshirani, 1996) was carried out with the Matlabs (Mathworks) *lasso* function. Both activities could be fitted well (R^2 of 0.8830 and 0.8462) considering the experimental error of up to 20 % (Figure 5.32, B).

5.2.3 Discussion

In order to compare different DOE strategies, LIP refolding conditions were subject to an in depth analysis with the proposed stochastic optimization strategy and a standard statistical DOE approach. The stochastic optimization approach was applied multiple times. In total, 968 experimental conditions were examined in 44 GENs. Additionally, its performance was juxtaposed to a standard statistical approach revealing the importance of variable interactions in the problem.

Robustness of the stochastic optimization

An experimental analysis regarding the robustness of a stochastic DOE has not been published as of February 2012. This study on LIP refolding represents the first experimental problem, in which a stochastic search algorithm was evaluated multiple times on the same problem. All three optimizations (OPT_{I-III}) which incorporated SDS achieved high refolded LIP activities of about 1000 U g^{-1} or greater in alkaline refolding buffers with GSSG and GSH and SDS. This close similarity indicates that the proposed stochastic optimization strategy is quite robust, as the optima in each independent approach were roughly identical. In contrast to statistical DOE which starts with a predefined list of experiments (the experimental design), a stochastic optimization is not deterministic: Non-deterministic steps occur both in the beginning (random starting population) and in each GEN (mutation, crossing-over). Nevertheless, all three optimizations with the GA resulted in similar optima for the refolded activities of LIP. Therefore, the above-detailed stochastic aspects of the GA seem to be limited, even for an experimental problem with small population sizes and few GENs.

OPT_{II} and OPT_{III} obtained similar optima and slightly outperformed OPT_{I}, which despite evaluating 10 GENs (220 experiments) identified only a suboptimal solution. SDS occurred late in OPT_{I} and the percentage of buffers containing SDS in the entire optimization and the last GEN was drastically lower compared to the other two optimizations (Table 5.5). This indicated further potential and that higher activities in OPT_{I} could have probably been obtained by performing more GENs.

Despite the comprehensive dataset for the LIP, it is not justified to define the best refolding conditions identified in OPT_{II} and OPT_{III} as global optima. By principle, a stochastic optimization is only able to identify local optima: The method is non-deterministic and thus further experiments might identify better refolding conditions. The fact that both independent optimizations yielded very similar optima is a hint but

no proof. A more detailed analysis, either with a fine resolution statistical DOE or by substituting the applied GA with a hybrid optimization algorithm (Grosan and Abraham, 2007), might be able to determine even better activities. However, considering the large experimental error of up to 20 %, it is questionable if the measurements would provide enough resolution to distinguish between optimal and suboptimal solutions in close proximity.

Comparison to a standard two-step statistical DOE

In the initial step of the statistical DOE, the D-optimal screening used a simplified linear model to estimate the most important variables. Despite this simplification, most process variables that affected refolding were "correctly" identified. "Correctly" meaning that the entire LIP dataset exhibited similar results (section 5.3.2). However, another detergent outperformed SDS, which had tremendous impact on the activity in the stochastic optimizations (section 5.1.6) in the screening. Therefore, it was not included and successively optimized in the RSM. Although, the RSM increased native and refolded activities compared to the screening, it was not possible to identify the high activities which the GA discovered. The computational optima of the RSM models were outside the design space and the highest activities were measured for a point at the corner. In these outlying regions of the problem space, the model quality of the RSM is bound to be rather poor. A further optimization experiment with adjusted borders and an experimental validation would be necessary to evaluate this region of experimental space. The interactions of the pH and GSSG, which were identified in the RSM model, were observed previously (Ahn *et al.*, 1997; Willis *et al.*, 2005). Alkaline buffers (pH 7.5 to pH 9.0) are standard for oxidative refolding (Fischer *et al.*, 1993).

The importance of including interactions into the process model was highlighted by a regression analysis on the entire LIP dataset. A linear regression model was unsuitable to correctly estimate LIP refolding. Small correlation coefficients (R^2) of 0.52 and 0.68 compared well to Weuster-Botz (2000), who reported 0.45 to 0.6 for the linear regression analysis of datasets from stochastic optimizations. A second order polynomial, that incorporated interaction and quadratic terms (Equation 8) correctly estimated LIP activities (R^2 0.88 and 0.85). This model was in principle able to estimate LIP activities, but was overly complex (335 coefficients) and basically unfit for prediction purposes due to the generalization error (results not shown). However, it demonstrated in conjunction with the linear approach, that a correct estimation of LIP activity could only be achieved by integrating interaction and quadratic terms. Hence, a simple linear model for

screening and a successive optimization was inadequate to identify the "global" optimum conditions found by the GA. In summary, it was not possible to obtain the high activities previously determined in the stochastic DOE because SDS was excluded after the screening experiment. This highlighted the limits of the classic two-step statistical DOE: The linear model used to identify the most important process variables did not incorporate interactions, which proved to be integral for this optimization.

Regarding the experimental effort, the statistical DOE was far superior to the GA. It amounted to 30 experiments in the screening and 27 in the subsequent optimization. On the other hand, 22 experiments were performed in each GEN for the GA with a total effort $22 \cdot 10 = 220$ experiments for the first optimization. However, this comparison is only valid for a DOE approach using a linear screening as the first step. This approach however, failed to identify optimal conditions. Performing a "non-linear screening" that includes interaction terms would require far more experiments. For LIP, a problem with 27 variables, the model featured 378 terms and was overly complex (see above). It was difficult to fit, even with the entire dataset of the stochastic optimizations (679 experiments). Hence, the experimental effort of the stochastic optimization, which at first seemed quite high (220 experiments) is actually moderate compared to the statistical approach. Thus, the analyzed problem (LIP refolding) exemplified the efficiency of GAs for complicated multidimensional problems.

5.3 Modeling of the refolding conditions

Results from the experimental work (section 5.1 and 5.2) were subjected to a detailed analysis, which aimed to correlate the composition of the refolding buffer and the refolding success. More specifically, native and refolded activities as well as the relative refolding yield were to be described as a function of the refolding condition. The aim was to analyze all compounds of the refolding buffer and their effect on the activity and refolding yield. Analog to the regression analysis in the previous section (5.2.2), the dataset of the stochastic optimization of one protein was used as input, however more sophisticated modeling approaches were pursued: artificial neural networks (ANN) and bagged decision trees (BDT).

5.3.1 Preliminary models of LIP refolding

Modeling focused on lysozyme (LYZ) and the lipase from *Thermomyces lanuginosus* (LIP) as they provided far more experimental data than the other proteins. Work on LYZ yielded promising preliminary ANN models for the native activity measurements from the first two optimizations (results not shown). However, subsequent experiments with LYZ were carried out with a different enzyme batch and the resulting enzymatic activities differed significantly (compare 5.1.4). Therefore, an experimental validation was not possible and the remaining dataset (new enzyme batch, third optimization) was not large enough for new models. Consequently, further work concentrated on LIP. For this protein, five different optimizations had been performed resulting in a detailed picture of refolding conditions (see 5.2.1). In total, the LIP dataset contained 459 (preliminary model) to 767 (refined model) individual refolding experiments. Modeling variables and inputs are summarized in Table 5.11 and Table 5.12. Both datasets were normalized to obtain values between zero and one.

Table 5.11: Modeling variables for LIP. Native and refolded activities measured in the respective refolding conditions and the derived relative yield. All variables were normalized.

Modeling variable	Comment
Native activity	Measured only once for most refolding conditions, 15 % to 20 % standard error for multiple measurements
Refolded activity	Measured three-fold, standard error smaller than 20 %
Refolding yield	Calculated as the quotient of native and refolded activity (Equation 15), up to 35 % error

Table 5.12: Input variables of the LIP models. (Features) Normalized compounds of the refolding buffer ordered according to the functional classes (pH and buffer, salts, various additives, detergents, redox agents).

Nr.	Features of the refolding buffer
1	pH
2	sodium phosphate buffer, PB
3	hydroxylethyl-piperazine-ethanesulfonic acid, HEPES
4	morpholino-propanesulfonic acid, MOPS
5	tris-carboxyethyl-phosphine, TRIS·HCl
6	NaCl
7	KCl
8	glycerol
9	polyethylene glycol 4000, PEG 4000
10	arginine
11	glycine
12	glutamine
13	glutamate
14	Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2+}, mineral ions supplemented as sulfates
15	ethylenediaminetetraacetic acid, EDTA
16	cholamidopropyl-dimethylammonium-propanesulfonate, CHAPS
17	non-detergent sulfobetaine 201, NDSB 201
18	dodecyldimethyl-ammonio-propanesulfonate, ZWITTERGENT 3-12
19	polyethylene glycol sorbitan-monolaurate, TWEEN 20
20	polyethylene glycol tert-octylphenyl ether, TRITON-X 100
21	sodium dodecyl sulfate, SDS
22	polyethylene glycol dodecyl ether, BRIJ 35
23	dithiothreitol, DTT

Table 5.12 (continued):

Nr.	Features of the refolding buffer
24	tris-carboxyethyl-phosphine, TCEP
25	reduced L-glutathione, GSH
26	oxidized L-glutathione, GSSG

Artificial neural network (ANN) models

First, normalized data from LIP refolding experiments were used as input for ANNs (see section 3.4.1). While the variation of inputs, network size, architecture and training algorithms showed some positive effects (results not shown), ANN models generally performed well in training, but were unable to predict new data correctly. This high generalization error is exemplified in Figure 5.33 for two feed-forward ANN with 10 and 20 neurons in the hidden layer, respectively. In this example the native activity of a preliminary dataset (600 LIP refolding conditions) was modeled. The dataset was divided randomly using 70 % for training to adjust the network weights using the Levenberg-Marquard algorithm with backpropagation, while the remaining 188 experiments were used for an independent test and validation. A more complicated network increased the correlation coefficient for the training data (R^2) from 0.6846 to 0.9399, but it caused the quality of the validation results to decrease even further (R^2 0.4961 to 0.2003). Thus, the ANN was not able to predict new data. The key problem was the distribution of the dataset. LIP refolding was optimized by using a genetic algorithm (GA), a heuristic and stochastic optimization method. Therefore, the experimental dataset was not uniformly distributed. In some regions of the search space data were very sparse. Especially for highly active refolded protein, only few experiments had been performed. Consequently, the division of the dataset strongly influenced the model performance. The high level of noise in the dataset, that is the error of the activity measurements (Table 5.11), further complicated modeling. In addition to the native activity, refolded activity and yield were examined as well. For these two model variables the ANN performance was even lower compared to the native activity (results not shown).

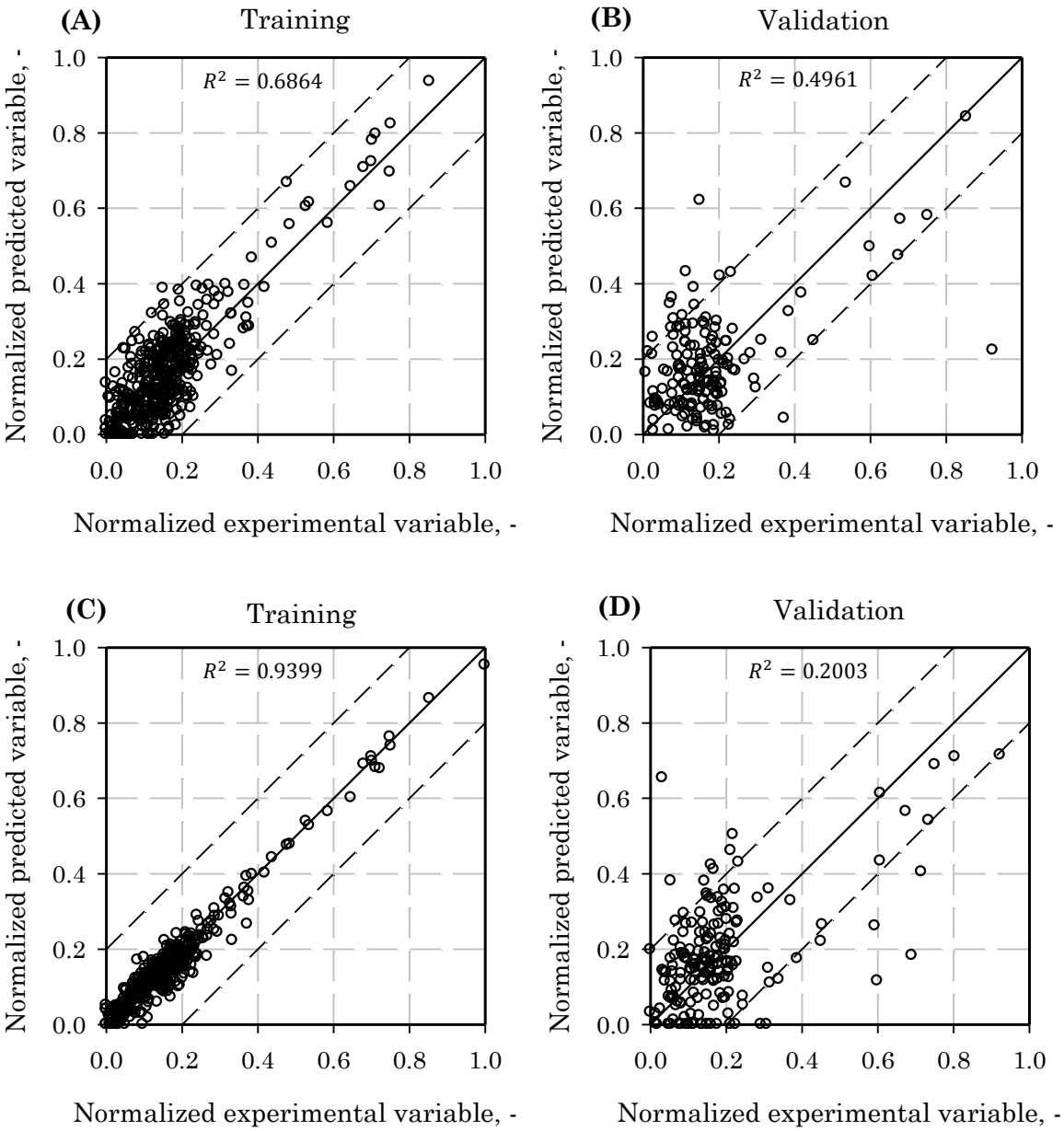


Figure 5.33: Training and validation of ANN models with 10 (A, B) or 20 (C, D) neurons in the hidden layer for LIP native activity. Training was performed with 413 experiments, test and validation with 188. (○) model data, (—) perfect fitting line, (– –) 20 % deviations limits.

Preliminary bagged decision tree (BDT) models without SDS

Due to the lack of performance of ANNs, other modeling approaches were pursued. As the data quality and distribution seemed to be problematic, ensemble system were focused on, which generally perform better on "non-ideal" datasets (Polikar, 2006). In particular, bagged decision trees (BDT) which are based on resampling the dataset, often perform well on noisy dataset and offer better generalization performances (see section 3.4.2).

BDT performance was evaluated in the training with the out-of bag mean square error: the prediction error of all observations that were not part of the bootstrap sample (out-of-bag). For the model generation, first the optimal minimum leaf size (minimum number of observations per tree leaf) was determined by comparing mean squared errors obtained by regression for various leaf sizes (Figure 5.34, A). A minimum leaf size of five was optimal for all BDT models. Additionally, the fraction of observations in the training data that were in-bag for all trees was monitored (Figure 5.34, B). Starting at about 2/3 for one tree (2/3 was the fraction selected by one bootstrap replica), the fraction of in-bag observations dropped to zero at approximately ten trees.

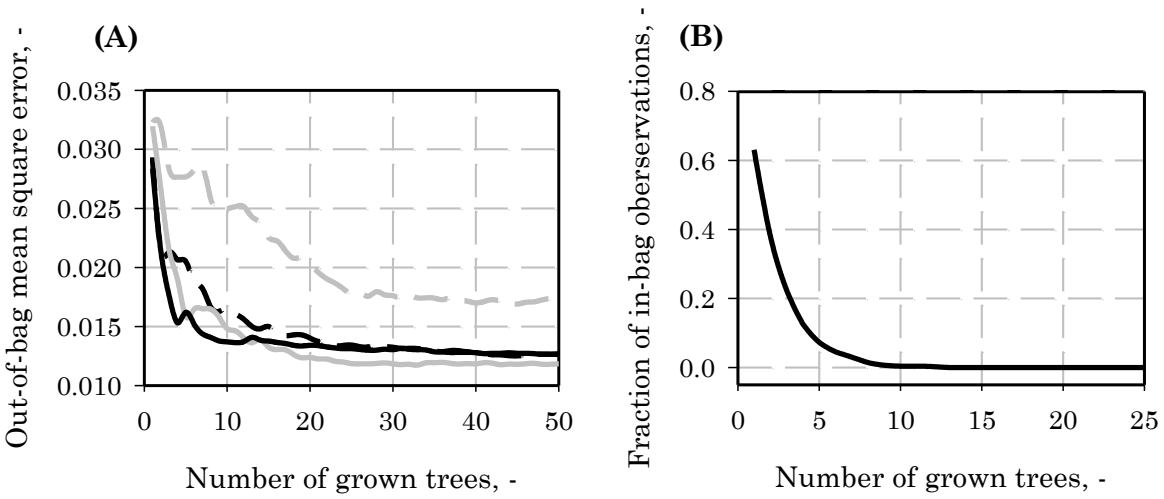


Figure 5.34: BDT model generation, example for LIP native activity. (A) Mean square error for models with different minimum leaf size (black, 1, grey 5, dashed black 10, dashed grey 50). (B) Fraction of in-bag observations for all trees.

Refolding experiments with SDS exhibited a low reproducibility if the protein was denatured with Gdn·HCl due to precipitation (see 5.1.6). Therefore, the first BDT models excluded all experiments incorporating SDS. Additionally, refolding experiments with urea denaturation were filtered out as well. The aim was a consistent minimal dataset which was restricted to refolding experiments performed with exactly the same experimental setup. This preliminary dataset of 459 LIP refolding conditions was later expanded in subsequent modeling approaches. Three different models were generated with the preliminary dataset: one for the native activity, the refolded activity and the relative refolding yield, respectively (Figure 5.35 and Figure 5.36).

In the parity plots (Figure 5.35, A and C) of predicted and measured LIP activities, more than 98 % of the observations were located within the limits of the standard deviation (20 %). Correlation coefficients (0.6735 and 0.8072) and model performance were moderate considering the high experimental error of up to 20 %. The architecture of the BDT model enables an easy estimation of the impact of each input variable or feature. Each feature represents one compound of the refolding buffer, for example the NaCl concentration (compare Table 5.12). In general the BDT prediction ability should depend more on important and less on unimportant features. By permuting the values of one feature across all observations and measuring the impact on the mean square error it is possible to obtain the out-of-bag feature importance (Figure 5.35, B and D). This value represents a significance measure for the model. It offers no information about the effect (positive or negative influence). For the native activity of LIP, NaCl, arginine and two reductive compounds were most important for the model prediction (Figure 5.35, B). While DTT and TCEP decreased LIP activities (see 5.1.9), the other two compounds had a positive effect. In contrast, the activity of the refolded protein was mostly affected by the pH (alkaline pH increased activities), mineral ions (negative), DTT (negative) and the GSH / GSSG (positive) redox system (Figure 5.35, D).

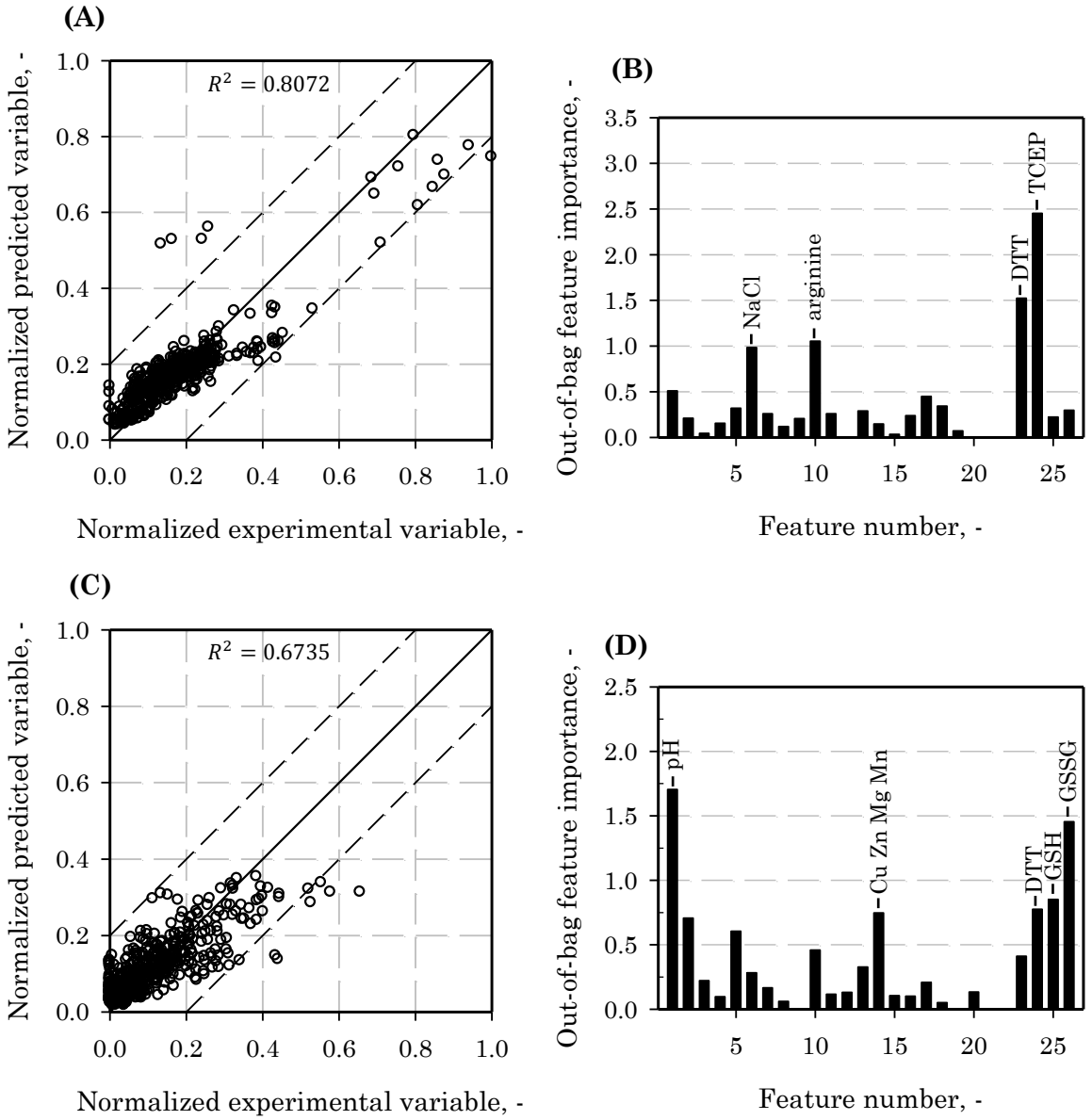


Figure 5.35: Preliminary BDT models for the native (A, B) and refolded (C, D) activity of LIP. The dataset contained 459 LIP refolding experiments without SDS. (A, C) Parity plots of experimental and predicted activity. (○) model data, (—) perfect fitting line, (– –) 20 % deviations limits. (B, D) Importance of the components of the refolding buffer (features) for the model performance.

The third model variable, the refolding yield, was calculated relative (Equation 15) as the quotient of the native and refolded activities. Due to the measurement errors in the LIP assay of 15 % to 20 % and the error propagation, yields were poorly defined (35 % standard deviation). In the BDT model an R^2 of 0.5573 was observed (Figure 5.36, A). Additionally, a clear bias was visible. Low refolding yields were overestimated and high refolding were underestimated by the model. The feature importance (Figure 5.36, B) was roughly equal to the refolded activity (compare Figure 5.35, D), though DTT and TCEP were less important.

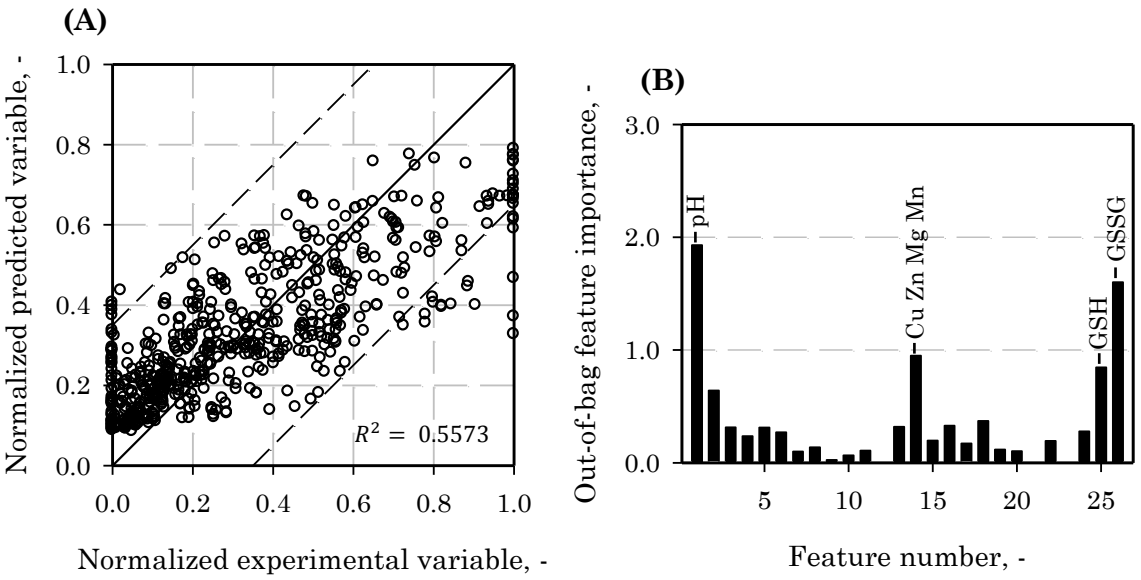


Figure 5.36: Preliminary BDT models for the relative refolding yield of LIP. The dataset contained 459 LIP refolding experiments without SDS. (A) Parity plot of experimental and predicted yield. (○) model data, (—) perfect fitting line, (– –) 35 % deviations limits. (B) Importance of the components of the refolding buffer (features) for the model performance.

Preliminary bagged decision tree (BDT) models with SDS

Only refolding conditions with SDS exhibited high LIP activities of up to 1750 U g^{-1} (section 5.2.1). Therefore, including SDS into the models was crucial. Hence, the previous dataset was merged with refolding experiments from the third stochastic optimization. The resulting 591 LIP refolding experiments were again modeled in respect to the three variables: native and refolded activities (Figure 5.37) and refolding yield (Figure 5.38).

Like before (Figure 5.35), refolding conditions with high activities were very sparse in the dataset and thus not uniformly distributed (Figure 5.37, A and C). Compared to the previous models, the R^2 values were slightly increased (0.8231 and 0.751). Most data were inside the limits of the experimental error. However, elevated refolded activities were underestimated (Figure 5.37, C). SDS was a very important feature for both models (Figure 5.37, B and D). In addition, DTT and TCEP strongly influenced the native activity comparable to the previous model. In contrast, the refolded activity was influenced by the pH and GSH / GSSG.

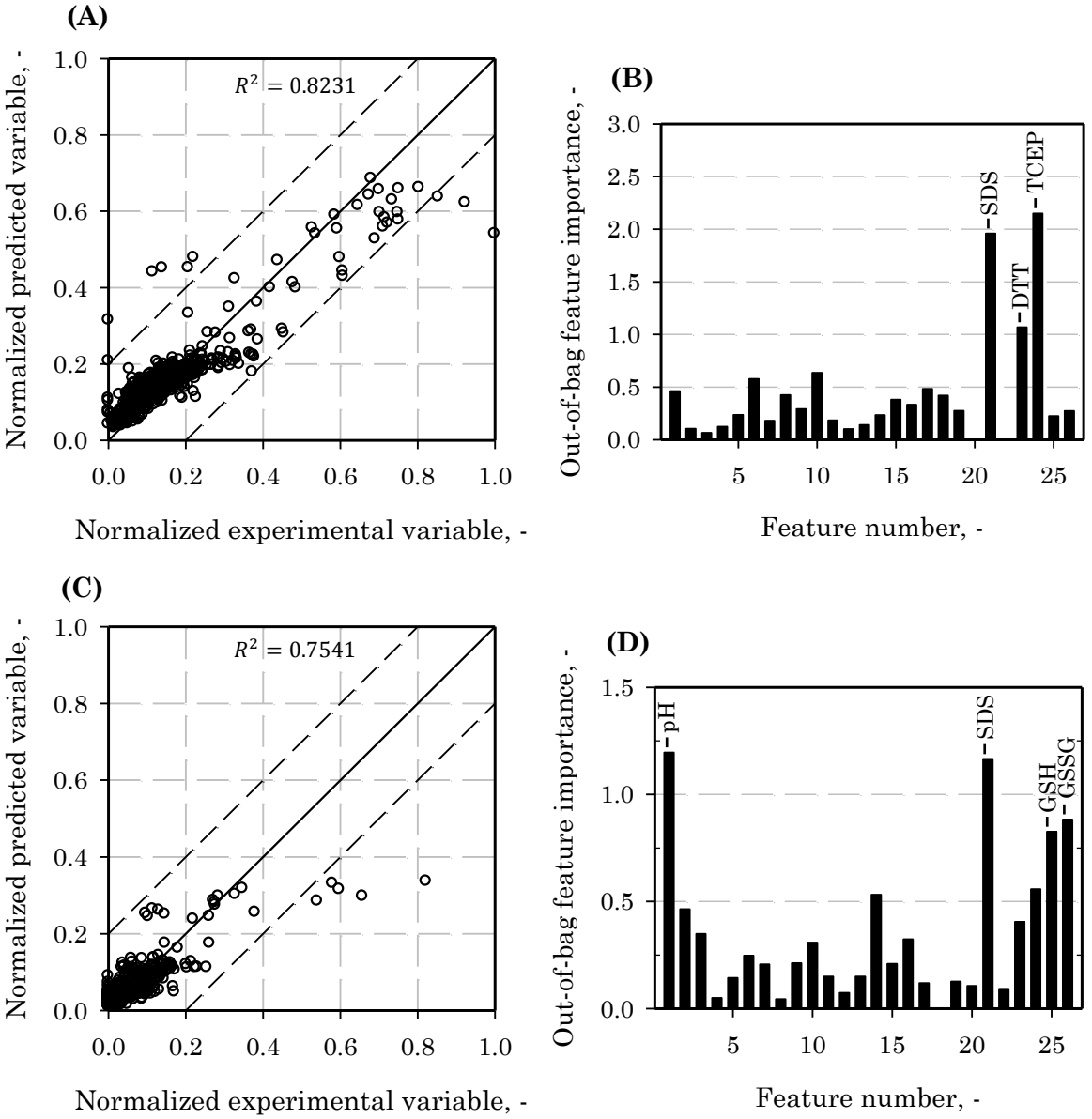


Figure 5.37: Preliminary BDT models for the native (A, B) and refolded (C, D) activity of LIP. The dataset contained 591 LIP refolding experiments including conditions with SDS. (A, C) Parity plots of experimental and predicted activity. (○) model data, (—) perfect fitting line, (– –) 20 % deviations limits. (B, D) Importance of the components of the refolding buffer (features) for the model performance.

Regarding the refolding yield, again a low correlation between experiments and model data (R^2 of 0.5106) and a bias occurred (Figure 5.38, A). However, the feature importance exhibited a very interesting picture (Figure 5.38, B) as SDS had no effect on the refolding yield itself. Only the pH, GSH and GSSG affected the prediction of the yield.

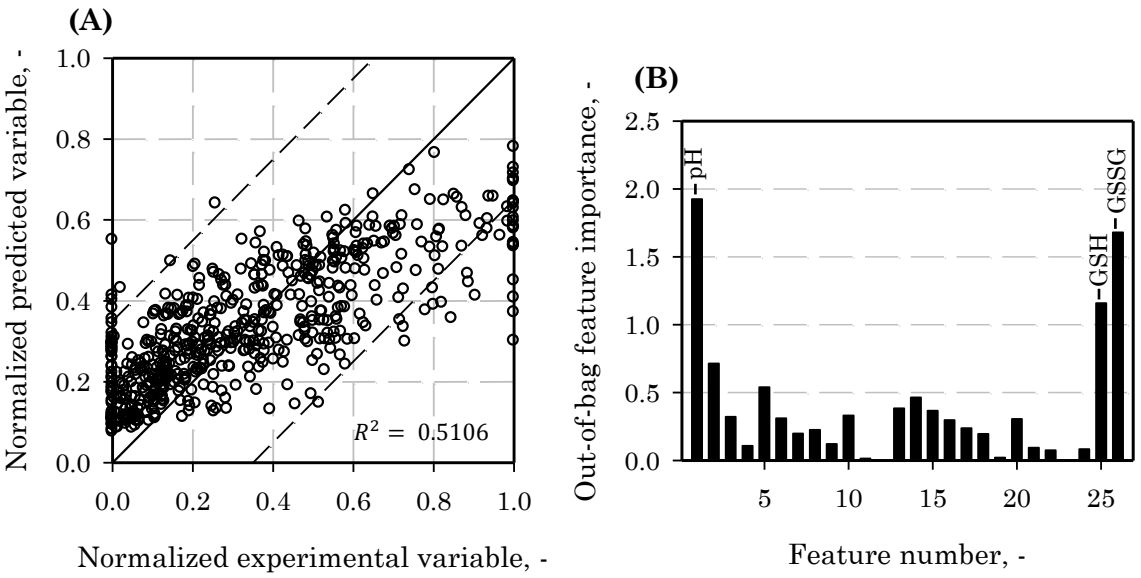


Figure 5.38: Preliminary BDT models for the relative refolding yield of LIP. The dataset contained 591 LIP refolding experiments including conditions with SDS. (A) Parity plot of experimental and predicted yield. (○) model data, (—) perfect fitting line, (– –) 35 % deviations limits. (B) Importance of the components of the refolding buffer (features) for the model performance.

5.3.2 Refined models of LIP refolding and experimental validation

The BDT models were refined and experimentally validated using the results of the continuation experiments of the first two stochastic optimizations (section 5.2.1). These contained many refolding conditions with SDS and high activities, thereby increasing the amount of observations with high activities and reducing the bias in the new dataset. Half of the new experiments were used for refinement (88), the other half for validation purposes (Figure 5.39 and Figure 5.40).

In the refined BDT models of the LIP activities (Figure 5.39, A and C), significantly higher R^2 of 0.8699 (native) and 0.8218 (refolded) were calculated. Thus, the refined models explained most of the variability in the dataset and over 99 % of the observations were within the boundaries of the standard deviation. Additionally, 88 refolding experiments were used for an independent validation. In the parity plots, these observations showed a similar distribution compared to the model data. R^2 for the validation data varied between 0.8205 (native) and 0.8292 (refolded) and were thus in good agreement with the model data. Feature importance was roughly equal to the previous models (compare Figure 5.37). However, the relative importance of SDS was increased (Figure 5.39, B and D).

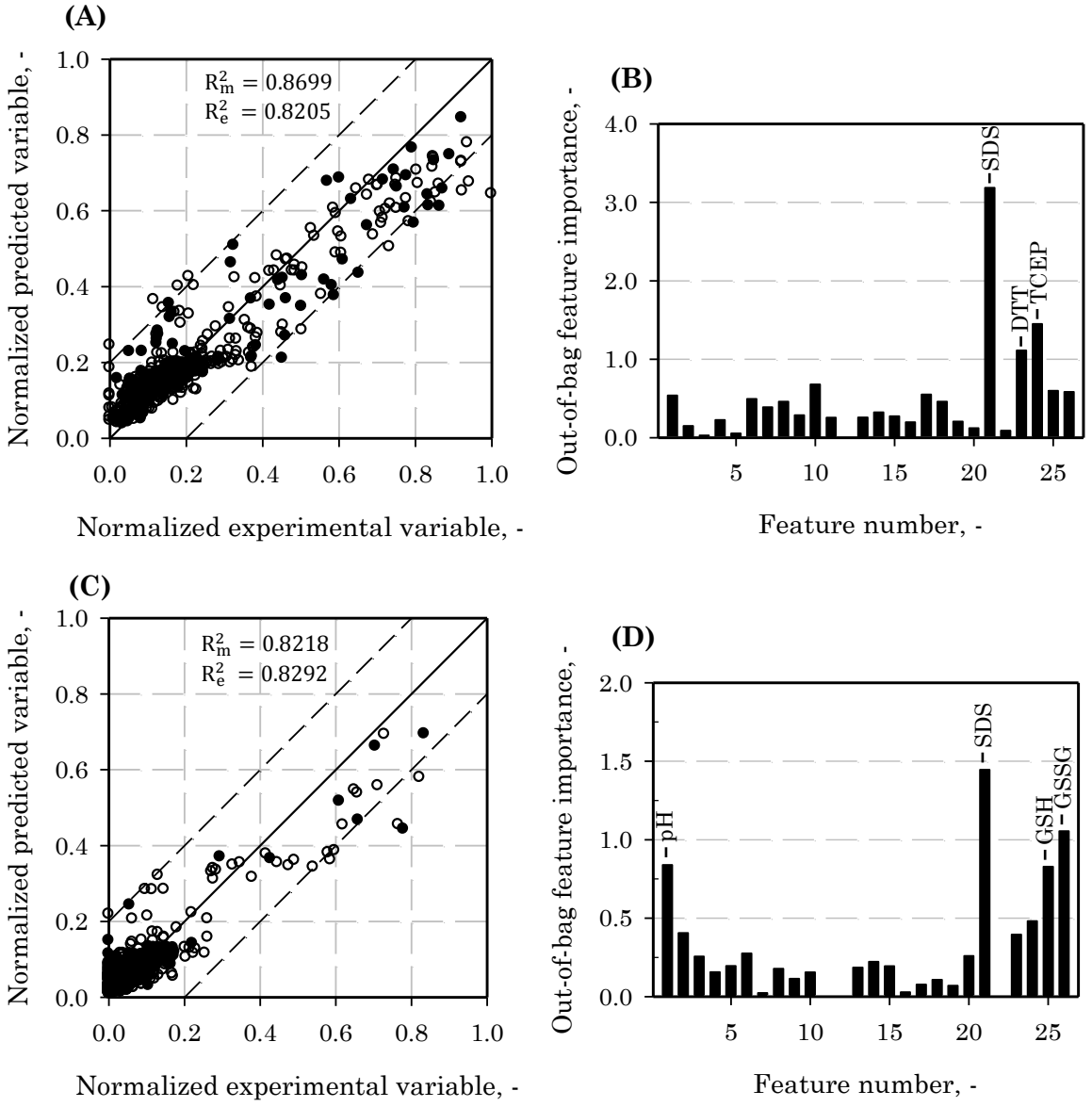


Figure 5.39: Refined BDT models for the native (A, B) and refolded (C, D) activity of LIP. (A, C) Parity plots of experimental and predicted activity. (○) model data 679 experiments, (●) independent validation 88 experiments, (—) perfect fitting line, (– –) 20 % deviations limits, (R^2_m, R^2_v) correlation coefficients for the model and validation data. (B, D) Importance of the components of the refolding buffer (features) for the model performance.

After refinement, the overall agreement of the predicted refolding yields to the experimental yields was moderate (Figure 5.40, A), both for model and validation data (R^2 of 0.7313 and 0.7165). Although the bias (to underestimate elevated yields and overestimate low yields) was reduced compared to the preliminary model (Figure 5.38, A), it was still visible. However, the statistical spread could be reduced. The feature importance remained constant, with the pH and the redox system (GSH and GSSG) as central components of the refolding buffer (Figure 5.40, B).

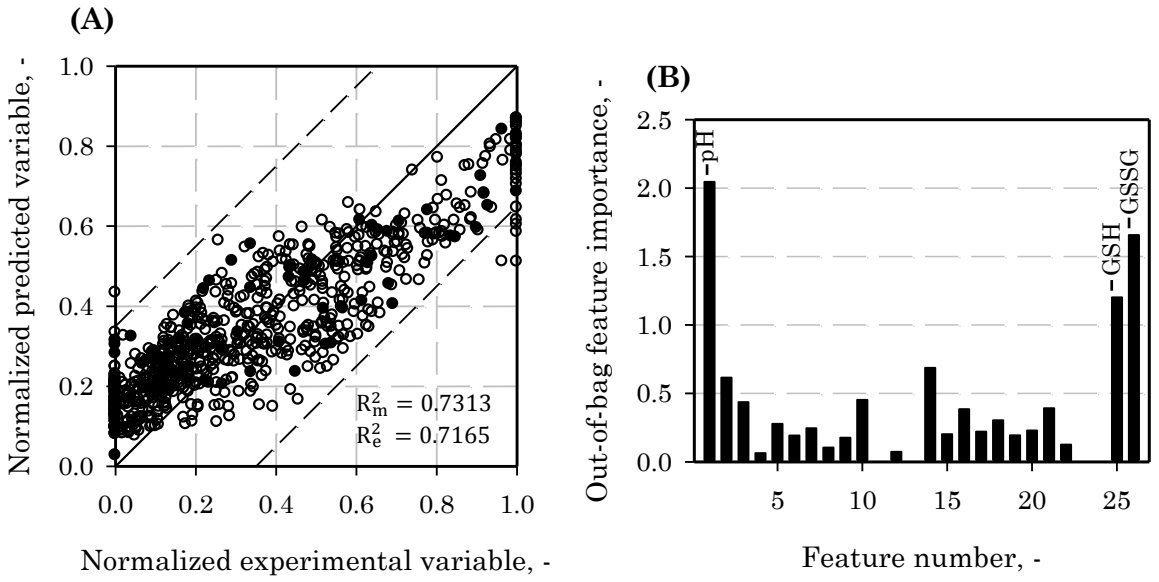


Figure 5.40: Refined BDT model for the relative refolding yield of LIP. (A) Parity plot of experimental and predicted yield. (○) model data 679 experiments, (●) independent validation 88 experiments, (—) perfect fitting line, (– –) 35 % deviations limits, (R^2_m, R^2_v) correlation coefficients for the model and validation data. (B) Importance of the components of the refolding buffer (features) for the model performance.

5.3.3 Quality of the experimental data and the model prediction

In the first stochastic optimization of LIP, the literature reference condition (Ahn *et al.*, 1997) showed only marginal refolding rendering it unsuitable as a standard condition (section 5.1.6). Hence, the best refolding condition from this optimization was chosen as a new refolding reference and measured in all further experiments. Consequently, a comprehensive overview of the experimental error and the reproducibility could be obtained. Figure 5.41 details the 38 measurements of the native and refolded activity of the stochastic optimizations (section 5.2.1) and the statistical DOE (section 5.2.2). In the majority of the experiments, 400 U g^{-1} to 500 U g^{-1} were measured for the native protein. Native activities tended to be lower in the first experiments (second and third optimization). Compared to the activity of the native enzyme, the refolded protein exhibited activities between 200 U g^{-1} and 300 U g^{-1} with a moderate statistical spread.

BDT models are based on many individual models (trees). Prediction is carried out for all individual models, afterwards the overall model prediction (prediction of the ensemble) is calculated by averaging. Hence, a BDT model offers a standard error that is easy to deduce. The three model predictions (M1, M2 and M3) for the reference condition are illustrated together with the mean experimental values (E) in Figure 5.41. Regarding the native activity, all three models predicted 400 U g^{-1} to 500 U g^{-1} with an error of

about 100 U g^{-1}. The magnitude was comparable to the experimental mean, though the predictions were slightly higher. In contrast, the agreement of model and experimental data was lower for the refolded activity. The unrefined model without SDS estimated significantly higher activities than the experiments. The other two models predicted approximately 200 U g^{-1} refolded activity, slightly lower than the measured LIP refolding. However, the relatively large experimental error has to be factored into the interpretation, hence the difference is only by trend and not significant.

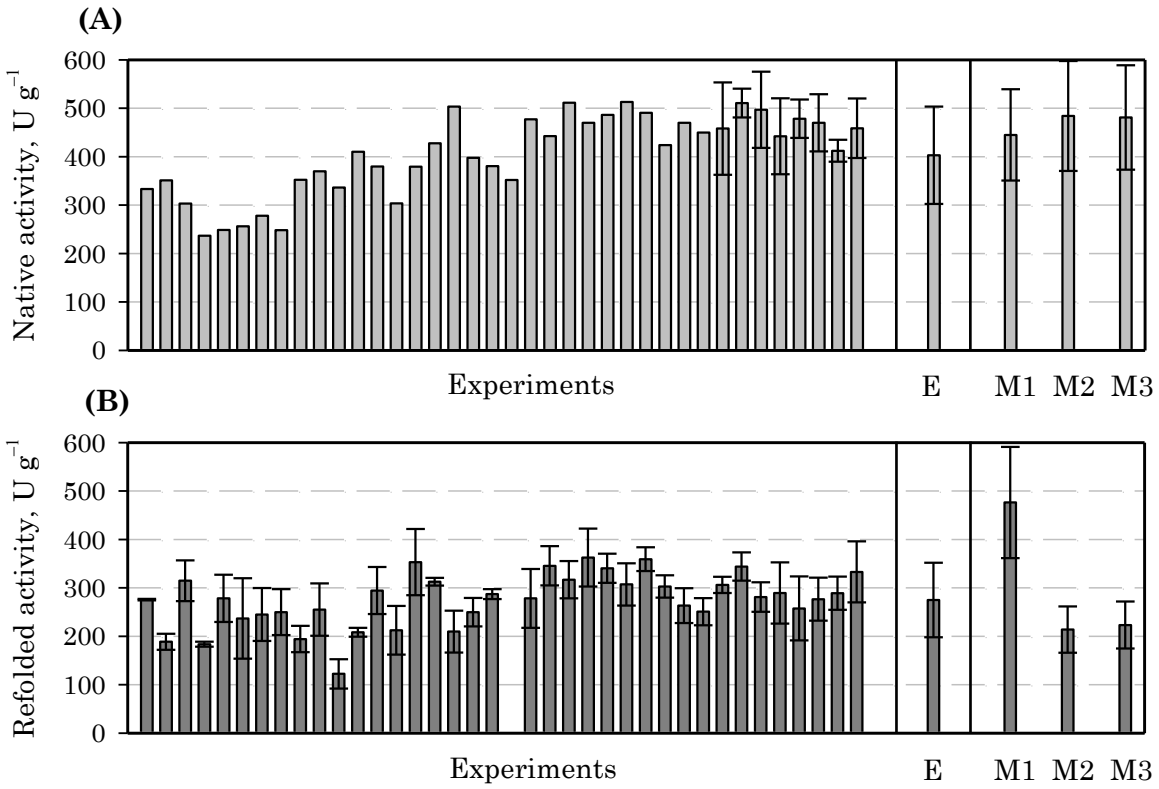


Figure 5.41: Native (A) and refolded (B) activity of the LIP reference conditions (100 mM MOPS, pH 9.25, 100 mM NaCl, 10 mM KCl, 5 % w/v PEG 4000, 300 mM arginine, 1000 mM NDSB, 1 mM GSH 2.5 mM GSSG) in all refolding experiments, the experimental mean value (E) and the three model predictions: M1 (unrefined model without SDS), M2 (unrefined model with SDS), M3 (refined model).

5.3.4 Discussion

Modeling focused on LIP for which roughly 1000 refolding conditions were experimentally evaluated, thus providing a detailed picture of the refolding buffer. This dataset was used to generate a model of the refolding success as a function of the buffer.

Insights

Modeling was severely affected by the large experimental error and the biased data distribution from the stochastic optimization. Only LIP, the protein with the largest dataset, was successfully modeled. Performance, that is the agreement between model and experimental data, was much better for the BDT models, especially the generalization error. The high generalization error of ANNs is one of the main drawbacks of this approach and has been studied thoroughly (Meireles *et al.*, 2003; Razi and Athappilly, 2005). Another advantage of the BDT models is the embedded error of the ensemble. Many individual models are averaged for the overall prediction, providing an easy to calculate standard error. For LIP, this ensemble model error was in most cases in good agreement with the experimental error observed for LIP refolding.

The available dataset of the stochastic optimization was pivotal for the performance of the model. Roughly 400 refolding experiments were necessary to generate adequate preliminary models for LIP. Consequently, more than 18 GENs with the GA would be necessary to generate sufficient data for a new target protein. This requirement could probably be drastically reduced if the experimental error of the measurements is decreased.

By measuring both native and refolded activity in the refolding experiments, it was possible to model three critical variables: the native activity, the refolded activity and the refolding yield. This differentiation enabled a more detailed look at the refolding buffer and made it possible to compare the impact on native protein and refolding. For example, SDS strongly influenced the activity of both native and refolded LIP, but had no effect on the refolding yield itself.

The feature importance of the BDT models was in good agreement with the composition of the best refolding conditions (compare Table 5.6) and the results of the statistical DOE (section 5.2.2). For the refolded LIP activity, the pH was important and alkaline buffers were advantageous. SDS strongly influenced the activity as well. GSH and GSSG were required for refolding, while purely reductive substances drastically reduced the activities.

Only LIP was successfully modeled with the BDT approach and the total number of proteins analyzed in this projects was small compared to the overall protein diversity. Therefore, it is difficult to generalize the results to other proteins. The major trend was the importance of the redox agents for disulfide-bridged (oxidative) and other proteins (reductive). In contrast to other studies (Ho and Middelberg, 2004; Zhang *et al.*, 2009), the isoelectric point (pI) of the protein seemed to have less importance. Almost all proteins preferred alkaline refolding conditions regardless of the pI. A general model for protein refolding requires data on more proteins and was not attainable within the scope of this thesis.

Application of the BDT model

The refined models of LIP were used for an *in silico* optimization of the refolding conditions. As an example application, the experimental optimizations of LIP (section 5.2.1) were analyzed in more detail. In this experiment one of the previous optimizations of LIP (the first optimization) was continued *in silico* (results not shown). By substituting the experimental evaluation with the BDT model of the refolding buffer the experimental effort could be avoided. This experiment allowed an estimation of the performance of the stochastic optimization in further GENs. According to the prediction, the slightly suboptimal optimization could achieve refolded activities that were identical to the other optimizations after three more GENs (1450 U g^{-1}, compare section 5.2.1).

This highlighted the potential application of the model in computational evaluations saving experimental effort and costs (Figure 5.42). Of course, the application is not limited to this aspect. It is possible to use the model to examine various similar aspects or questions. For example, a cost optimization or a restriction of the parameters to a subset is possible. Thereby, costly or unavailable compounds of the refolding buffer could be excluded. In addition, other downstream processing requirements like pH thresholds or the absence of detergents could be incorporated and optimized computationally. In both cases the optimization would proceed *in silico* and identify potential solutions for a sub-space of the search space. Afterwards, these refolding conditions can be verified experimentally. Hence, the costly experimental evaluation could be drastically reduced.

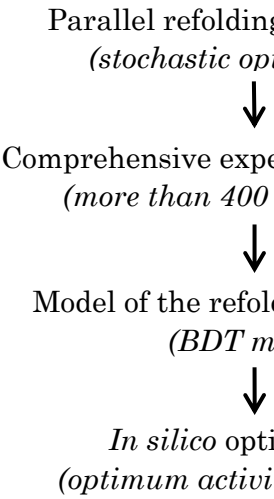


Figure 5.42: Generation and application of the model for protein refolding conditions. Protein refolding is optimized with the stochastic DOE. The generated data is used to train BDT models which can be applied for *in silico* optimization.

6 Conclusions

Soluble expression of recombinant proteins in *Escherichia coli* is often limited. To obtain the biologically active native form, additional processing steps are required, as the protein of interest aggregates inside the cell. The critical step is the reconstitution of the native structure (refolding), which represents a bottleneck in process development, as suitable conditions have to be evaluated in large screening experiments (Clark, 2001).

In this work, a new optimization strategy for protein refolding was developed, which used a genetic algorithm (GA) to optimize refolding in a standardized experimental design. For this purpose, experimental parameters were extracted from the refolding literature and combined with the information on roughly 1100 experiments from the REFOLD database to establish a comprehensive design. The new optimization strategy iteratively optimizes refolding conditions, specifically the 26 variables of the refolding buffer, using evolutionary principles (Figure 6.1).

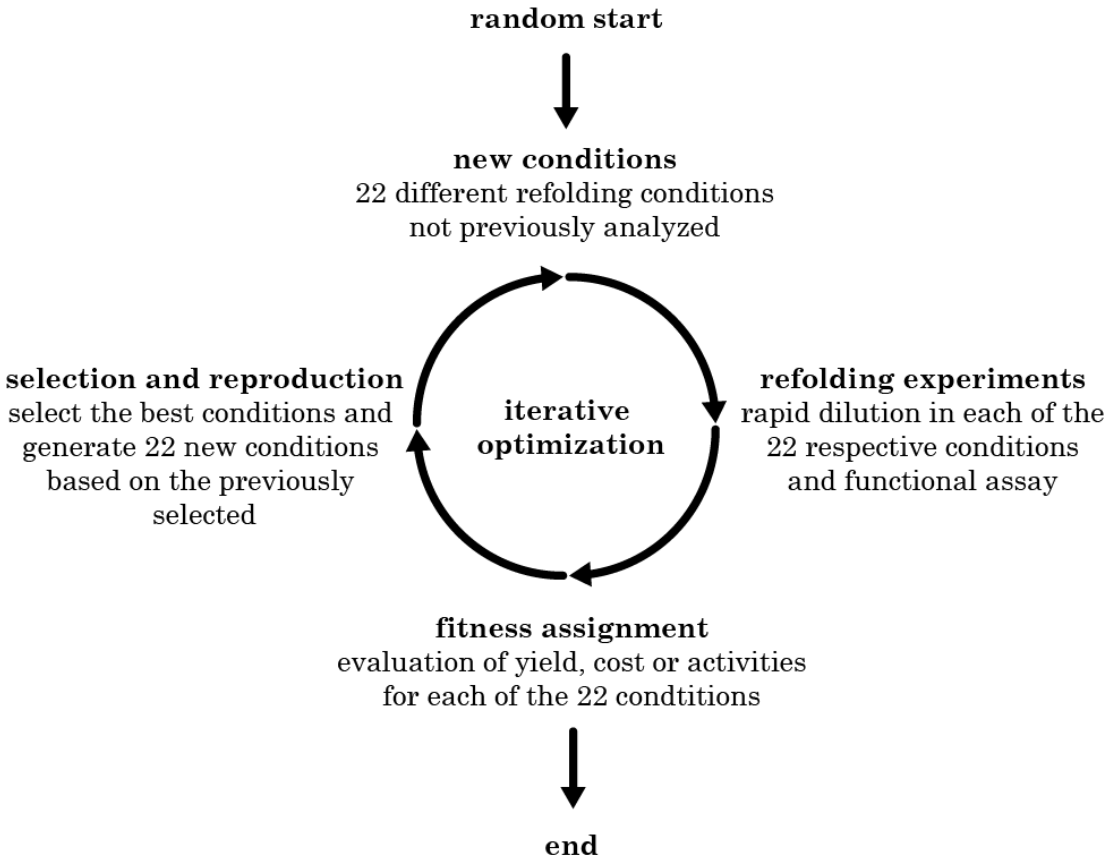


Figure 6.1: Scheme of the novel stochastic optimization strategy for protein refolding in mL-scale and 96-well plate format.

In order to demonstrate the suitability of this new approach, six structurally and functionally different model proteins were evaluated in a series of refolding experiments. For all proteins, the stochastic optimization identified comparable or better refolding conditions than the reference from the literature. For green fluorescent protein (GFP) and glutathione reductase (GLR), refolding yields and experimental costs were optimized, achieving 100 % refolding yield and costs of the refolding buffer between 0.006 € mL^{-1} and 0.012 € mL^{-1}. GLR and four other proteins, namely glucokinase (GLK), lysozyme (LYZ), lactate dehydrogenase (LDH) and the lipase from *Thermomyces lanuginosus* (LIP) were optimized with regards to the activity of the native and refolded protein. For all five proteins, the optimization of the enzymatic activities was successful, obtaining a 1.3-fold to 30.6-fold increase compared to the activity of the reference (Table 6.1). In addition, 100 % refolding yield could be achieved for all proteins except LDH.

Table 6.1: Performance of the new stochastic optimization strategy for protein refolding. Maximal measured native and refolded activities relative to the experimentally verified reference refolding conditions from the literature. (GFP) green fluorescent protein, (GLR) glutathione reductase, (GLK) glucokinase, (LYZ) lysozyme, (LDH) lactate dehydrogenase, (LIP) lipase from *Thermomyces lanuginosus*.

	GFP	**GLR**	**GLK**	**LYZ**	**LDH**	**LIP**
Improvement of native activity	n/a	130 %	130 %	40 %	40 %	5.3-fold
Improvement of refolded activity	n/a	250 %	30 %	40 %	40 %	30.6-fold
100 % yield	yes	yes	yes	yes	no	yes

Next, the proposed optimization strategy was characterized by analyzing the stochastic aspects of the GA and thus the overall robustness of the process. This was deemed necessary, as the optimization algorithm is not deterministic: Non-deterministic steps occur in the beginning (random start) and in each iteration (mutation, crossing-overs). Experiments were performed with LIP, a disulfide-bridged protein with a mass of 29 kDa. Three independent optimizations were evaluated. Within 10 iterations (generations, GEN) all optimizations achieved activities of approximately 1000 U g^{-1} or greater in alkaline refolding buffers with a mixture of oxidized and reduced glutathione (GSSG and GSH) and sodium dodecyl sulfate (SDS) (Table 6.2). The similar optima determined in each independent optimization approach indicated that the stochastic

optimization strategy is quite robust. Therefore, the above-detailed stochastic aspects of the GA seem to be limited, even for an experimental problem with small population sizes (22) and few GENs (10). To the best of our knowledge, this study represents the first case in which a GA was evaluated multiple times on the same experimental problem. According to these results, stochastic optimization strategies seem to be quite robust and suitable for experimental problems.

Table 6.2: Robustness of the stochastic optimization strategy for protein refolding. Optimal refolding conditions of LIP identified in three independent optimizations. Composition, individual activities of the native and refolded protein (* U g^{-1}) and yield in the refolding conditions with the highest refolded activity for each optimization.

Best LIP refolding condition (highest refolded activity) in each optimization	**Native activity***	**Refolded activity***	**Relative yield, %**
Optimization one 100 mM MOPS, pH 9.25, 350 mM NaCl, 25 mM glutamate, 7.5 mM EDTA, 3 mM SDS, 3.75 mM GSH, 0.5 mM GSSG	1062 ± 296	977 ± 33	92 ± 29
Optimization two 750 mM TRIS·HCl, pH 7.5, 50 mM KCl, 25 mM arginine, 50 mM glutamine, 12 mM SDS, 5 mM GSH, 5 mM GSSG	1451 ± 286	1335 ± 172	92 ± 32
Optimization three 500 mM TRIS·HCl, pH 8.5, 175 mM NaCl, 50 mM KCl, 0.05 % w/v PEG 4000, 250 mM arginine, 200 mM glutamate, 12 mM SDS, 0.5 mM GSH, 5 mM GSSG	1306 ± Na	1430 ± 175	100 ± 12

MOPS, morpholino-propanesulfonic acid; **TRIS**, tris(hydroxymethyl)aminomethane; **EDTA**, ethylenediaminetetraacetic acid; **GSH**, reduced glutathione; **GSSG**, oxidized glutathione; **PEG**, polyethylene glycol

In the next step, the stochastic optimization was compared to a standard two-step statistical design of experiments (DOE), which included a D-optimal screening experiment and the subsequent optimization by response surface methodology (RSM). The D-optimal screening used a simplified linear process model to estimate the most important process variables. Although many variables which affected LIP refolding were correctly identified, the importance of SDS was underestimated. Therefore, SDS was not optimized in the subsequent RSM and it was not possible to obtain the high activities determined in the stochastic DOE (508 U g^{-1} versus 1430 U g^{-1} refolded activity).

Applying the linear model on the entire LIP dataset gave poor estimates for the activities and highlighted the limitations of the linear process model. The regression models featured small correlation coefficients (R^2) of 0.52 and 0.68, which compared well to Weuster-Botz (2000), who reported 0.45 to 0.60 for the linear regression analysis of datasets from stochastic optimizations. Thus, the linear approach was insufficient and interactions seemed to be essential to give a correct estimate of the activity. This assumption was reinforced by the regression analysis with a second order polynomial, that incorporated interaction (325) and quadratic (26) terms, which correctly estimated (R^2 0.88 and 0.85) LIP activities.

Using this complex model (378 terms) for a non-linear statistical DOE would drastically increase the experimental effort from 30 to at least 379 experiments. Hence, the experimental effort of the GA, which at first seemed quite high (22 experiments in 10 GENs = 220 experiments), is actually moderate. Therefore, LIP refolding seems to be a good example for the efficiency of GAs for complicated multidimensional problems.

Finally, black box models were trained on the experimental data to predict the refolding success dependent on the composition of the refolding buffer. Modeling was severely affected by the large experimental error (up to 35 % standard deviation) and the biased data distribution from the stochastic optimization. Only LIP, which was subjected to an in depth analysis within this thesis (about 1000 refolding experiments), could be successfully modeled with the robust bagged decision tree (BDT) approach. For LIP, experimental and predicted values were in good agreement even for the independent validation (R^2 greater than 0.8). Almost all results were within the limits of the standard deviation (Figure 6.2, A). BDT models include an embedded estimate of the variable importance which allows visualization of the significance for the prediction (Figure 6.2, B). In the LIP model, an alkaline pH, SDS and the presence of glutathione were most important for high activities of the refolded protein. This was in good agreement with composition of the best refolding conditions identified by the stochastic optimization (compare Table 6.2). The LIP model was successfully applied for an *in silico* optimization. In this experiment one of the previous optimizations of LIP (the first optimization) was continued *in silico*. Costly experiments could be avoided by substituting the experimental evaluation with the model of the refolding buffer. According to the prediction, the slightly suboptimal first optimization could achieve refolded activities that were identical to the other optimizations after three more GENs (1450 U g^{-1}, compare Table 6.2). This highlighted the potential applications of the model in computational evaluations saving experimental effort and costs.

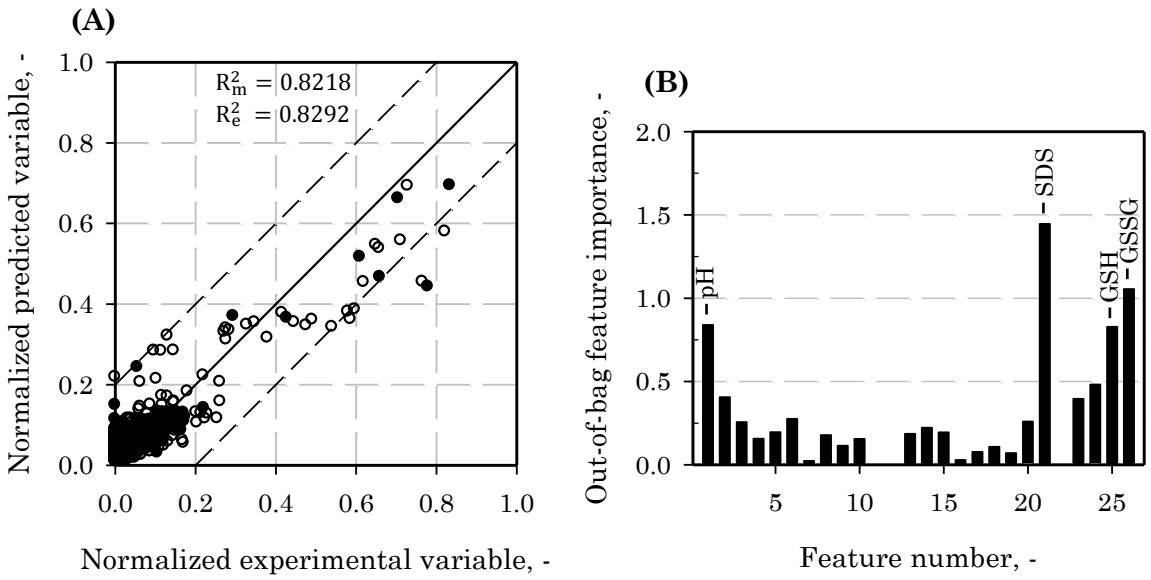


Figure 6.2: BDT model for the refolded activity of LIP. (A) Parity plot of experimental and predicted refolded activity. (○) model data 679 experiments, (●) independent validation 88 experiments, (—) perfect fitting line, (– –) 20 % deviations limits, (R^2_m, R^2_v) correlation coefficients for the model and validation data. (B) Importance of the components of the refolding buffer (features) for the model performance.

7 Outlook

There are two possibilities to integrate the results into the optimization of new target proteins. On the one hand, bagged decision tree (BDT) models could be embedded into the stochastic optimization by evaluating several generations in experiments in the first place. Afterwards, a model is trained on the acquired data and in turn, refolding is optimized *in silico*, thus saving experimental effort (Franco-Lara *et al.*, 2006). The principle was successfully examined for the refolding of LIP by continuing the first optimization *in silico*. However, the large datasets required for an adequate model (roughly 1000 experiments for LIP) pose a serious drawback. On the other hand, it is possible to adjust the search space of the stochastic optimization according to the protein of interest. One of the major trends was the preference of oxidative conditions (GSH and GSSG) for extracellular proteins with disulfide bonds (LYZ, LIP) and reductive conditions for the rest. Hence, purely reductive conditions could be excluded for proteins with disulfide bonds in order to limit the search space and speed up the optimization. This approach was evaluated for LYZ, which was optimized with the normal GA setup, a modified setup for redox conditions and modified redox conditions in conjunction with a threshold for ionic strength.

The search space for new target proteins can be constrained in a similar way by using sequence information or sequence based predictions (Liu, 2007; Sankararaman *et al.*, 2010). Alternatively, if the protein is closely related to a protein that has been previously optimized, it would also be possible to adjust the search space accordingly. For proteins which were already successfully refolded but not optimized, it would probably be more efficient to substitute the global GA optimization with a local optimization using either a standard optimization algorithm or a RSM approach. The proposed workflow for the optimization of new target proteins is illustrated in Figure 7.1.

As the total number of proteins analyzed in this projects is small compared to the protein diversity, it is difficult to generalize these results. The major trend was the importance of the redox agents for disulfide-bridged (oxidative) and other proteins (reductive). In contrast to other studies (Ho and Middelberg, 2004; Zhang *et al.*, 2009), the isoelectric point (pI) of the protein seemed to have less importance. Almost all proteins preferred alkaline refolding conditions regardless of the pI. A general model that predicts refolding, requires data on more proteins and was not attainable within the scope of this

thesis. Further work should focus on examining more proteins and in parallel integrate additional information from REFOLD (Amin *et al.*, 2006; Buckle *et al.*, 2005) and other databases.

Next to direct application for protein refolding, it might also be possible to use the generated BDT model as a "real world" search space in order to evaluate other optimization algorithms or methods. The comprehensive dataset with realistic error data provides the potential to thoroughly test new methods in a realistic environment.

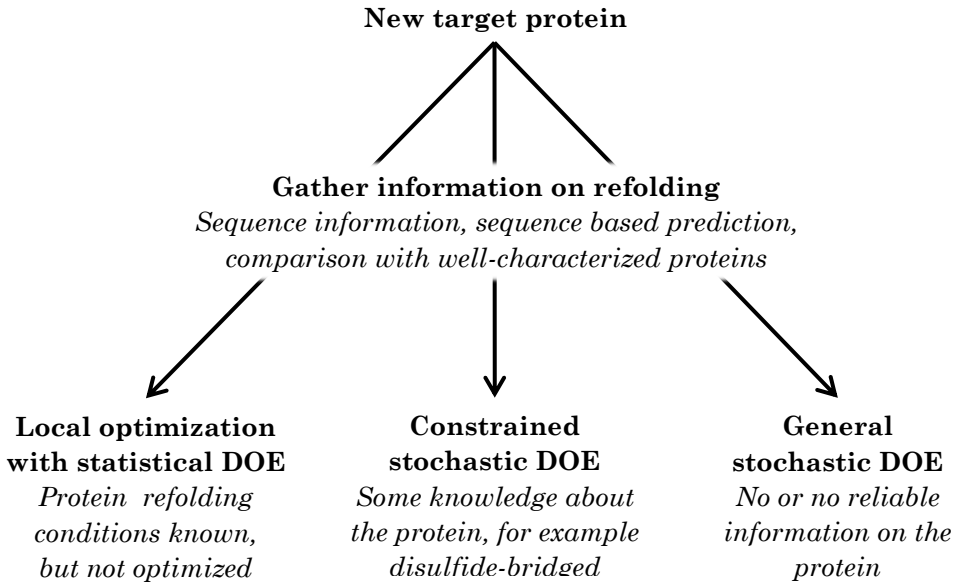


Figure 7.1: Proposed workflow for new target proteins. After detecting insolubility either experimentally or *in silico* (Diaz *et al.*, 2010; Magnan *et al.*, 2009), the protein is mapped to well characterized proteins and the protein sequence and sequence based prediction algorithms are utilized to obtain as much information as possible on refolding. Subsequently, three different DOE approaches are proposed depending on the amount of information on the target protein.

8 References

Agatonovic-Kustrin S, Beresford R. (2000): Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis* 22, 717-27.

Ahn JH, Lee YP, Rhee JS. (1997): Investigation of refolding condition for Pseudomonas fluorescens lipase by response surface methodology. *Journal of Biotechnology* 54, 151-160.

Altamirano MM, García C, Possani LD, Fersht a R. (1999): Oxidative refolding chromatography: folding of the scorpion toxin Cn5. *Nature biotechnology* 17, 187-91.

Altamirano MM, Woolfson a, Donda a, Shamshiev a, Briseño-Roa L, Foster NW, Veprintsev DB, De Libero G, Fersht a R, Milstein C. (2001): Ligand-independent assembly of recombinant human CD1 by using oxidative refolding chromatography. *Proceedings of the National Academy of Sciences of the United States of America* 98, 3288-93.

Amin AA, Bottomley SP, Chow MKM, Fulton KF, Whisstock JC, Buckle AM. (2006): REFOLD: an analytical database of protein refolding methods. *Protein expression and purification* 46, 166-71.

Andersen DC, Krummen L. (2002): Recombinant protein expression for therapeutic applications. *Current opinion in biotechnology* 13, 117-23.

Anfinsen CB. (1972): The formation and stabilization of protein structure. *The Biochemical journal* 128, 737-49.

Arakawa T, Timasheff SN. (1984): Protein stabilization and destabilization by guanidinium salts. *Biochemistry* 23, 5924-9.

Arakawa T, Ejima D, Tsumoto K, Obeyama N, Tanaka Y, Kita Y, Timasheffe SN. (2007): Suppression of protein interactions by arginine: a proposed mechanism of the arginine. *Biophysical chemistry* 127, 1 - 8.

Armstrong N, de Lencastre A, Gouaux E. (1999): A new protein folding screen: application to the ligand binding domains of a glutamate and kainate receptor and to lysozyme and carbonic anhydrase. *Protein science : a publication of the Protein Society* 8, 1475-83.

Ayling A, Baneyx F. (1996): Influence of the GroE molecular chaperone machine on the in vitro refolding of Escherichia coli beta-galactosidase. *Protein science : a publication of the Protein Society* 5, 478-87.

Back T, Hammel U, Schwefel H-P. (1997a): Evolutionary computation: comments on the history and current state. *IEEE Transactions on Evolutionary Computation* 1, 3-17.

Back T, Fogel DB, Michalewicz Z. (1997b): Handbook of Evolutionary Computation. IOP Publishing Ltd.

Balbach J, Forge V, van Nuland NA, Winder SL, Hore PJ, Dobson CM. (1995): Following protein folding in real time using NMR spectroscopy. *Nature structural biology* 2, 865-70.

Basu A, Li X, Leong SSJ. (2011): Refolding of proteins from inclusion bodies: rational design and recipes. *Applied microbiology and biotechnology* 92, 241-51.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. (2000): The Protein Data Bank. *Nucleic acids research* 28, 235-42.

De Bernardez Clark E, Schwarz E, Rudolph R. (1999): Inhibition of aggregation side reactions during in vitro protein folding. *Methods in enzymology* 309, 217-36.

Bessette PH, Aslund F, Beckwith J, Georgiou G. (1999): Efficient folding of proteins with multiple disulfide bonds in the Escherichia coli cytoplasm. *Proceedings of the National Academy of Sciences of the United States of America* 96, 13703-8.

Betts SD, King J. (1998): Cold rescue of the thermolabile tailspike intermediate at the junction between productive folding and off-pathway aggregation. *Protein science : a publication of the Protein Society* 7, 1516-23.

Bhat R, Timasheff SN. (1992): Steric exclusion is the principal source of the preferential hydration of proteins in the presence of polyethylene glycols. *Protein science : a publication of the Protein Society* 1, 1133-43.

Bianchi L, Dorigo M, Gambardella LM, Gutjahr WJ. (2008): A survey on metaheuristics for stochastic combinatorial optimization. *Natural Computing* 8, 239-287.

Boyle DM, Buckley JJ, Johnson GV, Gustafson ME, Rathore A. (2009): Use of the design-of-experiments approach for the development of a refolding technology for progenipoietin-1, a recombinant human cytokine fusion protein from Escherichia coli inclusion bodies. *Biotechnology and applied biochemistry* 54, 85-92.

Boyle DM, Johnson GV, Heeren RA, Shell RE, Banerjee A, Gustafson ME. (2008): Evaluation of refolding conditions for a human recombinant fusion cytokine protein, promegapoietin-1a. *Biotechnology and applied biochemistry* 49, 73-83.

Bradford MM. (1976): A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry* 72, 248-54.

Breiman L. (2001): Random Forests. *Machine Learning* 45, 5–32.

Breiman L, Friedman J, Stone CJ, Olshen RA. (1993): Classification and Regression Trees. Chapman and Hall / CRC.

Brzozowski a M, Savage H, Verma CS, Turkenburg JP, Lawson DM, Svendsen A, Patkar S. (2000): Structural origins of the interfacial activation in Thermomyces (Humicola) lanuginosa lipase. *Biochemistry* 39, 15071-82.

Buchfink R, Tischer A, Patil G, Rudolph R, Lange C. (2010): Ionic liquids as refolding additives: variation of the anion. *Journal of biotechnology* 150, 64-72.

Buckle AM, Devlin GL, Jodun RA, Fulton KF, Faux N, Whisstock JC, Bottomley SP. (2005): The matrix refolded. *Nature methods* 2, 3.

Cabrita LD, Bottomley SP. (2004): Protein expression and refolding--a practical guide to getting the most out of inclusion bodies. *Biotechnology annual review* 10, 31-50.

Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. (1994): Green fluorescent protein as a marker for gene expression. *Science (New York, N.Y.)* 263, 802-5.

Chi EY, Krishnan S, Randolph TW, Carpenter JF. (2003): Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharmaceutical research* 20, 1325-36.

Cho H-J, Lee Y, Chang RS, Hahm M-S, Kim M-K, Kim YB, Oh Y-K. (2008): Maltose binding protein facilitates high-level expression and functional purification of the chemokines RANTES and SDF-1alpha from Escherichia coli. *Protein expression and purification* 60, 37-45.

Choe W, Nian R, Lai W. (2006): Recent advances in biomolecular process intensification. *Chemical Engineering Science* 61, 886-906.

Clark ED. (2001): Protein refolding for industrial processes. *Current opinion in biotechnology* 12, 202-7.

Coello Coello CA. (2006): Evolutionary multi-objective optimization: a historical view of the field. *IEEE Computational Intelligence Magazine* 1, 28-36.

Collinson LP, Dawes IW. (1995): Isolation, characterization and overexpression of the yeast gene, GLR1, encoding glutathione reductase. *Gene* 156, 123-7.

Covas JA, Cunha AG, Oliveira P. (1999): An optimization approach to practical problems in plasticating single screw extrusion. *Polymer Engineering & Science* 39, 443-456.

Cowan RH, Davies RA, Pinheiro TTJ. (2008): A screening system for the identification of refolding conditions for a model protein kinase, p38alpha. *Analytical biochemistry* 376, 25-38.

Cowieson NP, Wensley B, Listwan P, Hume DA, Kobe B, Martin JL. (2006): An automatable screen for the rapid identification of proteins amenable to refolding. *Proteomics* 6, 1750-7.

Cregg JM, Cereghino JL, Shi J, Higgins DR. (2000): Recombinant protein expression in Pichia pastoris. *Molecular biotechnology* 16, 23-52.

Crisman RL, Randolph TW. (2009): Refolding of proteins from inclusion bodies is favored by a diminished hydrophobic effect at elevated pressures. *Biotechnology and bioengineering* 102, 483-92.

Danielsen S, Eklund M, Deussen HJ, Gräslund T, Nygren P a, Borchert TV. (2001): In vitro selection of enzymatically active lipase variants from phage libraries using a mechanism-based inhibitor. *Gene* 272, 267-74.

Dashivets T, Wood N, Hergersberg C, Buchner J, Haslbeck M. (2009): Rapid matrix-assisted refolding of histidine-tagged proteins. *Chembiochem : a European journal of chemical biology* 10, 869-76.

Davies RC, Neuberger A, Wilson BM. (1969): The dependence of lysozyme activity on pH and ionic strength. *Biochimica et biophysica acta* 178, 294-305.

Dechavanne V, Barrillat N, Borlat F, Hermant A, Magnenat L, Paquet M, Antonsson B, Chevalet L. (2011): A high-throughput protein refolding screen in 96-well format combined with design of experiments to optimize the refolding conditions. *Protein expression and purification* 75, 192-203.

Dejaegher B, Heyden YV. (2011): Experimental designs and their recent advances in set-up, data interpretation, and analytical applications. *Journal of pharmaceutical and biomedical analysis* 56, 141-58.

Derewenda U, Swenson L, Wei Y, Green R, Kobos PM, Joerger R, Haas MJ, Derewenda ZS. (1994): Conformational lability of lipases observed in the absence of an oil-water interface: crystallographic studies of enzymes from the fungi Humicola lanuginosa and Rhizopus delemar. *Journal of lipid research* 35, 524-34.

Diamond R. (1974): Real-space refinement of the structure of hen egg-white lysozyme. *Journal of molecular biology* 82, 371-91.

Diaz A a, Tomba E, Lennarson R, Richard R, Bagajewicz MJ, Harrison RG. (2010): Prediction of protein solubility in Escherichia coli using logistic regression. *Biotechnology and bioengineering* 105, 374-83.

Dill KA, Chan HS. (1997): From Levinthal to pathways to funnels. *Nature structural biology* 4, 10-9.

Dill KA, Ozkan SB, Shell MS, Weikl TR. (2008): The protein folding problem. *Annual review of biophysics* 37, 289-316.

Doglia SM, Ami D, Natalello A, Gatti-Lafranconi P, Lotti M. (2008): Fourier transform infrared spectroscopy analysis of the conformational quality of recombinant proteins within inclusion bodies. *Biotechnology journal* 3, 193-201.

Ennis, Layne. (1957): Spectrophotometric and turbidimetric methods for measuring proteins. *Methods in enzymology* 3, 447-454.

Fahey EM, Chaudhuri JB, Binding P. (2000): Refolding and purification of a urokinase plasminogen activator fragment by chromatography. *Journal of chromatography. B, Biomedical sciences and applications* 737, 225-35.

Fairbanks G, Steck TL, Wallach DFH. (1971): Electrophoretic analysis of the major polypeptides of the human erythrocyte membrane. *Biochemistry* 10, 2606-2617.

Fischer B, Sumner I, Goodenough P. (1993): Isolation, renaturation, and formation of disulfide bonds of eukaryotic proteins expressed in Escherichia coli as inclusion bodies. *Biotechnology and bioengineering* 41, 3-13.

Fisher R. (1971): The Design of Experiments 9th ed. Macmillan.

Fling SP, Gregerson DS. (1986): Peptide and protein molecular weight determination by electrophoresis using a high-molarity tris buffer system without urea. *Analytical biochemistry* 155, 83-8.

Franco-Lara E, Link H, Weuster-Botz D. (2006): Evaluation of artificial neural networks for modelling and optimization of medium composition with a genetic algorithm: From Biochemical Engineering to Systems Biology. *Process Biochemistry* 41, 2200-2206.

Freund Y, Schapire RE. (1997): A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,. *Journal of Computer and System Sciences* 55, 119-139.

Freydell EJ, van der Wielen LAM, Eppink MHM, Ottens M. (2011): Techno-economic evaluation of an inclusion body solubilization and recombinant protein refolding process. *Biotechnology progress* 27, 1315-28.

Galat A. (1982): Statistics of disulfide bond formation in proteins. *The International journal of biochemistry* 14, 363-5.

Gekko K, Timasheff SN. (1981): Mechanism of protein stabilization by glycerol: preferential hydration in glycerol-water mixtures. *Biochemistry* 20, 4667-76.

Gerngross TU. (2004): Advances in the production of human therapeutic proteins in yeasts and filamentous fungi. *Nature biotechnology* 22, 1409-14.

Gobin OC, Schüth F. (2008): On the suitability of different representations of solid catalysts for combinatorial library design by genetic algorithms. *Journal of combinatorial chemistry* 10, 835-46.

Gobin OC, Joaristi AM, Schüth F. (2007): Multi-objective optimization in combinatorial chemistry applied to the selective catalytic reduction of NO with C3H6. *Journal of Catalysis* 252, 205-214.

Graumann K, Premstaller A. (2006): Manufacturing of recombinant therapeutic proteins in microbial systems. *Biotechnology journal* 1, 164-86.

Grosan C, Abraham A. (2007): Hybrid Evolutionary Algorithms: Methodologies, Architectures, and Reviews. In: . *Hybrid Evolutionary Algorithms*. Springer-Verlag GmbH, Vol. 17, pp. 1-17.

Hacker DL, De Jesus M, Wurm FM. (2009): 25 years of recombinant proteins from reactor-grown cells - where do we go from here? *Biotechnology advances* 27, 1023-7.

Hamada H, Arakawa T, Shiraki K. (2009): Effect of additives on protein aggregation. *Current pharmaceutical biotechnology* 10, 400-7.

Hartl FU, Hayer-Hartl M. (2002): Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science (New York, N.Y.)* 295, 1852-8.

Havel J, Link H, Hofinger M, Franco-Lara E, Weuster-Botz D. (2006): Comparison of genetic algorithms for experimental multi-objective optimization on the example of medium design for cyanobacteria. *Biotechnology journal* 1, 549-55.

Heiring C, Muller Y a. (2001): Folding screening assayed by proteolysis: application to various cystine deletion mutants of vascular endothelial growth factor. *Protein engineering* 14, 183-8.

Hevehan DL, De Bernardez Clark E. (1997): Oxidative renaturation of lysozyme at high concentrations. *Biotechnology and bioengineering* 54, 221-30.

Ho JGS, Middelberg APJ. (2004): Estimating the potential refolding yield of recombinant proteins expressed as inclusion bodies. *Biotechnology and bioengineering* 87, 584-92.

Hochuli E, Döbeli H, Schacher A. (1987): New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. *Journal of chromatography* 411, 177-84.

Hofmann A, Tai M, Wong W, Glabe CG. (1995): A sparse matrix screen to establish initial conditions for protein renaturation. *Analytical biochemistry* 230, 8-15.

Hoover DM, Lubkowski J. (2002): DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic acids research* 30, e43.

Ishibashi M, Ida K, Tatsuda S, Arakawa T, Tokunaga M. (2011): Interaction of hexa-His tag with acidic amino acids results in facilitated refolding of halophilic nucleoside diphosphate kinase. *International journal of biological macromolecules* 49, 778-83.

Jaeger KE, Reetz MT. (1998): Microbial lipases form versatile tools for biotechnology. *Trends in biotechnology* 16, 396-403.

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E. (2009): Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC bioinformatics* 10, 136.

De Jesus M, Wurm FM. (2011): Manufacturing recombinant proteins in kg-ton quantities using animal cells in bioreactors. *European journal of pharmaceutics and biopharmaceutics : official journal of Arbeitsgemeinschaft für Pharmazeutische Verfahrenstechnik e.V* 78, 184-8.

Jevsevar S, Gaberc-Porekar V, Fonda I, Podobnik B, Grdadolnik J, Menart V. (2005): Production of nonclassical inclusion bodies from which correctly folded protein can be extracted. *Biotechnology progress* 21, 632-9.

Jin K, Thomas OR, Dunnill P. (1994): Monitoring recombinant inclusion body recovery in an industrial disc stack centrifuge. *Biotechnology and bioengineering* 43, 455-60.

Jungbauer A, Kaar W. (2007): Current status of technical protein refolding. *Journal of biotechnology* 128, 587-96.

Jungbauer A, Kaar W, Schlegl R. (2004): Folding and refolding of proteins in chromatographic beds. *Current opinion in biotechnology* 15, 487-94.

Katoh S, Katoh Y. (2000): Continuous refolding of lysozyme with fed-batch addition of denatured protein solution. *Process Biochemistry* 35, 1119 - 1124.

Kiefhaber T, Rudolph R, Kohler HH, Buchner J. (1991): Protein aggregation in vitro and in vivo: a quantitative model of the kinetic competition between folding and aggregation. *Bio/technology (Nature Publishing Company)* 9, 825-9.

Kim CS, Lee EK. (2000): Effects of operating parameters in in vitro renaturation of a fusion protein of human growth hormone and glutathione S transferase from inclusion body. *Process Biochemistry* 36, 111-117.

Koolaee SMV, Shojaosadati SA, Babaeipour V, Ghaemi N. (2006): Physiological and morphological changes of recombinant γ E . coli during over-expression of human interferon-γ in HCDC. *Iranian Journal of Biotechnology* 4, 230-238.

Kunz W, Henle J, Ninham BW. (2004): "Zur Lehre von der Wirkung der Salze" (about the science of the effect of salts): Franz Hofmeister's historical papers. *Current Opinion in Colloid & Interface Science* 9, 19-37.

Kweon D-H, Lee D-H, Han N-S, Seo J-H. (2004): Solid-phase refolding of cyclodextrin glycosyltransferase adsorbed on cation-exchange resin. *Biotechnology progress* 20, 277-83.

Laemmli UK. (1970): Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 227, 680-5.

Lange C, Patil G, Rudolph R. (2005): Ionic liquids as refolding additives: N'-alkyl and N'-(omega-hydroxyalkyl) N-methylimidazolium chlorides. *Protein science : a publication of the Protein Society* 14, 2693-701.

Lasky FD, Boser RB. (1997): Designing in quality through design control: a manufacturer's perspective. *Clinical chemistry* 43, 866-72.

Lee GH, Cooney D, Middelberg a PJ, Choe WS. (2006): The economics of inclusion body processing. *Bioprocess and biosystems engineering* 29, 73-90.

Levinthal C. (1969): How to fold graciously. In: DeBrunner, JTP, Munck, E, editors. *Mossbauer Spectroscopy in Biological Systems Proceedings of a meeting held at Allerton House*. University of Illinois Press, Vol. 24, pp. 22-24.

Li M, Poliakov A, Danielson UH, Su Z, Janson J-C. (2003): Refolding of a recombinant full-length non-structural (NS3) protein from hepatitis C virus by chromatographic procedures. *Biotechnology letters* 25, 1729-34.

Lilie H, Schwarz E, Rudolph R. (1998): Advances in refolding of proteins produced in E. coli. *Current opinion in biotechnology* 9, 497-501.

Lin L, Seehra J, Stahl ML. (2006): High-throughput identification of refolding conditions for LXRbeta without a functional assay. *Protein expression and purification* 47, 355-66.

Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. (2011): How Fast-Folding Proteins Fold. *Science* 334, 517-520.

Link H, Weuster-Botz D. (2006): Genetic algorithm for multi-objective experimental optimization. *Bioprocess and biosystems engineering* 29, 385-90.

Lionberger R a, Lee SL, Lee L, Raw A, Yu LX. (2008): Quality by design: concepts for ANDAs. *The AAPS journal* 10, 268-76.

Liu D, Schmid RD, Rusnak M. (2006): Functional expression of Candida antarctica lipase B in the Escherichia coli cytoplasm--a screening system for a frequently used biocatalyst. *Applied microbiology and biotechnology* 72, 1024-32.

Liu H-L. (2007): Recent Advances in Disulfide Connectivity Predictions. *Current Bioinformatics* 2, 31-47.

Lunin VV, Li Y, Schrag JD, Iannuzzi P, Cygler M, Matte A. (2004): Crystal Structures of Escherichia coli ATP-Dependent Glucokinase and Its Complex with Glucose. *Structure* 186, 6915-6927.

Magnan CN, Randall A, Baldi P. (2009): SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics (Oxford, England)* 25, 2200-7.

Makino T, Skretas G, Georgiou G. (2011): Strain engineering for improved expression of recombinant proteins in bacteria. *Microbial cell factories* 10, 32.

Mannall GJ, Myers JP, Liddell J, Titchener-Hooker NJ, Dalby PA. (2009): Ultra scale-down of protein refold screening in microwells: challenges, solutions and application. *Biotechnology and bioengineering* 103, 329-40.

de Marco A, Deuerling E, Mogk A, Tomoyasu T, Bukau B. (2007): Chaperone-based procedure to increase yields of soluble recombinant proteins produced in E. coli. *BMC biotechnology* 7, 32.

Mattingly JR, Iriarte A, Martinez-Carrion M. (1995): Homologous Proteins with Different Affinities for groEL. *Journal of Biological Chemistry* 270, 1138-1148.

Mavis D, Stellwagen E. (1968): Purification and Subunit Structure of Glutathione Reductase from Bakers ' Yeast *. *Biological Chemistry* 4, 809-814.

Mayer M, Buchner J. (2004): Refolding of inclusion body proteins. *Methods in molecular medicine* 94, 239-54.

Meireles MRG, Almeida PEM, Simoes MG. (2003): A comprehensive review for industrial applicability of artificial neural networks. *IEEE Transactions on Industrial Electronics* 50, 585-601.

Meyer D, Schneider-Fresenius C, Horlacher R, Peist R, Boos W. (1997): Molecular characterization of glucokinase from Escherichia coli K-12. *Journal of bacteriology* 179, 1298-306.

Michalewicz Z. (1999): Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag.

Middelberg APJ. (2002): Preparative protein refolding. *Trends in biotechnology* 20, 437-43.

Montgomery DC. (2009): Design and Analysis of Experiments. John Wiley & Sons.

Mountain A, Rehm H-J, Ney U, Schomburg D. (1999): Biotechnology, Recombinant Proteins, Monoclonal Antibodies, and Therapeutic Genes. Ed. Dietmar Schomburg A. Mountain, Hans-Jürgen Rehm, U. Ney. *Biotechnology and bioengineering* 2nd ed. Weinheim, Germany: Wiley-VCH. Vol. 109.

Nelder BJA, Meadf R, Nelder JA, Mead R. (1965): A Simplex Method for Function Minimization. *The Computer Journal* 7, 308-313.

Nordhoff A, Tziatzios C, van Den Broek JA, Schott MK, Kalbitzer HR, Becker K, Schubert D, Schirmer RH. (1997): Denaturation and reactivation of dimeric human glutathione reductase--an assay for folding inhibitors. *European journal of biochemistry / FEBS* 245, 273-82.

Nørby JG. (1988): Coupled assay of Na+,K+-ATPase activity. *Methods in enzymology* 156, 116-9.

Oberg K, Chrunyk BA, Wetzel R, Fink AL. (1994): Nativelike secondary structure in interleukin-1 beta inclusion bodies by attenuated total reflectance FTIR. *Biochemistry* 33, 2628-34.

Onuchic JN, Wolynes PG. (2004): Theory of protein folding. *Current Opinion in Structural Biology* 14, 70-75.

Ormö M, Cubitt a B, Kallio K, Gross L a, Tsien RY, Remington SJ. (1996): Crystal structure of the Aequorea victoria green fluorescent protein. *Science (New York, N.Y.)* 273, 1392-5.

O'Brien EP, Dima RI, Brooks B, Thirumalai D. (2007): Interactions between hydrophobic and ionic solutes in aqueous guanidinium chloride and urea solutions: lessons for protein denaturation mechanism. *Journal of the American Chemical Society* 129, 7346-53.

Patnaik PR. (1998): Bioseparation and Bioprocessing. Ed. G. Subramanian. Weinheim, Germany: Wiley-VCH Verlag GmbH.

Phan J, Yamout N, Schmidberger J, Bottomley SP, Buckle AM. (2011): Refolding your protein with a little help from REFOLD. Ed. Andrew F. Hill, Kevin J. Barnham, Stephen P. Bottomley, Roberto Cappai. *Methods in molecular biology (Clifton, N.J.)* 752, 45-57.

Pineda JRET, Callender R, Schwartz SD. (2007): Ligand binding and protein dynamics in lactate dehydrogenase. *Biophysical journal* 93, 1474-83.

Plackett RL, Burman JP. (1946): The design of optimum multifactorial experiments. *Biometrika* 33, 305-325.

Polikar R. (2006): Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6, 21-45.

Qoronfleh MW, Hesterberg LK, Seefeldt MB. (2007): Confronting high-throughput protein refolding using high pressure and solution screens. *Protein expression and purification* 55, 209-24.

Rabhi-Essafi I, Sadok A, Khalaf N, Fathallah DM. (2007): A strategy for high-level expression of soluble and functional human interferon alpha as a GST-fusion protein in E. coli. *Protein engineering, design & selection : PEDS* 20, 201-9.

Razi M, Athappilly K. (2005): A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications* 29, 65-74.

Read J a, Winter VJ, Eszes CM, Sessions RB, Brady RL. (2001): Structural basis for altered activity of M- and H-isozyme forms of human lactate dehydrogenase. *Proteins* 43, 175-85.

Reeves CR. (1993): Using genetic algorithms with small populations sizes. *Proceedings of the Fifth International Conference on Genetic Algorithms*, 92-99.

Roberts MJ, Bentley MD, Harris JM. (2002): Chemistry for peptide and protein PEGylation. *Advanced drug delivery reviews* 54, 459-76.

Royer C a. (2006): Probing protein folding and conformational transitions with fluorescence. *Chemical reviews* 106, 1769-84.

Rudolph G. (1994): Convergence analysis of canonical genetic algorithms. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 5, 96-101.

Rudolph R, Heider I, Westhof E, Jaenicke R. (1977): Mechanism of refolding and reactivation of lactic dehydrogenase from pig heart after dissociation in various solvent media. *Biochemistry* 16, 3384-90.

Rudolph R, Zettlmeissl G, Jaenicke R. (1979): Reconstitution of lactic dehydrogenase. Noncovalent aggregation vs. reactivation. 2. Reactivation of irreversibly denatured aggregates. *Biochemistry* 18, 5572-5.

Rudolph R, Lilie H. (1996): In vitro folding of inclusion body proteins. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 10, 49-56.

Rumelhart DE, Hinton GE, Williams RJ. (1986): Learning representations by back-propagating errors. *Nature* 323, 533-536.

Rypniewski WR, Holden HM, Rayment I. (1993): Structural consequences of reductive methylation of lysine residues in hen egg white lysozyme: an X-ray analysis at 1.8-A resolution. *Biochemistry* 32, 9851-8.

Sakamoto R, Nishikori S, Shiraki K. (2004): High temperature increases the refolding yield of reduced lysozyme: implication for the productive process for folding. *Biotechnology progress* 20, 1128-33.

Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjölander K. (2010): Active site prediction using evolutionary and structural information. *Bioinformatics (Oxford, England)* 26, 617-24.

Sass C, Briand M, Benslimane S, Renaud M, Briand Y. (1989): Characterization of rabbit lactate dehydrogenase-M and lactate dehydrogenase-H cDNAs. Control of lactate dehydrogenase expression in rabbit muscle. *The Journal of biological chemistry* 264, 4076-81.

Scheich C, Niesen FH, Seckler R, Büssow K. (2004): An automated in vitro protein folding screen applied to a human dynactin subunit. *Protein science : a publication of the Protein Society* 13, 370-80.

Schmidt FR. (2004): Recombinant expression systems in the pharmaceutical industry. *Applied microbiology and biotechnology* 65, 363-72.

Schmitt L. (2001): Theory of genetic algorithms. *Theoretical Computer Science* 259, 1-61.

Schmoeger E, Wellhoefer M, Dürauer A, Jungbauer A, Hahn R. (2010): Matrix-assisted refolding of autoprotease fusion proteins on an ion exchange column: a kinetic investigation. *Journal of chromatography. A* 1217, 5950-6.

Schwarz E, Lilie H, Rudolph R. (1996): The effect of molecular chaperones on in vivo and in vitro folding processes. *Biological chemistry* 377, 411-6.

Sela M, White FH, Anfinsen CB. (1957): Reductive cleavage of disulfide bridges in ribonuclease. *Science (New York, N.Y.)* 125, 691-2.

Shukla D, Trout BL. (2010): Interaction of arginine with proteins and the mechanism by which it inhibits aggregation. *The journal of physical chemistry. B* 114, 13426-38.

Singh SM, Panda AK. (2005): Solubilization and refolding of bacterial inclusion body proteins. *Journal of bioscience and bioengineering* 99, 303-10.

Smith PK, Krohn RI, Hermanson GT, Mallia AK, Gartner FH, Provenzano MD, Fujimoto EK, Goeke NM, Olson BJ, Klenk DC. (1985): Measurement of protein using bicinchoninic acid. *Analytical biochemistry* 150, 76-85.

Sosnick TR, Hinshaw JR. (2011): How Proteins Fold. *Science* 334, 464-465.

Stambaugh R, Post D. (1966): Substrate and product inhibition of rabbit muscle lactic dehydrogenase heart (H4) and muscle (M4) isozymes. *The Journal of biological chemistry* 241, 1462-7.

Stellwagen E, Prantner JD, Stellwagen NC. (2008): Do zwitterions contribute to the ionic strength of a solution? *Analytical biochemistry* 373, 407-9.

Sun F, Lin DKJ, Liu M-Q. (2011): On construction of optimal mixed-level supersaturated designs. *The Annals of Statistics* 39, 1310-1333.

Takagi T, Tsujii K, Shirahama K. (1975): Binding isotherms of sodium dodecyl sulfate to protein polypeptides with special reference to SDS-polyacylamide gel electrophoresis. *Journal of biochemistry* 77, 939-47.

Tandon S, Horowitz PM. (1987): Detergent-assisted refolding of guanidinium chloride-denatured rhodanese. The effects of the concentration and type of detergent. *The Journal of biological chemistry* 262, 4486-91.

Tandoğan B, Ulusu NN. (2007): The inhibition kinetics of yeast glutathione reductase by some metal ions. *Journal of enzyme inhibition and medicinal chemistry* 22, 489-95.

Taylor G, Hoare M, Gray DR, Marston FAO. (1986): Size and Density of Protein Inclusion Bodies. *Nature biotechnology* 4, 553-557.

The Uniprot Consortium. (2012): Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research* 40, D71-5.

Tibshirani R. (1996): Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58, 267-288.

Timasheff SN. (1998): Control of protein stability and reactions by weakly interacting cosolvents: the simplicity of the complicated. *Advances in protein chemistry* 51, 355-432.

Timasheff SN. (2002): Protein-solvent preferential interactions, protein hydration, and the modulation of biochemical reactions by solvent components. *Proceedings of the National Academy of Sciences of the United States of America* 99, 9721-6.

Tobbell DA, Middleton BJ, Raines S, Needham MRC, Taylor IWF, Beveridge JY, Abbott WM. (2002): Identification of in vitro folding conditions for procathepsin S and cathepsin S using fractional factorial screens. *Protein expression and purification* 24, 242-54.

Topell S, Hennecke J, Glockshuber R. (1999): Circularly permuted variants of the green fluorescent protein. *FEBS letters* 457, 283-9.

Trésaugues L, Collinet B, Minard P, Henckes G, Aufrère R, Blondeau K, Liger D, Zhou C-Z, Janin J, Van Tilbeurgh H, Quevillon-Cheruel S. (2004): Refolding strategies from inclusion bodies in a structural genomics project. *Journal of structural and functional genomics* 5, 195-204.

Tsumoto K, Umetsu M, Yamada H, Ito T, Misawa S, Kumagai I. (2003a): Immobilized oxidoreductase as an additive for refolding inclusion bodies: application to antibody fragments. *Protein Engineering Design and Selection* 16, 535-541.

Tsumoto K, Arakawa T, Chen L. (2010): Step-wise refolding of recombinant proteins. *Current pharmaceutical biotechnology* 11, 285-8.

Tsumoto K, Ejima D, Kumagai I, Arakawa T. (2003b): Practical considerations in refolding proteins from inclusion bodies. *Protein expression and purification* 28, 1-8.

Tsumoto K, Umetsu M, Kumagai I, Ejima D, Philo JS, Arakawa T. (2004a): Role of arginine in protein refolding, solubilization, and purification. *Biotechnology progress* 20, 1301-8.

Tsumoto K, Umetsu M, Kumagai I, Ejima D, Philo JS, Arakawa T. (2004b): Role of arginine in protein refolding, solubilization, and purification. *Biotechnology progress* 20, 1301-8.

Vagenende V, Yap MGS, Trout BL. (2009): Mechanisms of protein stabilization and prevention of protein aggregation by glycerol. *Biochemistry* 48, 11084-96.

Valax P, Georgiou G. (1993): Molecular characterization of beta-lactamase inclusion bodies produced in Escherichia coli. 1. Composition. *Biotechnology progress* 9, 539-47.

Vallejo LF, Rinas U. (2004): Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microbial cell factories* 3, 11.

Van Veldhuizen D a, Lamont GB. (2000): Multiobjective evolutionary algorithms: analyzing the state-of-the-art. *Evolutionary computation* 8, 125-47.

Ventura S, Villaverde A. (2006): Protein quality in bacterial inclusion bodies. *Trends in biotechnology* 24, 179-85.

Vincentelli R, Canaan S, Campanacci V, Valencia C, Maurin D, Frassinetti F, Scappucini-Calvo L, Bourne Y, Cambillau C, Bignon C. (2004): High-throughput automated refolding screening of inclusion bodies. *Protein science : a publication of the Protein Society* 13, 2782-92.

Voet D, Voet J. (2004): Biochemistry 3rd ed. Hoboken, NJ.: Wiley.

Vuillard L, Rabilloud T, Goldberg ME. (1998): Interactions of non-detergent sulfobetaines with early folding intermediates facilitate in vitro protein renaturation. *European journal of biochemistry / FEBS* 256, 128-35.

Wacker M, Linton D, Hitchen PG, Nita-Lazar M, Haslam SM, North SJ, Panico M, Morris HR, Dell A, Wren BW, Aebi M. (2002): N-linked glycosylation in Campylobacter jejuni and its functional transfer into E. coli. *Science (New York, N.Y.)* 298, 1790-3.

Wang W. (2005): Protein aggregation and its inhibition in biopharmaceutics. *International journal of pharmaceutics* 289, 1-30.

Wang X-T, Engel PC. (2009): An optimised system for refolding of human glucose 6-phosphate dehydrogenase. *BMC biotechnology* 9, 19.

Werner MH, Clore GM, Gronenborn a M, Kondoh a, Fisher RJ. (1994): Refolding proteins by gel filtration chromatography. *FEBS letters* 345, 125-30.

West SM, Chaudhuri JB, Howell J a. (1998): Improved protein refolding using hollow-fibre membrane dialysis. *Biotechnology and bioengineering* 57, 590-9.

Wetlaufer DB, Xie Y. (1995): Control of aggregation in protein refolding: a variety of surfactants promote renaturation of carbonic anhydrase II. *Protein science : a publication of the Protein Society* 4, 1535-43.

Weuster-Botz D. (2000): Experimental design for fermentation media development: statistical design or global random search? *Journal of bioscience and bioengineering* 90, 473-483.

Wilkins MR, Gasteiger E, Bairoch a, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF. (1999): Protein identification and analysis tools in the ExPASy server. *Methods in molecular biology (Clifton, N.J.)* 112, 531-52.

Willis MS, Hogan JK, Prabhakar P, Liu X, Tsai K, Wei Y, Fox T. (2005): Investigation of protein refolding using a fractional factorial screen: a study of reagent effects and interactions. *Protein science : a publication of the Protein Society* 14, 1818-26.

Wolf D, Buyevskaya OV, Baerns M. (2000): An evolutionary approach in the combinatorial selection and optimization of catalytic materials. *Applied Catalysis A: General* 200, 63-77.

Xie Y, Wetlaufer DB. (1996): Control of aggregation in protein refolding: the temperature-leap tactic. *Protein science : a publication of the Protein Society* 5, 517-523.

Xu X, Kashima O, Saito A, Azakami H, Kato A. (2004): Structural and functional properties of chicken lysozyme fused serine-rich heptapeptides at the C-terminus. *Bioscience, biotechnology, and biochemistry* 68, 1273-8.

Yasuda M, Murakami Y, Sowa A, Ogino H, Ishikawa H. (1998): Effect of additives on refolding of a denatured protein. *Biotechnology progress* 14, 601-6.

Yu J, Zhou C-zhao. (2007): Crystal structure of glutathione reductase Glr1 from the yeast Saccharomyces cerevisiae. *Proteins*, 972-979.

Zhang T, Xu X, Shen L, Feng Y, Yang Z, Shen Y, Wang J, Jin W, Wang X. (2009): Modeling of protein refolding from inclusion bodies. *Acta biochimica et biophysica Sinica* 41, 1044-52.

Zheng Y, Guo S, Guo Z, Wang X. (2004): Effects of N-terminal deletion mutation on rabbit muscle lactate dehydrogenase. *Biochemistry. Biokhimiiâ* 69, 401-6.

Zitzler E. (1999): Evolutionary Algorithms for Multiobjective Optimization : Methods and Applications. Shaker Verlag.

Zitzler E, Laumanns M, Thiele L. (2002): SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization. *Evolutionary Methods for Design, Optimisation and Control*, 95-100.

9 Appendix

9.1 Abbreviations

ATP	adenosine-triphosphate
BDT	bagged decision trees
BRIJ 35	polyethylene glycol dodecyl ether
CD	circular dichroism spectroscopy
CDW	cell dry weight
CHAPS	cholamidopropyl-dimethylammonium-propanesulfonate
CHO	chinese hamster ovary cells
CMC	critical micellar concentration
Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2+}	mineral ions supplemented as sulfates
ddH_2O	bidistilled water
DOE	design of experiments
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
FAD	flavin adenine dinucleotide
GA	genetic algorithm
Gdn·HCl	guanidine hydrochloride
GEN	generation or iteration of the optimization
GFP	green fluorescent protein
GLK	glucokinase
GLR	glutathione reductase
GSH	reduced L-glutathione
GSSG	oxidized L-glutathione

GUI	graphical user interface
HEPES	hydroxylethyl-piperazine-ethanesulfonic acid
HPLC	high-performance liquid chromatography
IEC	ion exchange chromatography
IMAC	immobilized metal affinity chromatography
LB	Luria broth medium
LDH	lactate dehydrogenase
LYZ	lysozyme
MOPS	morpholino-propanesulfonic acid
MSE	mean square error
NADH	nicotinamide adenine dinucleotide
NADPH	nicotinamide adenine dinucleotide phosphate
NDSB 201	non-detergent sulfobetaine 201
NZY	NZY media
PB	sodium phosphate buffer
PCR	polymerase chain reaction
PEG (4000)	polyethylene glycol (4000)
pI	isoelectric point
RSM	response surface methodology
RT	room temperature
SDC	deoxycholic acid sodium salt
SDS	sodium dodecyl sulfate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	size exclusion chromatography
SPEA	strength pareto evolutionary algorithm
TAE	buffer system with TRIS, acetic acid and EDTA
TCEP	tris-carboxyethyl-phosphine

TRIS	tris(hydroxymethyl)aminomethane
TRITON-X 100	polyethylene glycol tert-octylphenyl ether
TWEEN 20	polyethylene glycol sorbitan-monolaurate
ZWITTERGENT 3-12	dodecyldimethyl-ammonio-propanesulfonate

9.2 Symbols and variables

b_0	zero order coefficients	-
b_i	first order coefficients	-
$b_{i\cdot i}$	interaction coefficients	-
$b_{i\cdot j}$	second order coefficients	-
costs	experimental costs of the refolding buffer	€ mL^{-1}
I	ionic strength	mol L^{-1}
native activity	specific activity of the native protein	(-, U mg^{-1}, U g^{-1}, s^{-1})
$OPT_{I–V}$	LIP optimization	
R^2	correlation coefficient	-
refolded activity	specific activity of the refolded protein	(-, U mg^{-1}, U g^{-1}, s^{-1})
refolding yield	relative refolding yield	
θ_{MRW}	molar ellipticity	deg cm^2 $dmol^{-1}$

9.3 Experimental design matrices

9.3.1 Standard experimental design matrices

Table 9.1: Two-level full factorial design for 3 factors with 8 experiments. These are the experimental data points of the cube illustrated in Figure 3.16.

Experiment	**Factors**		
	A	**B**	**C**
1	−1	−1	−1
2	1	−1	−1
3	−1	1	−1
4	1	1	−1
5	−1	−1	1
6	1	−1	1
7	−1	1	1
8	1	1	1

Table 9.2: Two-level fractional (2^{4-1}) factorial design for 4 factors with 8 experiments**.**

Experiment	**Factors**			
	A	**B**	**C**	**D**
1	−1	−1	−1	−1
2	1	−1	−1	1
3	−1	1	−1	1
4	1	1	−1	−1
5	−1	−1	1	1
6	1	−1	1	−1
7	−1	1	1	−1
8	1	1	1	1

Table 9.3: Plackett-Burman design with 7 factors.

Experiment	**Factors**						
	A	**B**	**C**	**D**	**E**	**F**	**G**
1	1	1	1	−1	1	−1	−1
2	−1	1	1	1	−1	1	−1
3	−1	−1	1	1	1	−1	1
4	1	−1	−1	1	1	1	−1
5	−1	1	−1	−1	1	1	1
6	1	−1	1	−1	−1	1	1
7	1	1	−1	1	−1	−1	1
8	−1	−1	−1	−1	−1	−1	−1

Table 9.4: Central composite design with 3 factors. For practical purposes the centre point is often measured several times, in order to get on error estimation. α defines the design type and varies for circumscribed (1.6818), faced (1) and inscribed (0.5946) designs.

Experiment	**Factors**			**Comment**
	A	**B**	**C**	
1	−1	−1	−1	factorial
2	1	−1	−1	factorial
3	−1	1	−1	factorial
4	1	1	−1	factorial
5	−1	−1	1	factorial
6	1	−1	1	factorial
7	−1	1	1	factorial
8	1	1	1	factorial
9	− α	0	0	star
10	+α	0	0	star
11	0	− α	0	star
12	0	+α	0	star
13	0	0	− α	star
14	0	0	+α	star
15	0	0	0	centre

9.3.2 Experimental design matrices of the statistical DOE

Table 9.5: Encoding of the D-optimal design. Factors (f) were two-level and numerical except for the categorical factor 2 (buffer conditions; 5 levels), 13 (detergents, 8 levels) and 14 (redox substances, 6 levels).

Factor	Levels	decoded, experimental values
f1	2	1: pH 6.0; 2: pH 9.75
f2	4	1: 100 mM PB; 2: 100 mM HEPES, 3: 100 mM MOPS, 4: 1000 mM TRIS·HCl; 5: -
f3	2	1: -; 2: 350 mM NaCl
f4	2	1: -; 2: 80 mM KCl
f5	2	1: -; 2: 15 % v/v glycerol
f6	2	1: -; 2: 0.25 % w/v PEG
f7	2	1: -; 2: 750 mM arginine
f8	2	1: -; 2: 350 mM glycine
f9	2	1: -; 2: 350 mM glutamine
f10	2	1: -; 2: 350 mM glutamate
f11	2	1: -; 2: 0.1 mM Cu^{2+} Zn^{2+} Mg^{2+} Mn^{2+}
f12	2	1: -; 2: 10 mM EDTA
f13	8	1: -; 2: 10.67 mM CHAPS; 3: 1500 mM NDSB 201; 4: 4 mM ZWITTERGENT 3-12; 5: 0.08 mM TWEEN; 6: 0.8 mM TRITON-X 100; 7: 12 mM SDS; 8: 0.12 mM BRIJ 35
f14	4	1: 10 mM DTT; 2: 10 mM TCEP; 3: 5 mM GSH; 4: 5 mM GSSG; 5: 5 mM GSH and 5 mM GSSG; 6: -

Table 9.6: D-optimal design with 30 experiments (E) and 14 factors (f). Factors are two-level except for factor 2 (buffer conditions;5 levels), 13 (detergents, 8 levels) and 14 (redox substances, 6 levels). Native and refolded LIP activity in U g^{-1} (nat and ref) with standard deviation ($std_{nat,ref}$).

E	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	nat	std_{nat}	ref	std_{ref}
1	1	2	1	2	1	1	1	2	1	2	2	1	5	6	420.6	67.8	140.0	41.4
2	1	3	2	2	1	1	1	1	2	2	1	1	6	2	204.1	69.7	208.5	64.3
3	1	5	2	1	2	2	1	1	2	2	1	1	2	6	44.6	4.4	46.3	23.5
4	2	2	2	1	2	1	2	1	1	2	2	2	4	6	134.4	32.8	133.6	40.6
5	1	1	1	2	2	1	2	1	1	1	1	1	3	6	368.5	88.2	75.6	8.4
6	2	3	2	1	2	2	1	1	2	1	2	1	5	3	426.4	97.7	133.2	19.5
7	2	5	1	1	2	1	1	1	2	2	2	2	8	2	27.3	23.4	30.1	15.6
8	2	4	1	1	1	2	2	2	2	1	2	2	6	6	344.9	78.2	194.5	33.2
9	1	2	1	1	2	2	1	2	2	1	1	2	7	2	36.5	9.9	29.0	15.6
10	2	2	1	2	2	1	2	2	2	2	1	1	2	3	81.4	30.3	89.9	30.6
11	1	5	2	2	1	1	2	1	1	1	1	2	7	3	200.2	64.7	92.3	35.0
12	1	2	2	1	1	2	2	1	1	2	1	1	8	4	172.8	65.7	274.5	56.5
13	2	2	2	2	1	1	1	2	2	1	2	2	3	1	240.9	67.2	100.3	6.3
14	2	1	2	2	1	2	2	2	1	1	2	2	2	2	3.6	7.1	4.2	12.1

15	2	5	1	1	1	2	2	2	1	2	1	1	5	5	255.9	55.8	133.3	40.2
16	1	4	2	2	2	1	2	2	2	1	2	1	8	5	415.7	52.0	278.0	60.6
17	1	3	1	1	1	1	2	2	2	1	1	1	4	1	123.9	6.9	200.8	2.9
18	2	4	2	1	2	1	1	2	1	1	1	1	1	2	4.1	5.9	43.7	21.0
19	2	1	1	1	1	1	1	1	2	1	2	1	1	3	239.1	18.0	72.8	18.9
20	2	4	1	2	1	2	1	1	1	2	2	1	7	1	78.8	4.7	170.3	14.2
21	1	4	1	1	2	2	2	2	1	2	2	2	3	3	320.3	102.8	235.0	62.9
22	2	4	1	2	2	1	2	1	2	1	1	2	5	4	354.5	51.2	183.2	10.3
23	1	3	1	1	1	1	1	1	1	1	2	2	2	5	97.5	5.5	173.1	15.3
24	1	5	1	2	2	2	1	2	1	1	2	1	4	4	413.3	59.9	453.2	91.0
25	2	2	1	2	2	2	2	1	1	1	2	1	6	5	425.3	59.2	275.8	23.8
26	2	3	2	1	2	1	2	2	2	2	2	1	7	4	99.9	8.4	147.3	20.1
27	2	1	2	2	1	2	1	2	2	2	1	2	4	5	265.2	57.4	206.4	16.2
28	1	1	2	1	2	1	1	2	1	2	1	2	6	1	119.0	63.8	140.0	38.9
29	1	3	1	2	2	2	2	1	2	2	2	2	1	1	59.8	20.1	58.6	11.4
30	2	3	1	2	2	2	1	2	1	1	1	2	8	6	283.5	37.7	144.5	58.3

Table 9.7: D-optimal design matrix for regression analysis. E experiments 1 to 30, 1 constant term, 2 to 28 linear terms.

E	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	1	1	0	1	0	0	1	0	1	1	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
2	1	1	0	0	1	0	0	0	1	1	1	1	0	0	1	1	0	0	0	0	0	1	0	0	1	0	0	0
3	1	1	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0
4	1	0	0	1	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
5	1	1	1	0	0	0	1	0	0	1	0	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0
6	1	0	0	0	1	0	0	1	0	0	1	1	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0
7	1	0	0	0	0	0	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
8	1	0	0	0	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
9	1	1	0	1	0	0	1	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0	0
10	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0
11	1	1	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0
12	1	1	0	1	0	0	0	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0
13	1	0	0	1	0	0	0	0	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0
14	1	0	1	0	0	0	0	0	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0
15	1	0	0	0	0	0	1	1	1	0	0	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1

16	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1
17	1	1	0	0	1	0	1	1	1	1	0	0	0	1	1	1	0	0	0	1	0	0	0	1	0	0	0	0
18	1	0	0	0	0	1	0	1	0	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0
19	1	0	1	0	0	0	1	1	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	1	0	0
20	1	0	0	0	0	1	1	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0
21	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
22	1	0	0	0	0	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0
23	1	1	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1
24	1	1	0	0	0	0	1	0	0	0	1	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0
25	1	0	0	1	0	0	1	0	0	0	0	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1
26	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0
27	1	0	1	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
28	1	1	1	0	0	0	0	1	0	1	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0
29	1	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
30	1	0	0	0	1	0	1	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 9.8: RSM, circumscribed central composite design. Experiments with coded variables (V_1 to V_5) and the experimental results for LIP refolding. Native and refolded LIP activity in U g^{-1} (nat and ref) with standard deviation ($std_{nat,ref}$).

	V_1	V_2	V_4	V_4	V_5	nat	std_{nat}	ref	std_{ref}
1	−1	−1	−1	−1	1	558.55	92.31	373.54	69.43
2	−1	−1	−1	1	−1	501.39	74.51	331.41	81.10
3	−1	−1	1	−1	−1	598.25	56.08	296.21	71.99
4	−1	−1	1	1	1	546.31	70.07	481.38	61.90
5	−1	1	−1	−1	−1	439.73	80.80	236.64	43.37
6	−1	1	−1	1	1	439.08	35.59	491.20	65.22
7	−1	1	1	−1	1	479.99	84.22	504.91	85.72
8	−1	1	1	1	−1	442.50	81.35	323.36	101.44
9	1	−1	−1	−1	−1	468.46	45.45	362.14	111.05
10	1	−1	−1	1	1	470.72	39.81	378.58	95.72
11	1	−1	1	−1	1	431.81	29.39	413.14	57.48
12	1	−1	1	1	−1	511.84	62.19	456.81	74.20
13	1	1	−1	−1	1	394.15	18.21	397.49	49.66
14	1	1	−1	1	−1	434.98	86.42	398.27	97,77

15	1	1	1	−1	−1	464.83	50.16	387.19	40.95
16	1	1	1	1	1	528.13	80.66	508.45	95.81
17	−2	0	0	0	0	662.56	86.83	440.79	110.29
18	2	0	0	0	0	440.67	48.67	476.87	74.52
19	0	−2	0	0	0	559.89	84.73	405.36	47.77
20	0	2	0	0	0	400.01	33.42	418.51	54.87
21	0	0	−2	0	0	483.73	21.24	301.48	79.75
22	0	0	2	0	0	556.82	48.09	412.66	86.32
23	0	0	0	−2	0	391.91	74.18	362.34	76.63
24	0	0	0	2	0	492.70	62.20	488.68	77.83
25	0	0	0	0	−2	367.55	56.32	229.94	42.74
26	0	0	0	0	2	377.95	80.27	392.46	90.33
27	0	0	0	0	0	347.48	52.83	363.83	46.33

9.4 Expression of the Lipase from *Thermomyces lanuginosus*

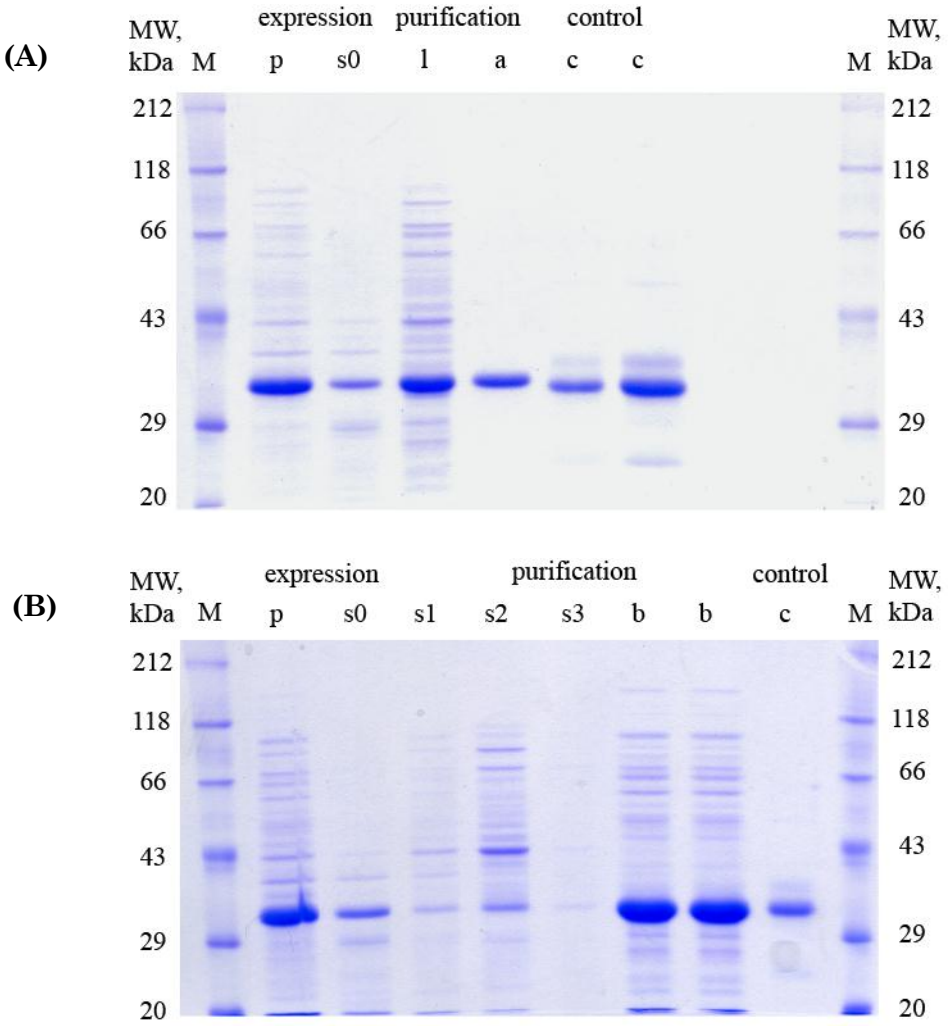


Figure 9.1: Purification of LIP. (A) with His-tag, (B) without His-tag. (p) pellet, (s_i) solube fractions, (a) purified His-tagged protein, (b) purified wild type, (c) control purchased by Sigma-Aldrich, (M) marker, 14 kDa to 212 kDa (Carl Roth).

Table 9.9: DNA oligomers for gene synthesis of LIP.

Name	Sequence	Length
d1	GATACACATATGGAAGTTTCCCAGGACCTGTTC	33
d2	CGCGAACAGATTGAACTGGTTGAACAGGTCCTGGGAAACTTC	42
d3	ACCAGTTCAATCTGTTCGCGCAATATTCTGCAGCCGCTTATT	42
d4	TGCGTCGTTGTTTTTACCACAATAAGCGGCTGCAGAATATTG	42
d5	GTGGTAAAAACAACGACGCACCAGCCGGCACGAACATTACCT	42
d6	CTTCCGGACAGGCGTTGCCCGTGCAGGTAATGTTCGTGCCGG	42
d7	CAACGCCTGTCCGGAAGTTGAAAAAGCGGACGCGACCTTCCT	42
d8	CGACACCGCTGTCCTCAAAAGAGTACAGGAAGGTCGCGTCCG	42
d9	TGAGGACAGCGGTGTCGGCGACGTTACTGGTTTCCTGGCGCT	42
d10	AACGATCAGTTTGTTGGTGTTGTCGAGCGCCAGGAAACCAGT	42
d11	CAACACCAACAAACTGATCGTTCTCTCTTTCCGTGGCTCTCG	42
d12	TTACCGATCCAATTCTCGATGGAACGAGAGCCACGGAAAGAG	42
d13	CCATCGAGAATTGGATCGGTAACCTGAACTTCGACCTGAAGG	42
d14	CCAGAGCAGATATCGTTGATCTCCTTCAGGTCGAAGTTCAGG	42
d15	AGATCAACGATATCTGCTCTGGTTGCCGTGGTCACGACGGTT	42
d16	GCTACAGAACGCCAAGAAGAGGTGAAACCGTCGTGACCACGG	42
d17	TCTTCTTGGCGTTCTGTAGCGGACACGCTGCGTCAGAAAGTA	42
d18	GGGTGTTCACGAACCGCGTCCTCTACTTTCTGACGCAGCGTG	42
d19	GCGGTTCGTGAACACCCGGACTATCGTGTAGTATTCACCGGT	42
d20	AGTGCACCACCGAGAGAGTGACCGGTGAATACTACACGATAG	42
d21	CTCTCTCGGTGGTGCACTCGCCACCGTTGCGGGTGCGGATCT	42
d22	CGTCAATGTCGTAGCCATTGCCGCGGAGATCCGCACCCGCAA	42
d23	CAATGGCTACGACATTGACGTTTTCTCTTACGGTGCGCCTCG	42
d24	CTCCGCAAACGCACGGTTACCTACGCGAGGCGCACCGTAAGA	42

Table 9.9 (continued):

Name	Sequence	Length
d25	CCGTGCGTTTGCGGAGTTCCTCACCGTTCAAACGGGTGGTAC	42
d26	TCGTGTGGGTGATACGGTACAGAGTACCACCCGTTTGAACGG	42
d27	TACCGTATCACCCACACGAATGACATTGTTCCGCGTCTGCCT	42
d28	TGTGAGAGTAACCGAATTCACGTGGAGGCAGACGCGGAACAA	42
d29	CGTGAATTCGGTTACTCTCACAGCAGCCCGGAGTACTGGATT	42
d30	ACTGGAACCAGGGTACCAGATTTAATCCAGTACTCCGGGCTG	42
d31	CTGGTACCCTGGTTCCAGTCACCCGTAACGACATCGTTAAAA	42
d32	AGTGGCATCGATACCTTCGATTTTAACGATGTCGTTACGGGT	42
d33	TCGAAGGTATCGATGCCACTGGCGGTAACAACCAGCCTAACA	42
d34	AGGTGCGCAGGGATGTCAGGAATGTTAGGCTGGTTGTTACCG	42
d35	GACATCCCTGCGCACCTCTGGTATTTCGGTCTGATCGGCACT	42
d36a	GGATATCTCGAGGCCGGAGCCCAGGCAAGTGCCGATCAGACCGAA	45
d36b	GGATATCTCGAGTCATTACTACAGGCAAGTGCCGATCAGACCGAA	45

Table 9.10: Assembly PCR reaction.

PCR mix	PCR cycle
1 µL oligomer mix (a or b)	1. 98 °C 30 s
5 µL phusion buffer	2. 98 °C 5 s
0.5 µL dNTPs	3. 63 °C 15 s
18.25 µL ddH_2O	4. 72 °C 15 s (go to step 2. 24 X)
0.25 µL phusion enzyme (high fidelity)	5. 72 °C 5 min
25 µL total volume	6. 4°C forever

Table 9.11: Amplification PCR reaction.

PCR mix	PCR cycle (gradient PCR)
1 µL assembly product (a or b)	1. 98 °C 30 s
5 µL phusion buffer	2. 98 °C 5 s
0.5 µL dNTPs	3. 63 °C 15 s
18.25 µL ddH2O	4. Gradient 72 to 78 °C 15 s (go to step 2. 29 X)
0.25 µL phusion enzyme (high fidelity)	5. 72 °C 5 min
25 µL total volume	6. 4 °C forever

Table 9.12: Digestion and dephosphorylation.

Digestion (construct)	Digestion (vector)	Dephosphorylation
1.5 µL construct	10.0 µL pET21-a(+)	50.0 µL digested vector
3.0 µL NE buffer	5.0 µL NE buffer	7.0 µL AP buffer
0.3 µL BSA (100X)	0.5 µL BSA (100X)	12.0 µL ddH2O
10.7 µL ddH2O	5.0 µL ddH2O	1.0 µL antarctic phosphatase
0.5 µL XhoI	0.5 µL XhoI	
0.5 µL NdeI	0.5 µL NdeI	
Incubate 2 h at 37 °C	Incubate 3 h at 37 °C	Incubate 1 h at 37 °C

Table 9.13: NZY medium: Dissolve peptone, yeast extract and NaCl, adjust the pH to 7.5 and sterilize. Afterwards, add Glucose, $MgCl_2$ and $MgSO_4$.

Substance	**Concentration**
Peptone	1.0 % w/v
Yeast extract	0.5 % w/v
NaCl	0.5 w/v
Glucose	0.02 M
$MgCl_2$	0.0125 M
$MgSO_4$	0.0125 M

Table 9.14: Colony PCR reaction.

PCR mix	**PCR cycle (gradient PCR)**
2.5 µL Taq buffer	1. 95 °C 30 s
0.5 µL dNTPs	2. 95 °C 5 s
0.13 µL Taq polymerase	3. 51 °C 30 s
1.0 µL T7 forward primer	4. 68 °C 90 s (go to step 2. 29 X)
1.0 µL T7 reverse primer	5. 68 °C 10 min
9.87 µL ddH_2O	6. 4 °C forever

9.5 Reagents, assays and kits

Table 9.15: Kits.

Kit	Source	Order number
BCA Protein Assay	Thermo Scientific	23225
QIAprep Spin Miniprep	Quiagen	27104
GeneElute™ PCR clean-up	Sigma-Aldrich	NA1020
Quick Ligation™	New England Biolabs	M2200S
EnzChek® Lysozyme Assay	Invitrogen	E22013

Table 9.16: Stock solutions for the refolding screen ([1] stored at −20 °C).

Substance	Concentration
PB (pH 6.0, 7.0)	0.4 M
HEPES (pH 6.0, 7.0, 8.0, 9.0)	0.5 M
MOPS (pH 6.0, 7.0, 8.0, 9.0)	0.5 M
TRIS·HC (pH 7.0, 8.0, 9.0)	2.5 M
NaCl	5 M
KCl	2.5 M
Glycerol	60 % v/v
PEG 4000	5 % w/v
EDTA	0.1 M
Cu/Zn/Mg/Mn ($CuSO_4 \cdot 5\ H_2O$, $ZnSO_4 \cdot 7\ H_2O$, $MnSO_4 \cdot H_2O$, $MgSO_4$)	0.02 M
Brij 35	5 mM
CHAPS	500 mM
ZWITTERGENT 3-12	200 mM

Table 9.16 (continued):

Substance	Concentration
TWEEN 20	5 mM
TRITON-X 100	50 mM
SDS	500 mM
DTT[1]	1 M
TCEP[1]	100 mM
GSH[1]	50 mM
GSSG[1]	50 mM

Table 9.17: Critical micellar concentration (CMC) of the applied detergents.

Detergent	CMC	Unit
CHAPS	8	mM
ZWITTERGENT 3-12	3	mM
NDSB 201	-	-
TWEEN 20	60	µM
TRITON-X 100	600	µM
BRIJ 35	90	µM
SDS	9	mM
SDC*	6 (-)	mM

CHAPS, cholamidopropyl-dimethylammonium-propanesulfonate; **ZWITTERGENT 3-12,** dodecyl-dimethyl-ammonio-propanesulfonate; **NDSB 201,** non-detergent sulfobetaine 201; **TWEEN 20**, polyethylene glycol sorbitan-monolaurate; **TRITON-X 100,** polyethylene glycol tert-octylphenyl ether; **BRIJ 35**, polyethylene glycol dodecyl ether; **SDS,** sodium dodecyl sulfate; **SDC**, deoxycholic acid sodium salt; **DTT**, dithiothreitol; **TCEP**, tris-carboxyethyl-phosphine; **GSH**, reduced glutathione; **GSSG**, oxidized glutathione

Table 9.18: Buffer solutions for the SDS-PAGE.

Substance	**Concentration**
2 X Stacking gel buffer	
TRIS·HCl	250 mM, pH 6.8
SDS	0.4 % (w/v)
4 X Separating gel buffer	
TRIS·HCl	1.5 M, pH 8.8
SDS	0.8 % (w/v)
5 X Laemmli buffer	
TRIS·HCl	300 mM, pH 6.8
Glycerol	50 % v/v
SDS	10 % w/v
2-Mercaptoethanol	5 % v/v
Bromophenol blue	0.05 % w/v
10 X Running buffer	
TRIS·HCl	250 mM
Glycine	1.92 M
SDS	1 % w/v

Table 9.19: Buffer solutions for Coomassie staining.

Substance	**Concentration**
Fairbanks A	
Isopropanol	25 % v/v
Acetic acid	10 % v/v
Coomassie Brilliant Blue R-250	0.05 % w/v
Fairbanks B	
Isopropanol	10 % v/v
Acetic acid	10 % v/v
Coomassie Brilliant Blue R-250	0.05 % w/v
Fairbanks C	50 % v/v
Acetic acid	10 % v/v

Table 9.20: TAE buffer solution (10x stock solution) for agarose gel electrophoresis.

Substance	**Concentration**
TRIS, pH 8.0	400 mM
EDTA	10 mM
Acetic acid	1.14 % v/v

Table 9.21: Cultivation media for E. coli (agar plates with an additional 15 g L^{-1} agar agar, selection media with 100 mg L^{-1} ampicillin).

Substance	**Concentration**
Luria broth (LB)	
Peptone	10 g L^{-1}
Yeast extract	5 g L^{-1}
NaCl	5 g L^{-1}
Terrific broth (TB)	
Peptone	12 g L^{-1}
Yeast extract	24 g L^{-1}
Glycerol	5.04 g L^{-1}
KH_2PO_4	2.13 g L^{-1}
K_2HPO_4	12.54 g L^{-1}

Table 9.22: Lactate dehydrogenase activity assay.

Substance	**Concentration**
Assay buffer (TRIS·HCL pH 7.3)	0.2 M
NADH	7 mM
Pyruvate	60 mM

Table 9.23: Lipase activity assay.

Substance	**Concentration**
Solution A	
4-nitrophenyl palmitate	3 g L^{-1} (in n-propanol)
Solution B	
Triton X-100	5 g L^{-1}
Gum arabic	1 g L^{-1}
TRIS·HCL pH 7.5	2.5 M