Impact of enzyme rigidity and flexibility on stability against environmental influences, promiscuity, and expression on large-scale

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Christina Gohlke, geb. Nutschel

aus Düsseldorf

Düsseldorf, Juli 2020

Aus dem Jülich Supercomputing Centre (JSC),
dem John von Neumann Institute for Computing (NIC),
der Computational Biophysical Chemistry group (CBClab),
dem Institute of Biological Information Processing (IBI-7)
des Forschungszentrum Jülichs GmbH

&

dem Institute of Molecular Enzyme Technology (IMET)
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:
1. Prof. Dr. Karl-Erich Jaeger
2. Prof. Dr. Birgit Strodel
Tag der mündlichen Prüfung: 14.04.2021

Eidesstattliche Erklärung

Ich versichere an Eides Statt, dass die vorliegende Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Diese Dissertation wurde in der vorgelegten oder einer ähnlichen Form noch bei keiner anderen Institution eingereicht, und es wurden bisher keine erfolglosen Promotionsversuche von mir unternommen.

_____

Düsseldorf, im Juli 2020

TABLE OF CONTENTS

LIST OF PUBLICATIONS

This thesis is based on the following four publications:

PUBLICATION I: Rigidity theory for biomolecules: concepts, software, and applications.

Hermans, S.M.A., Pfleger, C., Nutschel, C., Hanke, C.A., Gohlke, H.

WIREs Comput Mol Sc. 2017, 7, e1311.

(C.N. wrote the topic on rigidity theory-based thermostability predictions)

PUBLICATION II: Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for *Bacillus subtilis* lipase A

Nutschel, C., Fulton, A., Zimmermann, O., Schwaneberg, U., Jaeger, K.-E., Gohlke, H.

J Chem Inf Model. 2020, 60, 3, 1568-1584.

(C.N. analyzed the experimental data, performed MD simulations and CNA computations, analyzed the computational results, and wrote the manuscript)

PUBLICATION III: Promiscuous esterases counterintuitively are less flexible than specific ones

Nutschel,-C., Coscolín, C., Mulnaes, D., David, B., Ferrer, M., Jaeger, K.-E., Gohlke, H.

J Chem Inf Model. 2020, DOI: 10.1021/acs.jcim.1c00152.

(C.N. analyzed the experimental data, performed structure prediction, MD simulations and CNA computations, analyzed the computational data, and wrote the manuscript)

PUBLICATION IV: Contribution of single amino acid and codon substitutions to the production and secretion of a lipase by *Bacillus subtilis*

Skoczinski, P., Volkenborn, K., Fulton, A., Bhadauriya, A., Nutschel, C., Gohlke, H., Knapp, A., Jaeger, K.-E.

Microb Cell Fact. 2017, 16, 160.

(C.N. performed CNA computations and drafted the corresponding parts in the manuscript)

ABBREVIATIONS

AA	Amino acid
ABC	ATP-binding cassette
BRENDA	BRaunschweig ENzyme *DAtabase*
*Bs*LipA	Lipase A from *Bacillus subtilis*
B. subtilis	*Bacillus subtilis*
CalB	*Lipase B* from *Candida antarctica*
CAR	*Catalytically active residues*
CMC	Critical micelle concentration
CNA	Constraint Network Analysis approach
CYP	Cytochrome P450
$(\Delta)D$	Experimental detergent tolerance (in comparison to wild type)
ΔD_{max}	Highest maximum effects of experimental detergent tolerance in comparison to wild type
D_{cut}	Distance cutoff
DNN	Deep neural network
DOF	Degrees of freedom
EC	Enzyme Commission
E. coli	*Escherichia coli*
E_{cut}	Energy cutoff [kcal*mol^{-1}]
E_{HB}	Mayos's hydrogen bond potential energy [kcal*mol^{-1}]
ELISA	Enzyme-linked immunosorbent assay
ENT	Ensembles of network topologies
ENT^{FNC}	Ensembles of network topologies based on the concept of fuzzy noncovalent constraints
ESTs	Carboxylesterases (EC 3.1.1.1)
F	Floppy modes
F	Classification of bacterial lipolytic enzymes based on Arpigny and Jaeger[80]
FDA	Food and Drug Administration
FIRST	Floppy Inclusion and Rigid Substructure Topography software
F_{mxw}	Floppy modes according to Maxwell
GRAS	Generally recognized as safe
H/H_{type2}	Cluster configuration entropy (Type2)
lipA	Gene encoding lipase A of *Bacillus subtilis*
LIPs	Triacylglycerol hydrolases (EC 3.1.1.3)
LPS	Lipopolysaccharides
MCMC	Markov Chain Monte Carlo
MD	Molecular dynamics
meta-MQAP	meta Model Quality Assessment Program
N_c	Number of independent constraints
N_r	Number of redundant constraints
N_s	Number of constraints in a subgraph

OM	Outer cell membrane
p_i	Percolation index
P_∞	Rigidity order parameter
r_i	Rigidity index
rc_{ij}	Stability map
$rc_{ij,neighbor}$	Neighbor stability map
$P_{EST/LIP}$	Substrate promiscuity of esterase (EST) and lipase (LIP)
ProFASi	Protein Folding and Aggregation Simulator approach
$\widetilde{rc}_{ij,\ neighbor}$	Median neighbor stability map
S	Mean rigid cluster size
T	Temperature
T_{og}	Optimal growth temperature
T_p	Phase transition temperature
$(\Delta)T_{50}$	Experimental thermostability (in comparison to wild type)
$\Delta T_{50;\ max}$	Highest maximum effects of experimental thermostability in comparison to wild type
Vol_{eff}	Active site effective volume
wt*Bs*LipA	Wild type lipase A from *Bacillus subtilis*
$\{\sigma\}$	Network states

ZUSAMMENFASSUNG

Heutzutage sind Enzyme wegen ihrer vielfältigen Anwendungen in der Lebensmittel-, Waschmittel-, Medizin- oder Pharmaindustrie in unserem täglichen Leben immer allgegenwärtiger. Sie erfüllen jedoch nicht immer die erforderlichen Anforderungen industrieller Anwendungen in „rauen“ Umgebungen, wie hohen Temperaturen oder der Anwesenheit von Lösungs- und Reinigungsmitteln. Um industrielle Anwendungen effizienter zu gestalten, werden außerdem Enzyme mit einem breiten Substratspektrum und hohen Produktausbeuten bevorzugt. Die moderne Enzymtechnologie weist ein zunehmendes Potenzial für eine Vielzahl interdisziplinärer Verfahren zur Entwicklung neuartiger maßgeschneiderter Enzyme für menschliche Zwecke auf. Insbesondere das „Protein Engineering“ hat sich als nützliches Werkzeug für die Entwicklung neuartiger maßgeschneiderter Enzyme mit verbesserten Eigenschaften herausgestellt. Am gebräuchlichsten sind wissensbasierte Strategien, bei denen das „Wissen“ aus Informationen über die Proteinstruktur und / oder -sequenz sowie Computertechniken mit Experimenten kombiniert wird. Da es jedoch an umfassenden experimentellen Daten, die auf einheitliche Weise gemessen wurden, mangelt, ist die Entwicklung und Validierung von Algorithmen für wissensbasierte Strategien unbefriedigend. Im Vergleich zu früheren Studien habe ich in meiner Dissertation zum ersten Mal wissensbasierte Strategien angewendet, um den Einfluss der Enzymflexibilität und -rigidität auf die Protein-Thermostabilität und / oder -detergenzien-Toleranz, -substratpromiskuität und -expression mit unserer internen Constraint Network Analysis (CNA)-Software in großem Maßstab für biotechnologisch hochrelevante bakterielle lipolytische Enzyme zu untersuchen.

ABSTRACT

Nowadays, enzymes are becoming ever more ubiquitous in our daily lives because of their diverse applications such as in the food, detergent, and medical or pharmaceutical industries. However, they do not always meet the required demands of industrial applications in terms of harsh environments, such as high temperatures or the presence of solvents and detergents. In addition, to make industrial applications more efficient, enzymes with a broad substrate spectrum and high product yields are preferred. Modern enzyme technology offers an increasing potential of a wide range of interdisciplinary processes for designing novel tailor-made enzymes according to human purposes. Especially, protein engineering has emerged as a useful tool for developing novel tailor-made enzymes with improved properties. Most common are knowledge-driven strategies, where the "knowledge" from information about the protein structure and / or sequence as well as computational techniques is combined with experiments. However, as there is a lack of available experimental large-scale data measured in a uniform way, the development and validation of algorithms for knowledge-driven strategies has remained unsatisfactory. Here, compared to previous studies, for the first time, I applied knowledge-driven strategies to rationalize the impact of enzyme flexibility and rigidity on protein thermostability and / or detergent tolerance, substrate promiscuity, and expression with our in-house Constraint Network Analysis (CNA) software at large-scale for biotechnologically highly relevant bacterial lipolytic enzymes.

1 INTRODUCTION

Enzymes are biomolecules, typically *proteins* made up by building blocks called *amino acids* (AAs), which are essential for nearly all biochemical reactions within living cells such as energy storage, cellular respiration, and signal transduction[1-3]. However, enzymes do not only play an important role within living cells. Already thousands of years ago, as nobody was aware of the existence of enzymes, people have used microorganisms for the production as well as preservation of food and feed, e.g., yeast dough, alcoholic beverages, vinegar, cheese, and silage[4] **(Figure 1)**.

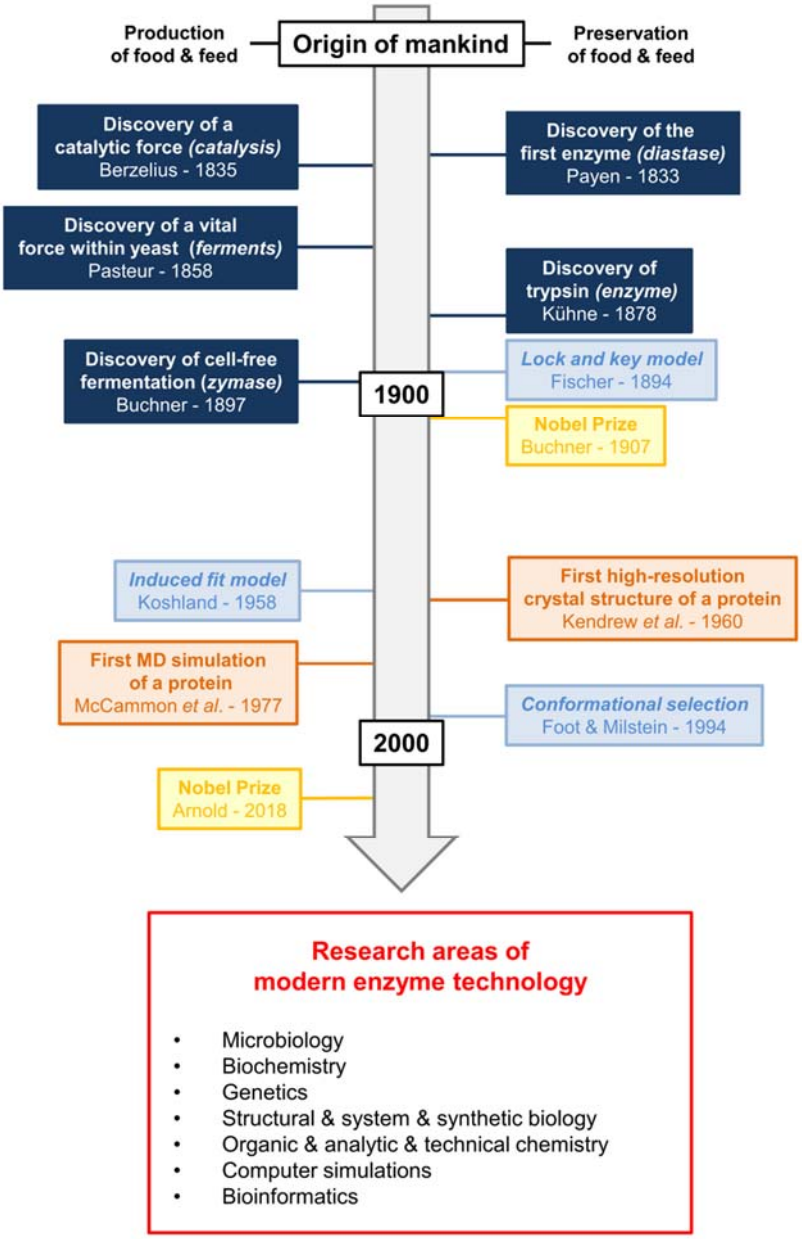


Figure 1: History of enzyme technology. Since the beginning of mankind, microorganisms have been used for the production and preservation of food and feed (top). In the 19th century, the first enzymes and their properties were discovered (dark blue boxes). Later, three theories rationalizing how substrates bind to enzymes, i.e. the

lock and key model, the *induced fit model*, and *conformational selection*, were postulated (light blue boxes). The first and the latest Nobel Prizes in Chemistry with respect to enzyme technology are shown in yellow boxes. Further milestones, i.e. the determination of the first high-resolution crystal structure of a protein[5] and the implementation of the first molecular dynamics (MD)-simulation of a protein[6], are shown in orange boxes. Nowadays, a variety of research areas contribute to modern enzyme technology (red box).

In the early 19th century, the discovery of enzymes and their properties began **(Figure 1)**. The first enzyme, the so-called *diastase* (from Greek: diastasis, "separation"), was discovered by the French chemist Anselme Payen in 1833[7, 8]. In 1835, the Swedish chemist Jöns Jakob Berzelius proposed the existence of a catalytic force and introduced the term *catalysis* (from Greek: kata and lyein, "down" and "loosen")[2, 9, 10]. Later, in 1858, when studying the fermentation of sugar to alcohol by yeast, the French chemist Louis Pasteur postulated that it was catalyzed by a vital force contained within the yeast cells, so-called *ferments* (from Latin: fermentum, "yeast"), which were thought to function only within living organisms[11, 12]. Finally, the term *enzyme* (from Greek: énzymon, "in yeast") was first used by the German physiologist Wilhelm Kühne in 1878[2, 3]. Around 20 years later the foundation of modern enzymology was laid by the German chemist Eduard Buchner, who demonstrated that sugar was fermented by *zymase*, a protein-containing substance in yeast, even without living cells[2, 13]. In 1907, he received the Nobel Prize in Chemistry for his work **(Figure 1)**.

The reason why nature has evolved a variety of enzymes is that the majority of the abovementioned cellular processes would not take place spontaneously. Almost all enzymes follow the same principle: The so-called *active site* of an enzyme binds a *substrate*, catalyzes a reaction by which *products* are formed, and then allows the products to dissociate. Meanwhile, the enzyme increases the reaction rate by lowering the *activation energy* (E_a) of the reaction, the energy that is required to start the reaction. The lower E_a, the faster a reaction happens. There are three theories proposing three distinct models of the mechanism of enzyme-substrate binding: The *lock and key model*, the *induced fit model*, and the *conformational selection*[14-20] **(Figures 1 and 2)**. Already in 1894, the chemist Emil Fischer postulated with the *lock and key model* that only substrates (the keys) with the correct shape would fit into the active site (the key hole) of the enzyme (the lock)[14] **(Figures 1 and 2A)**. Later, in 1958, Daniel Koshland's *induced fit model* suggested that the shape of the active site changed until the substrate is completely bound[15] **(Figures 1 and 2B)**. Finally, in 1994, Jefferson Foot and Cesar Milstein proposed the *conformational selection* assuming that all enzymes are inherently dynamic and sample a vast ensemble of conformations of which substrates bind to the most favored one[16-19] **(Figures 1 and 2C)**. Hence, unlike the lock and key model, the induced fit model and the conformational selection assume that enzymes are

rather flexible structures. This finding is in agreement with today's scientific knowledge. Many evidences have been found that enzyme flexibility is linked to biomolecular structure, (thermo-)stability, and function. One interesting example is that thermophilic enzymes are generally less flexible than their mesophilic homologues[21, 22] **(section 2.2.2.3.1)**. Hence, the increased structural rigidity of thermophilic enzymes can explain how they can maintain their functional integrity at high temperatures. Another example is that promiscuous human cytochrome P450 (CYP) enzymes that are involved in drug metabolism have more flexible active sites[23-25]. Thus, the induced fit model and the conformational selection are generally considered in such cases to be the more correct ones.

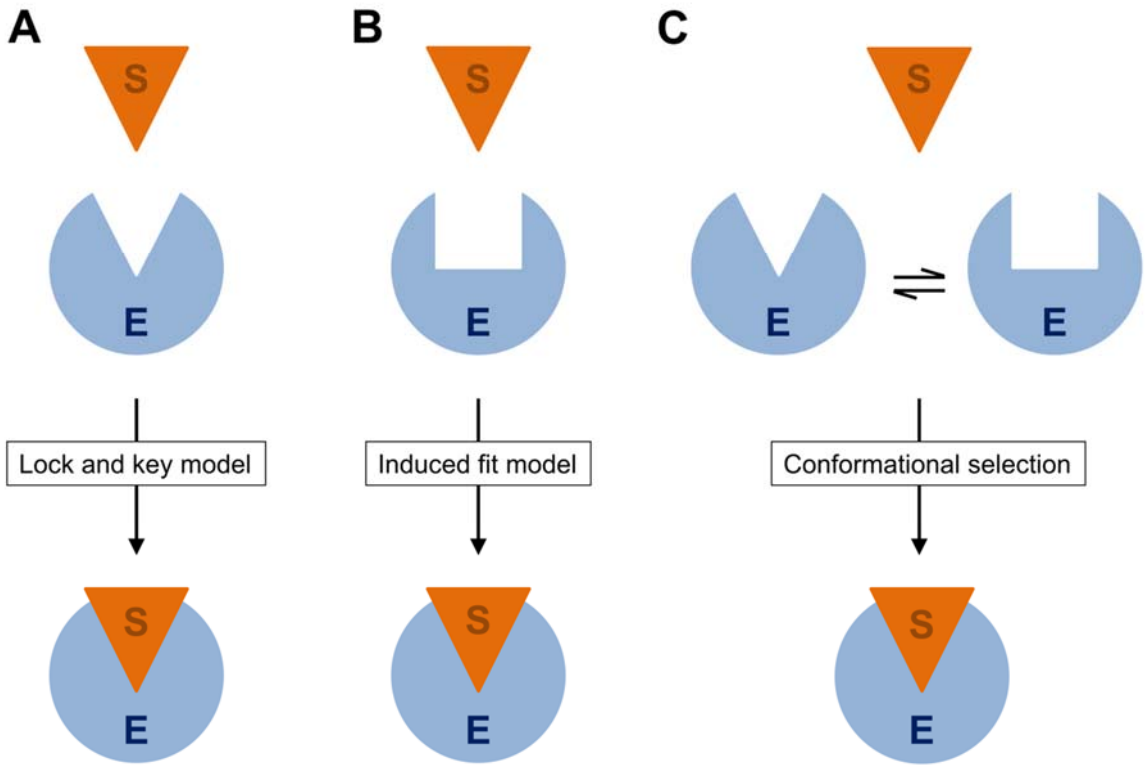


Figure 2: The different models of enzyme-substrate binding. (A) The *lock and key model* proposes that only substrates with the correct shape would fit into the active site of the enzyme. (B) The *induced fit model* suggests that the shape of the active site changes until the substrate is completely bound. (C) *Conformational selection* assumes that enzymes are inherently dynamic and sample a vast ensemble of conformations of which substrates bind to the most favored one. The substrate (abbreviated as S) is shown as orange triangle, whereas the enzyme with its active site (abbreviated as E) is shown as light blue circle. Figure was taken and adapted from Savir *et al.*[20].

Nowadays, enzymes are becoming ever more ubiquitous in our daily lives because of their diverse applications such as in the food, detergent, and medical or pharmaceutical industries[4] **(section 2.3.3)**. Indeed, the increasing demand of enzymes can be seen by the *global industrial enzyme market* that has been forecast to reach US$ 7.0 billion by 2023 from US$ 5.5 billion in 2018[26]. Enzymes catalyzing a chemical reaction are so-called *biocatalysts*,

whereas the usage of isolated biocatalysts or whole cells (bacteria, fungi, microalgae and plants, among others) is referred to as *biocatalysis*[27]. In the context of *green chemistry*, an approach that aims at developing more sustainable chemical processes with less hazardous substances, biocatalysis has shown many advantages compared to traditional chemical synthesis[4, 28]. With this respect, the most important advantages are: (I) Biocatalysts can operate at mild conditions in aqueous media at close to room temperatures and low pressures, (II) biocatalysts are substrate specific, (III) biocatalysts remain unchanged by the catalyzed reactions and can be reused, (IV) biocatalysts are generally highly chemo-, regio-, and enantioselective, (V) the usage of whole cells as biocatalysts enables cofactor recycling[4, 29-31]. However, despite all these advantages, natural enzymes do not always meet the required demands of industrial applications in terms of harsh environments, such as high temperatures or the presence of solvents and detergents[32, 33]. In addition, to make industrial applications more efficient, enzymes with a broad substrate spectrum and high product yields are often preferred[34, 35]. Modern *enzyme technology* offers an increasing potential of a wide range of interdisciplinary processes for designing novel tailor-made enzymes according to human purposes[2] **(Figure 1)**. Therefore, a broad variety of research areas contribute to modern enzyme technology, e.g., microbiology, biochemistry, and bioinformatics. Especially, *protein engineering* **(section 2.1)** has emerged as a useful tool in enzyme technology. The timeliness of protein engineering can be seen by the award of the Nobel Prize in Chemistry to Frances H. Arnold for pioneering the use of *directed evolution* **(section 2.1.1)** to engineer enzymes in 2018 **(Figure 1)**. However, the most common approach of protein engineering is based on *knowledge-driven strategies* **(section 2.1.3)**, where the "knowledge" obtained from information about the protein structure and / or sequence as well as from computational predictions is combined with experiments[36-39]. Nevertheless, as there is a lack of available experimental large-scale data measured in a uniform way the development and validation of algorithms for data analysis in knowledge-driven strategies is often unsatisfactory[40-43].

Here, I applied knowledge-driven strategies to rationalize the impact of enzyme flexibility and rigidity on protein stability against environmental influences, i.e. protein thermostability and / or detergent tolerance **(section 5, PUBLICATION II[44])**, substrate promiscuity **(section 6, PUBLICATION III[45])**, and expression **(section 7, PUBLICATION IV[35])** using our in-house Constraint Network Analysis (CNA) software[46] **(section 2.2.2)**. CNA is a graph theory-based rigidity analysis approach for linking a biomolecular structure, flexibility, (thermo)stability and function. Until now, CNA has been successfully applied to small-scale data sets of proteins to investigate protein thermostability **(section 2.2.2.3)**. However, CNA

has not been applied to a large-scale data set of protein variants to investigate either protein thermostability or other protein properties. In this respect, in order to validate my knowledge-driven strategies based on CNA predictions, I systematically investigated for the first time large-scale data sets of biotechnologically highly relevant bacterial lipolytic enzymes **(section 2.3)**.

2 BACKGROUND

2.1 Protein engineering strategies

The following section was taken and adapted from **PUBLICATION II**[44].

Natural enzymes are versatile biocatalysts catalyzing a wide variety of reactions. However, they do not always meet the required demands of industrial applications in terms of harsh environments, such as high temperatures or the presence of solvents and detergents[32, 33]. Hence, protein engineering has emerged as a useful tool for developing novel tailor-made enzymes. There are two major strategies for protein engineering: directed evolution and rational design, both of which can be combined into knowledge-driven strategies **(Figure 3)**[47].

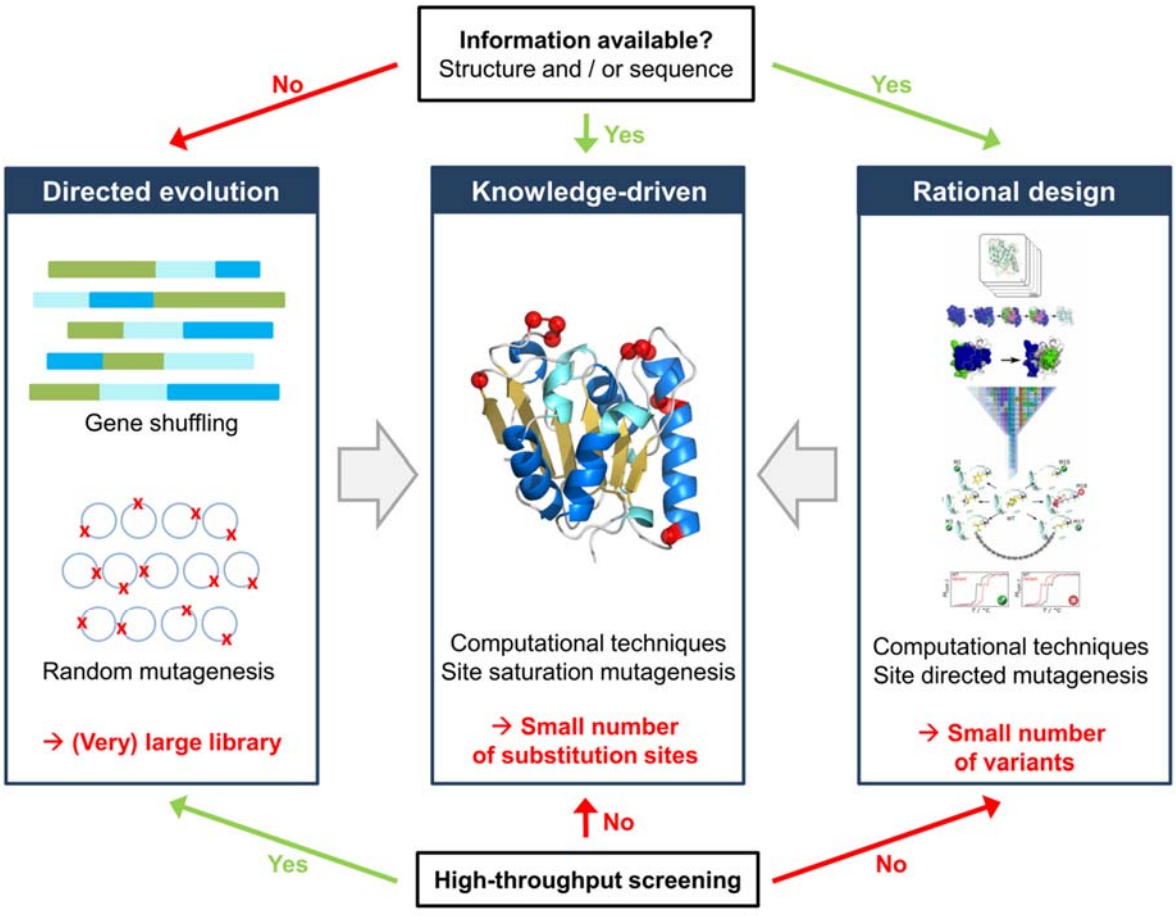


Figure 3: Overview of protein engineering strategies. Selection of protein engineering strategies based on the availability of protein structure and / or sequence as well as high-throughput screening (HTS) methods. Directed evolution **(left)** improves protein functions through iterative cycles of mutagenesis and screening or selection. Hence, (very) large protein libraries are generated by either random recombination of a set of related sequences, such as gene shuffling, or random mutagenesis. Rational design **(right)** applies computational techniques to predict the effect of specific substitutions, which are introduced by site directed mutagenesis (SDM). This results in a small number of variants. As an example, a strategy by Rathi *et al.*[48] is shown, where specific substitutions of *Bacillus subtilis* lipase A (*Bs*LipA) **(section 2.3.4)** are rationally predicted and experimentally validated with respect to increased protein thermostability **(section 2.2.2.3.2)**. By combining the advantages of directed

evolution and rational design, knowledge-driven strategies **(middle)** lead to a small number of substitution sites. In **PUBLICATION II**[44] I predicted beneficial substitution sites of *Bs*LipA (PDB ID: 1ISP) **(section 2.3.4)** with respect to increased protein thermostability and/or detergent tolerance, which are shown as red spheres. The predicted substitution sites were validated against a complete experimental site saturation mutagenesis (SSM) library. Figure adapted from Steiner *et al.*[47].

In the following I will provide an overview about the different protein engineering strategies and emphasize the advantages of the knowledge-driven strategies, which I applied in **PUBLICATIONS II-IV**[35, 44, 45].

2.1.1 Directed evolution

Following the principles of natural evolution, albeit on a reduced timescale, protein engineering by directed evolution **(Figure 3)** has become an attractive strategy to improve protein functions through iterative cycles of mutagenesis and screening or selection[30, 32, 49-51]. The main advantage of directed evolution is that no information about the protein sequence and / or structure is needed. Common methods for library generation are based on either random recombination of a set of related sequences, e.g., gene shuffling, or the introduction of random mutations in single sequences, e.g., error-prone PCR (epPCR), Sequence Saturation Mutagenesis (SeSaM), and Phage-Assisted Continuous Evolution (PACE)[52, 53]. To successfully investigate (very) large protein libraries, powerful automated techniques for rapid high-throughput screenings (HTS) were established, such as fluorescence-activated cell sorting (FACS) or automated liquid handling[32, 49, 50, 54-56]. However, the highly labor-intensive methods can become technically challenging if beneficial mutations need to be accumulated over generations of mutagenesis and screening or selection to reach a desired effect[50]. After all, directed evolution is not good for problems that require multiple, simultaneous, low-probability events[57]. Many examples for the successful application of directed evolution are provided by Frances H. Arnold[50, 58, 59], who was honored with the Nobel Prize in Chemistry for pioneering the use of directed evolution **(Figure 3)**.

2.1.2 Rational design

Alternatively, protein functions can be modified by rational design **(Figure 3)** based upon the ability to predict the effect of a specific substitution by numerous computational techniques[47, 60, 61]. In contrast to directed evolution, information about the protein structure and / or sequence is evaluated to propose specific substitutions, which are introduced by site directed mutagenesis (SDM)[62]. Using rational design, the following three questions must be answered: (I) *Where* to substitute?, (II) *Which* substitution should be introduced?, and (III) *How* to

evaluate the effect of the substitution? The main advantage of rational design over directed evolution is an increased probability of beneficial protein variants and a significant reduction of the protein library size[60]. Thus, this strategy avoids the time- and cost-intensive generation and screening of large protein libraries, especially, if no HTS is available. An example for the successful application of rational design is given by a prospective study from Rathi *et al.*[48], where specific substitutions of *Bs*LipA that lead to increased thermostability are rationally predicted by the Constraint Network Analysis (CNA) approach **(section 2.2.2.3.2)**. Finally, the results were experimentally validated with respect to increased protein thermostability. Despite successful applications in single cases the general reliability of rational design is still unsatisfactory[40, 63-66]. One reason is that multiple attempts to identify key features in protein sequences and/or structures associated with protein function have failed to paint a clear picture, which makes it difficult to define rules of universal validity and general applicability[32, 41]. Another reason lies in the data used in the design and evaluation of computational techniques. For example, the ProTherm database[67, 68], a collection of thermodynamic data of proteins, contains on average ~12 single, ~12 double, and ~1 multiple substitution for each of the ~1000 proteins stored[32]. Thus, while overall exhaustive, the data may not include a sufficient number of variants per protein to compensate for outliers and, therefore, may not allow a stratification of the data to derive a generally applicable set of rules. As such data, furthermore, originates from different experimental methods, it is not surprising that different thermodynamic data have been found associated with the same variant[42]. In addition, the data is strongly biased towards substitutions to alanine, whereas it is very limited for some other substitutions[43].

2.1.3 Knowledge-driven strategies

As an intermediate, third route recent developments have tended towards knowledge-driven strategies **(Figure 3)**, which combine the advantages of directed evolution and rational design[36, 47, 69]. The "knowledge" generally arises from information about the protein structure and / or sequence as well as computational techniques[36-39]. First, substitution sites with high potential to yield beneficial protein variants are predicted; second, substitutions are engineered by SSM or SDM[41]. By such knowledge-driven strategies, the challenge of accurately predicting the effect of a substitution on protein function is circumvented, and substitution efforts are guided to a few, distinguished sequence positions, making subsequent combinations feasible. This strategy usually leads to smaller "smart" libraries with a higher probability of the desired improvement[69, 70]. However, even with HTS it is difficult to handle

all variants based on combinations of the 20 proteinogenic AAs at more than six substitution sites (i.e., more than $20^6 = 6.4 * 10^7$ variants)[32, 39, 49, 71]. In **PUBLICATIONS II-IV**[35, 44, 45] I applied knowledge-driven strategies based on CNA **(section 2.2.2)** to rationalize the impact of enzyme rigidity and flexibility on different protein properties.

2.2 Rigidity theory for biomolecules

The following part was taken and adapted from **PUBLICATION I**[72] in which we reviewed fundamental concepts in rigidity theory, ways to represent biomolecules as constraint networks, and methodological and algorithmic developments for analysing such networks and linking the results to biomolecular function. These applications include investigating large biomolecules such as the ribosome[73], understanding allostery[73-75], predicting thermodynamic properties[76], assessing the structural stability of complexes[77, 78], identifying folding cores of proteins[79, 80], sampling of biomolecular conformational spaces[81-84], finding putative binding sites[85], and analyzing structural determinants of thermostability[22, 86]. To automate and improve the efficiency of the analysis, several software packages have been developed[87, 88], including CNA[46]. In **PUBLICATIONS II-IV**[35, 44, 45] we performed rigidity analyses of proteins in various contexts based on CNA[46]. Assuming that proteins follow the same laws of physics as do mechanical structures, protein and mechanical rigidity are strongly interlinked. Hence, the basis of rigidity theory will be the focus in the following.

2.2.1 Basic concepts of rigidity theory

2.2.1.1 Constraint counting: Maxwell's rules for rigidity

Analyzing network rigidity was already of scientific interest in the 19th century when Maxwell investigated the stability of mechanical structures, e.g., bridges, consisting of struts (distance constraints) connected by joints **(Figure 4)**[89].

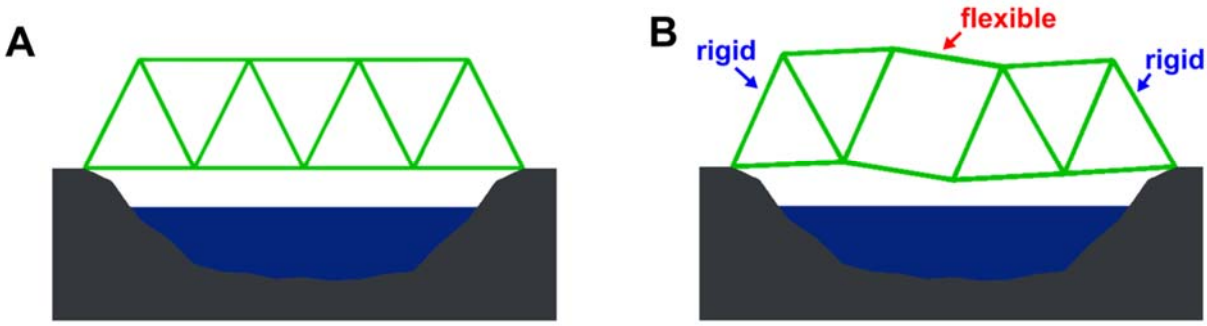


Figure 4: Network rigidity of mechanical structures. Schematic representation of a bridge consisting of struts (distance constraints) connected by joints. **(A)** In 2D, the triangle is the smallest rigid unit. Hence, if all constraints are in place, the bridge is *isostatically* or *minimally rigid*. **(B)** Removing one constraint divides the bridge into two *rigid* clusters with a *flexible* region in between. Figure taken and adapted from **PUBLICATION I**[72].

Maxwell introduced constraint counting as mean field approach to assign the number of independent internal degrees of freedom (abbreviated as DOF), called 'floppy modes'

(abbreviated as *F*). *F* determines possible movements of a structure in the *d*-dimensional space without violating any of the constraints. For a network with *N* sites, lacking any constraints, *F* is given by **Eq. 1**. The latter term denotes the global DOF.

$$F = dN - d(d+1)/2 \qquad \textbf{Eq. 1}$$

Maxwell assumed that in a system with independent constraints N_c each constraint removes one *F*. This results in the number of *F* according to Maxwell (abbreviated as F_{mxw}) given by **Eq. 2**.

$$F_{mxw} = dN - N_C - d(d+1)/2 \qquad \textbf{Eq. 2}$$

If not all constraints are independent, using Maxwell's equation will lead to an underestimation of *F*. This is corrected for by considering the number of redundant constraints N_r[90]. The number of *F* is thus given by **Eq. 3**.

$$F = dN - (N_c - N_r) - d(d+1)/2 \qquad \textbf{Eq. 3}$$

Redundant constraints introduce stress in the network and do not add to the stability of the network anymore[91]. A network region with redundant constraints is *overconstrained* or *stressed*. If the number of constraints and internal DOF is the same, the region is *isostatically* or *minimally rigid* **(Figure 4A)**[92]. A region with fewer constraints than internal DOF is defined as *underconstrained* or *flexible* **(Figure 4B)**. The principles to determine flexibility in mechanical structures can further be used in proteins.

2.2.1.2 Constraint network representations for proteins

Applying a constraint network representation to proteins reduces its complexity to the question of connectivity as no geometric details are considered. There are several types of constraint networks in which atoms are transformed into nodes and (non)covalent bonds into constraints in between[93]. Due to the fact that CNA **(section 2.2.2)** models a protein as a *body-and-bar* network[46], in the following, the focus is on this type of constraint network representation. Alternatively, proteins can be modeled as *bond-bending* network (also called *bar-and-joint* network or *molecular framework*)[94, 95] and *body-bar-hinge* network[88, 95].

In *body-and-bar* networks[95, 96], atoms are considered as rigid bodies having six DOF, which are connected by bars. Two rigid bodies have in total 12 DOF. Disregarding the six global DOF, six bars are needed to lock in the internal DOF and, hence, to model double and peptide

bonds. A single bond is modeled with five constraints, leaving one DOF for the dihedral rotation. Exemplarily the *body-and-bar* network representation of propene is shown **(Figure 5)**.

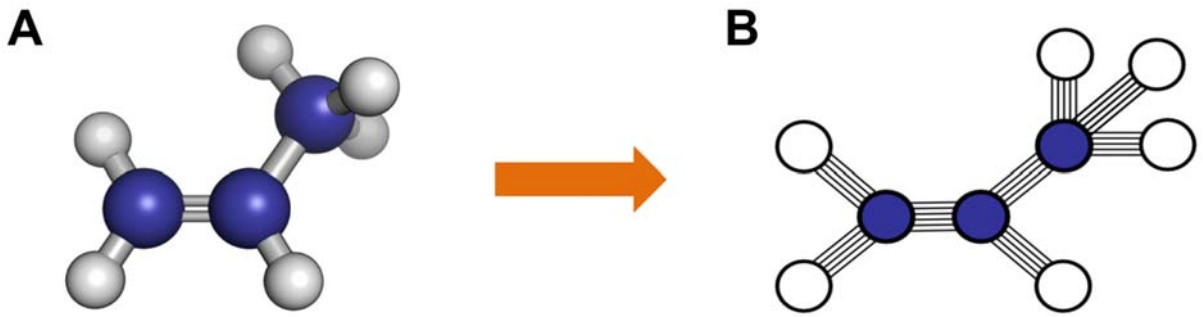


Figure 5: *Body-and-bar* network representation of propene. **(A)** Ball-and-stick representation of propene with carbon atoms shown in blue and hydrogen atoms shown in light gray. **(B)** In the *body-and-bar* network, atoms are modeled as bodies with six DOF, a single bond as five constraints between two bodies, and a double bond as six constraints between two bodies. Figure taken and adapted from **PUBLICATION I**[72].

Stronger noncovalent interactions, such as hydrogen bonds (including salt bridges) and hydrophobic interactions, are essential for the stability of proteins and, thus, require accurate modeling in the constraint network. In contrast, weaker interactions such as van der Waals or electrostatic forces are not included in the constraint network. In all types of constraint networks, modeling of different interaction strengths is possible by including a different number of constraints/bars[96, 97]. In *body-and bar* networks, hydrogen bonds are modeled with five bars, as are covalent bonds, and hydrophobic interactions with two bars[46, 96, 98] although lower and higher numbers of bars have been used for hydrophobic interactions, too[98].

Deciding which noncovalent interactions to include in the network is decisive for an accurate representation of the flexibility of the system[78, 99]. For this, the strength of hydrogen bonds is evaluated according to Mayos's hydrogen bond potential energy (E_{HB})[100]. Only hydrogen bonds with $E_{HB} \geq E_{cut}$, where E_{cut} represents an energy cutoff **(section 2.2.2.1)**, are included in the constraint network[87, 101]. Hydrophobic interactions are often included in the constraint network according to the criterion that the distance between carbon and/or sulfur atoms is less than the sum of their van der Waals radii (C: 1.7 Å, S: 1.8 Å) plus a distance cutoff D_{cut} = 0.25 Å[46]. Alternatively, Fox *et al.*[98] introduced a parameter to describe the strength of hydrophobic interactions based on pairwise van der Waals energies derived from the Lennard-Jones potential of the AMBER parm99 force field[102, 103]. Furthermore, it should be taken into account that the results of the rigidity analyses can be affected by additional factors such as water molecules[78, 87, 104], ions[105], small-molecule ligands[85, 96], and other biomolecules[78]. These

methods to represent proteins as constrained networks can now be implemented into algorithms to be used at a large scale. One common implementation is Laman's theorem and the pebble game algorithm.

2.2.1.3 Constraint Counting: Laman's theorem and pebble game algorithm

For a given constraint network, **Eq. 3** yields F in terms of a mean field approximation[90]. In 1970, *Laman's theorem*[90] had a major impact in that it allows to determine the DOF locally in generic (i.e., lacking any special symmetries) 2D constraint networks by applying constraint counting to all subgraphs within the network. Laman's theorem reads as follows: A generic 2D network is *minimally rigid* if and only if the number of constraints is $2N - 3$, and every non-empty subgraph s induced by $N_s \geq 2$ sites spans at most $2N_s - 3$ constraints. Based on Laman's theorem, Hendrickson[91] suggested an algorithm that exactly counts the number of F in a generic 2D network and, hence, is appropriate to decompose it into rigid regions and flexible links in between. Further developments on this algorithm led to the efficient combinatorial 2D pebble game algorithm implemented by Thorpe and Jacobs[106]. However, its generalization is not sufficient in higher dimensions, e.g., in the 3D double banana network **(Figure 6)**[107]. This network has overall $3N - 6$ constraints, and none of the subgraphs has more than $3N_s - 6$ constraints connecting N_s sites. Applying the 3D analog of Laman's theorem would thus lead to the conclusion that this network is minimally rigid, which is wrong as there is an implied-hinge joint between the two 'banana' subgraphs.

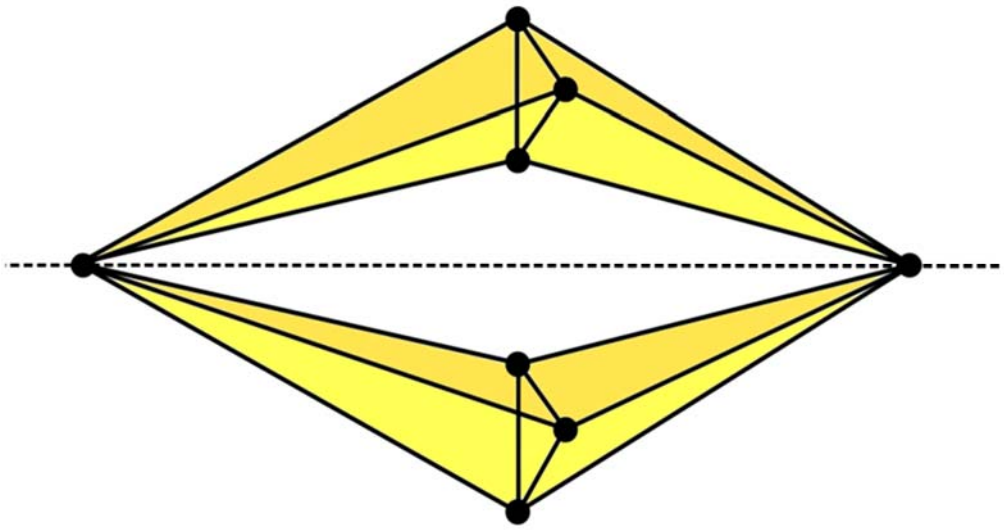


Figure 6: Double banana network. Constraint counting implies that the 3D double banana network is rigid because it satisfies the $3N - 6$ counting condition considering that the nodes have three DOF. However, internal motion within this network is possible along the implied-hinge joint between the two 'banana' subgraphs (dashed line). Figure taken and adapted from **PUBLICATION I**[72].

With the *molecular framework conjecture*, Tay and Whiteley[95] proposed that the constraint counting can be extended to a certain subtype of 3D networks with a molecule-like character, the *bond-bending* networks **(section 2.2.1.2)**. Based on this proposition, Jacobs constructed a 3D pebble game algorithm for these networks, the computational time complexity of which is, in a worst case scenario, $O(N^2)$; in practice, the algorithm runs in linear time[92, 101]. In comparison, brute force numerical techniques can give the same result as the pebble game algorithm, but are generally unfeasible for large systems due to a computational complexity of $O(N^3)$[101].

The pebble game algorithm for *bond-bending* networks **(section 2.2.1.2)** has been implemented in early versions of the Floppy Inclusion and Rigid Substructure Topography (FIRST) software[87]. CNA functions as a front- and back-end to FIRST[46] **(section 2.2.2)**. In 2004, Hespenheide *et al.*[96] implemented a 3D pebble game algorithm using a $6N - 6$ count applied on *body-and-bar* representations of molecules **(section 2.2.1.2)**. In 2008, Lee and Streinu[108, 109] described a family of pebble game algorithms, the (k,l)-pebble games, where k is the initial number of pebbles on each node and l is the acceptance condition, that is, the global degrees of freedom of the system. The original 2D pebble game algorithm of Jacobs and Hendrickson[110] is a (*2*,*3*)-pebble game in this terminology[108]. A (*6*,*6*)-pebble game implemented by Fox *et al.*[88] for analyzing *body-bar-hinge* networks **(section 2.2.1.2)** is equal to the 3D pebble game algorithm introduced by Hespenheide *et al.*[96] for analyzing *body-and-bar* networks **(section 2.2.1.2)**. Notably, the family of (k,l)-pebble games were proven to be correct by Katoh and Tanigawa in 2011[111].

When applying a 3D pebble game algorithm using a $6N - 6$ count on a *body-and-bar* network **(section 2.2.1.2)**, initially, each node in a network is assigned six pebbles corresponding to the six DOF in 3D. In order to decompose the network into flexible and rigid regions, the pebble game algorithm follows two rules[109]:

I. Define a constraint between the nodes: If the nodes i and j have at least seven pebbles in total, place a pebble on the constraint from i to j to define the constraint in the direction of j **(Figures 7A, B, E, F)**.

II. Slide a pebble: If there is a defined constraint between i and j and there is a pebble on j, reverse the direction of the constraint and move the pebble from j to i **(Figures 7C, D)**.

Exemplarily, a 3D pebble game algorithm using a $6N - 6$ count on a *body-and-bar* network of a biomolecule **(section 2.2.1.2)** is shown **(Figure 7)**.

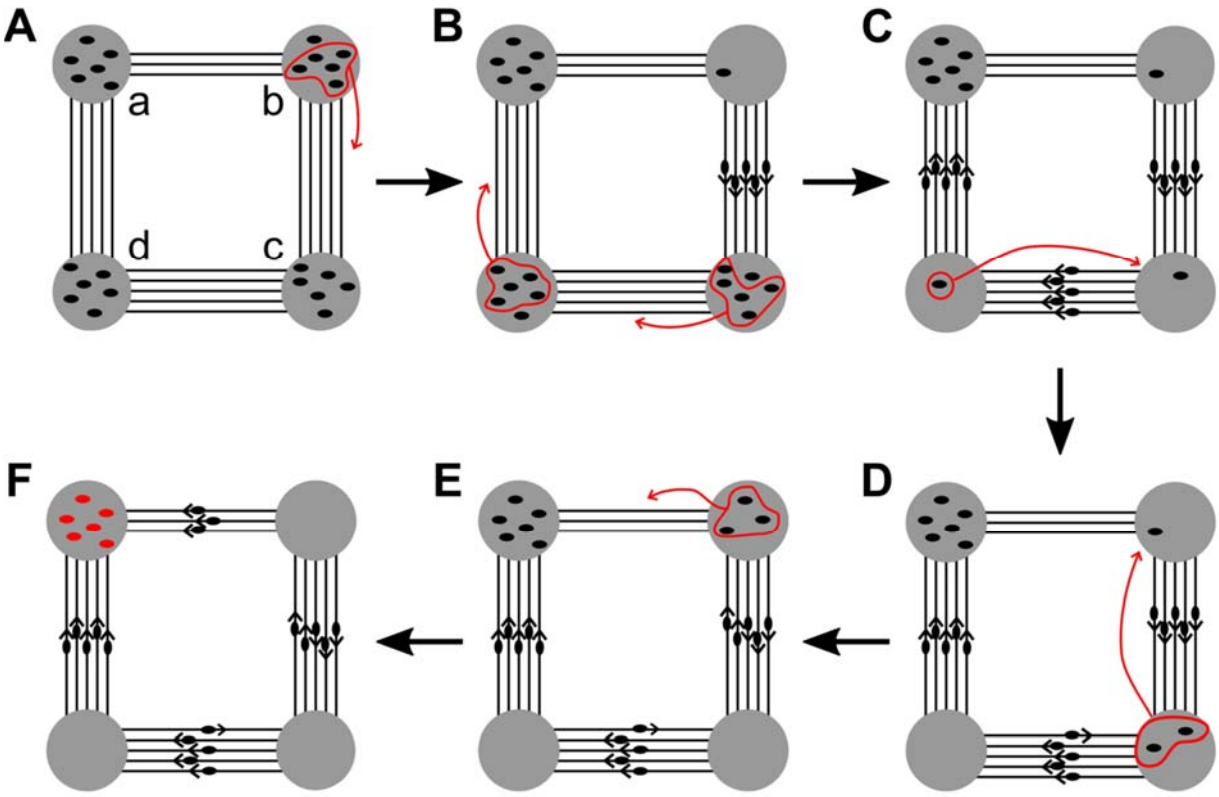


Figure 7: The 3D pebble game algorithm. An exemplary biomolecule is modeled as a *body-and-bar* network with four nodes connected by a total of 18 constraints. **(A)** Five pebbles are first placed on the constraints between **b** and **c** defining all five constraints in the same direction. **(B)** Then, five pebbles are placed on the constraints from **c** to **d** and from **d** to **a**. This leaves six pebbles on **a** and one pebble on **b**, **c**, and **d**, respectively. **(C, D)** All single pebbles are now collected on **b**. **(E)** There are now six pebbles on **a** and three pebbles on **b**; **c**, and **d** are empty. Finally, the last three constraints are defined by placing the three pebbles on the constraints between **b** and **a**. **(F)** Now 18 pebbles are used, and all constraints are defined. The remaining six pebbles on **a** represent the six global DOF, demonstrating that this graph is *minimally rigid*. Figure taken and adapted from **PUBLICATION I**[72].

2.2.2 Constraint Network Analysis

The Constraint Network Analysis (CNA) approach[46] was first introduced by Radestock and Gohlke[22] and aims at linking information from rigidity analysis derived from the Floppy Inclusion and Rigid Substructure Topography (FIRST) software[87] with biomolecular structure, (thermo-)stability, and function. FIRST, developed by Jacobs *et al.*[87], was the first implementation of a pebble game algorithm **(section 2.2.1.3)** together with code for generating constraint networks for proteins **(section 2.2.1.2)**. CNA functions as a front- and back-end to FIRST[46].

Going beyond the mere identification of flexible and rigid regions in a biomolecule, CNA allows for (I) performing constraint dilution simulations **(Figure 8A)** that consider a temperature dependence of hydrophobic tethers[112, 113], in addition to that of hydrogen bonds **(section 2.2.2.1)**, (II) computing a comprehensive set of global and local indices for

quantifying biomolecular stability[105] **(section 2.2.2.2) (Figure 8B-D)**, and (III) performing rigidity analysis on ensembles of network topologies (ENT). For the latter, structural ensembles obtained from molecular dynamics (MD) simulations and ensembles based on the concept of fuzzy noncovalent constraints (ENT^{FNC})[114] can be used. In short, ENT^{FNC} performs rigidity analyses of biomolecules on ENT generated from a single input structure. Here, the ENT are based on fuzzy noncovalent constraints, which considers thermal fluctuations of biomolecules without actually sampling conformations. That way, information on the influence of a finite temperature on constraint network representations is implicitly included without the need to derive system-specific parameters. As we[114, 115] and others[104, 116] observed, performing rigidity analysis on ENT instead of single networks greatly improves the robustness of the results. In **PUBLICATIONS II-IV**[35, 44, 45] constraint dilution simulations **(section 2.2.2.1)** were performed by CNA either on ENT generated from MD simulations or on ENT^{FNC}.

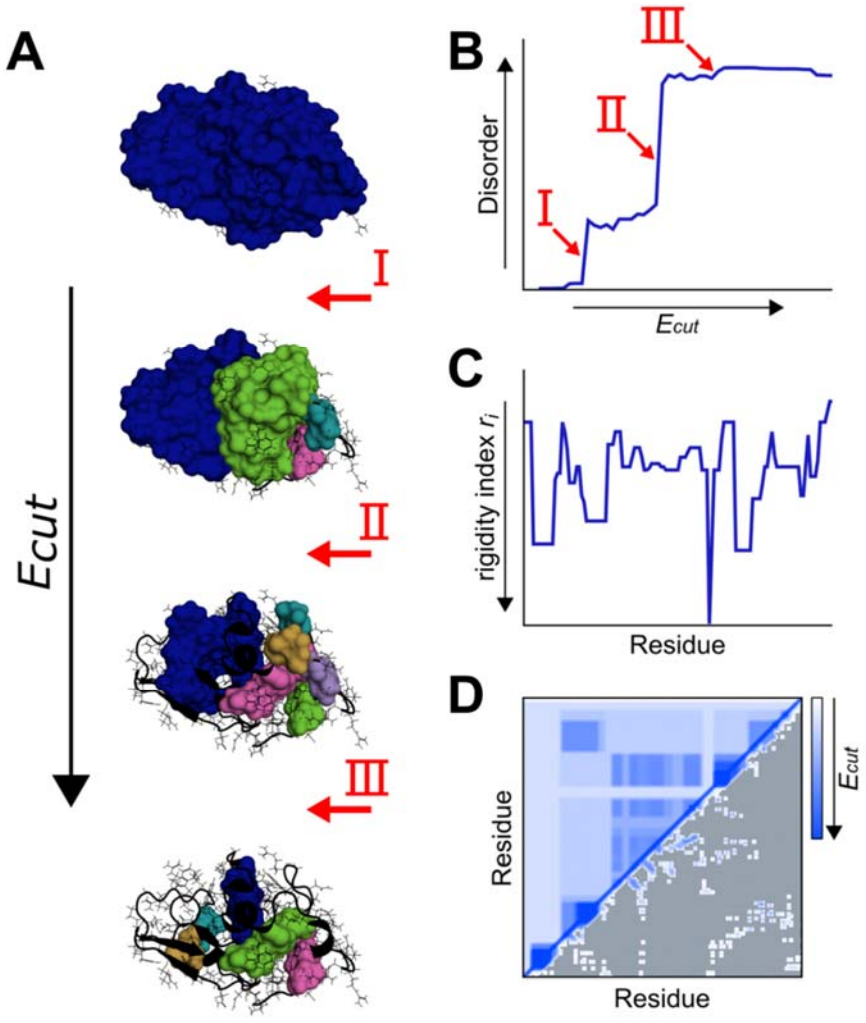


Figure 8: Results of a constraint dilution simulation of hen egg white lysozyme with CNA. (A) In the constraint dilution simulation, a stepwise decrease in the cutoff energy (E_{cut}) removes hydrogen bonds from the constraint network in the order of increasing strength. The colored surfaces represent the rigid clusters, and the black lines represent the flexible regions of the protein. **(B)** Degree of disorder along a constraint dilution

simulation as revealed form the cluster configuration entropy H. The disorder is low when a single rigid cluster dominates and increases when the cluster falls apart into smaller subclusters of different sizes. **(C)** The rigidity index r_i characterizes the per-residue stability as it monitors when a residue i segregates from any rigid cluster during a constraint dilution simulation. A lower r_i value indicates that the residue resides in a region of higher stability. **(D)** The stability map rc_{ij} represents when a 'rigid contact' between two residues of the network (both residues belong to the same rigid cluster) vanishes during the thermal unfolding simulation **(upper triangle)**; the neighbor stability map $rc_{ij,neighbor}$ considers only the rigid contacts between two residues that are at most 5 Å apart from each other, with values for all other residue pairs colored gray **(lower triangle)**. Note that arrows at axes labeled with E_{cut} point in the direction of more negative values. A blue (white) color indicates that contacts between residue pairs are more (less) rigid. Figure taken and adapted from **PUBLICATION I**[72].

In order to facilitate the processing of the highly information-rich results obtained from CNA, the VisualCNA plugin for PyMOL[117] and the CNA web server[118] have been developed. Both provide user-friendly interfaces around the CNA software for easily setting up CNA runs and analyzing results.

The CNA software and VisualCNA are available under academic licenses from https://cpclab.uni-duesseldorf.de/index.php/Software, and the CNA web server is accessible at https://cpclab.uni-duesseldorf.de/cna/. In **PUBLICATIONS II-IV**[35, 44, 45] constraint dilution trajectories **(section 2.2.2.1)** were visually inspected by VisualCNA.

2.2.2.1 Analyzing network states along constraint dilution trajectories

By gradually removing noncovalent constraints from an initial network representation of a biomolecule, a succession of network states {σ} is generated (constraint dilution trajectory). Analyzing such a trajectory by rigidity analysis reveals a hierarchy of rigidity that reflects the modular structure of biomolecules in terms of secondary, tertiary, and supertertiary structure[21, 22, 79, 119, 120]. In particular, constraint dilution allows simulating the loss of structural stability of a biomolecule with increasing temperature[121, 122]. For this, hydrogen bonds are removed from the constraint network if $E_{HB} > E_{cut,\sigma}$, where $\sigma = f(T)$ is the state of the network at temperature T and $E_{cut,\sigma 1} > E_{cut,\sigma 2}$ for $T_1 < T_2$ **(Figure 8A)**. Hydrophobic interactions are generally not removed along the constraint dilution trajectory because they remain constant in strength or become even stronger with increasing T. Alternatively, a modified method for accounting for the temperature dependence of hydrophobic interactions has been introduced that adds more constraints to the network with increasing temperature by linearly increasing the distance cutoff D_{cut}[112]. The hierarchy of rigidity of biomolecules leads to a percolation behavior that is often more complex than that of network glasses[93], and multiple phase transition points can be identified along the constraint dilution trajectory at which rigid clusters decompose **(Figure 8B)**[46]. The *rigidity percolation threshold* is then defined as the

phase transition when the network changes from an overall rigid to an overall flexible state and thus loses its ability to transmit stress[22].

Phase transitions can be related to the protein`s (thermo)stability **(section 2.2.2.3)**. Therefore, the computed E_{cut} values can be converted to a temperature T using the linear equation introduced by Radestock *et al*.[22] **(Eq. 4)**.

$$T = \frac{-20\ K}{kcal \cdot\ mol^{-1}}\ E_{cut} + 300\ K \qquad \textbf{Eq. 4}$$

In **PUBLICATIONS II-IV**[35, 44, 45] we applied **Eq. 4** to provide insights into a protein`s (thermo)stability during constraint dilution simulations.

2.2.2.2 Global and local indices for characterizing biomolecular stability

For having maximal advantage from rigidity analysis, the results need to be linked to biologically relevant characteristics of a structure. For this, CNA computes a comprehensive set of indices from the constraint dilution trajectory[105] **(section 2.2.2.1)**.

Global indices monitor the degree of flexibility and rigidity within constraint networks at the macroscopic level. They include the rigidity order parameter P_∞[123], which monitors the decay of the largest rigid cluster, the mean rigid cluster size S[124], which monitors the decay of all but the largest rigid cluster[123, 124], and the cluster configuration entropy H, a Shannon-type entropy[125] that is a morphological descriptor of the network heterogeneity[126] **(Figure 8B) (section 2.2.2.2.1)**.

Local indices characterize the network flexibility and rigidity down to the bond level. The percolation index p_i is a local analog to P_∞ and is most suitable to monitor the percolation behavior of a biomolecule locally[105]. The rigidity index r_i is a generalization of p_i as it monitors when a residue segregates from any rigid cluster[105] **(Figure 8C)**. Another set of local indices characterizes correlations of stability between pairs of residues[105]. As such, stability maps rc_{ij} are 2D generalizations of r_i[21] **(Figure 8D) (section 2.2.2.2.2)**. In addition, CNA computes *unfolding nuclei* as structural features from which macroscopic (in)stability originates[22] **(section 2.2.2.2.3)**. These can be used to predict structural *weak spots* for improving protein's stability.

The following sections focus on the indices that are used in **PUBLICATIONS II-IV**[35, 44, 45]. For further details about the other *global* and *local* indices see ref. [105].

2.2.2.2.1 Cluster configuration entropy

The following part was taken and adapted from **PUBLICATION II**[44]**.**

The cluster configuration entropy H_{type2} is a *global* index, which has been introduced by Radestock and Gohlke[22]. In **PUBLICATIONS II**[44] and **III**[45] H_{type2} is used to identify the phase transition temperature T_p at which a biomolecule switches from a rigid to a floppy state and the largest cluster stops to dominate the whole network. As long as the largest rigid cluster dominates the whole protein network, H_{type2} is low because of the limited number of possible ways to configure a system with a very large cluster. When the largest rigid cluster starts to decay or stops to dominate the protein network, H_{type2} jumps. There, the network is in a partially flexible state with many ways to configure a system consisting of many small clusters. The percolation behavior of protein networks is usually complex, and multiple phase transitions can be observed. In order to identify T_p, a double sigmoid fit is applied to an H_{type2} *versus* $T(E_{cut})$ curve as done previously[21, 22, 48, 112, 127], and T_p taken as that T value associated with the largest slope of the fit.

2.2.2.2.2 Stability maps

The following part was taken and adapted from **PUBLICATIONS II-IV**[35, 44, 45].

Since the percolation behavior of a protein network is complex due to the protein's structural hierarchy and composition of different modules, it is often challenging to assign a phase transition with H_{type2}. Thus, in **PUBLICATIONS II-IV**[35, 44, 45], in addition to using H_{type2}, we also characterized the hierarchy of rigid and flexible regions of wt*Bs*LipA at a *local* level by calculating stability maps.

The stability map rc_{ij} is a *local* index, which has been introduced by Radestock and Gohlke[21]. rc_{ij} represents the local stability within a protein structure for all residue pairs at which a rigid contact *rc* between two residues *i* and *j* (represented by their C_α atoms) is lost during the constraint dilution. *rc* exists if *i* and *j* belong to the same rigid cluster *c* of the set of rigid clusters $C^{E_{cut}}$[105]. Thus, rc_{ij} contains information cumulated over all network states along the constraint dilution trajectory as to which parts of the network are (locally) mechanically stable at a given σ, and which are not[127]. This stability information is not only available in a qualitative manner but also quantitatively in that each rc_{ij} has been associated with E_{cut} at

which the rigid contact is lost. The sum over all entries in rc_{ij} represents the chemical potential energy due to noncovalent bonding, obtained from the coarse-grained, residue-wise network representation of the underlying protein structure. To focus only on the stability of *rc* between structurally close residues, rc_{ij} was filtered such that only rigid contacts between two residues that are at most 5 Å apart from each other were considered (neighbor stability map $rc_{ij,neighbor}$). As done previously[127], to suppress the influence of extreme values in the double summation on the outcome of the unfolding energy, the median neighbor stability map $\tilde{rc}_{ij,neighbor}$ can be computed as the median of $rc_{ij,neighbor}$ averaged over the ensemble instead.

2.2.2.2.3 Unfolding nuclei

The following part was taken and adapted from **PUBLICATION II**[44].

Unfolding nuclei are represented by residues that percolate from the largest rigid cluster at the latest phase transition[22]. If such residues become flexible, it will have a detrimental effect on protein stability. Fringe residues of the *unfolding nuclei* percolate from the largest rigid cluster during earlier steps of the thermal unfolding. In **PUBLICATION II**[44], we follow the hypothesis that the more structurally stable the fringes of *unfolding nuclei* are, the more structurally stable will those *unfolding nuclei* be. Therefore, if such fringe residues (termed *weak spots*) are targeted by substitutions, the likelihood to stabilize the rigid core of a protein should be high. If two *unfolding nuclei* were only separated by one residue, this residue was also considered a *weak spot*. This procedure of identifying *weak spots* is in agreement with a previous study by us[22].

2.2.2.3 Applications of CNA

As in previous studies, monitoring the decay of network rigidity along a constraint dilution trajectory **(section 2.2.2.1)** generated by CNA was mainly used to provide insights into protein's thermostability. The following sections focus on these applications. Biomolecular thermostability can have a thermodynamic or kinetic origin[128]. Thermodynamic stability is a function of the change in free energy between the folded and unfolded state of a protein, whereas kinetic stability is determined by the height of the free energy barrier on the pathway of the time-dependent irreversible transition between folded and denatured state[128, 129]. In all studies reported below, rigidity analysis was used to investigate only the effect of mutations on the folded state. This was done because rigidity analysis cannot account for the time-dependency of processes[104], and it is very challenging to generate realistic structural models

of the unfolded state of a protein[130]. Still, applying rigidity analysis that way provides a wide range of applicability for studying thermostability because increased structural rigidity of the folded state is in 60% of the cases responsible for increased thermostability[48].

Initially, CNA has been applied to small-scale data sets of pairs of homologous proteins from psychrophilic to (hyper)thermophilic organisms **(section 2.2.2.3.1)**. Subsequently, series of protein variants were investigated **(section 2.2.2.3.2)**. However, CNA has not been applied to a large-scale data set of protein variants to investigate either protein thermostability or multiple types of protein stability. This is why I rationalize the impact of enzyme rigidity and flexibility on different protein properties with CNA at large-scale in **PUBLICATIONS II-IV**[35, 44, 45].

2.2.2.3.1 Constraint dilution simulations to investigate protein thermostability

Radestock *et al.*[21, 22] analyzed protein thermostability of pairs of homologous proteins from mesophilic and thermophilic organisms using CNA. The authors described the macroscopic percolation behavior and predicted T_p by monitoring H and P_∞ **(section 2.2.2.2)** during constraint dilution simulations **(section 2.2.2.3.1)**. The comparison between predicted T_p values and optimal growth temperatures of the corresponding organisms (T_{og}) revealed that in two-thirds of the pairs, a higher T_p was predicted for the thermophilic than for the mesophilic homolog[22]. At the microscopic level, the authors identified structural features from which a destabilization originates (abbreviated as *weak spots*), which is very helpful for guiding mutation experiments when prospectively engineering thermostability (see below). From both global and local stability characteristics the authors provided direct evidence for the 'principle of corresponding states,' according to which mesophilic/thermophilic homologs have similar flexibility and rigidity characteristics at the respective T_{og}[21, 22]. In addition, by monitoring the local distribution of flexible and rigid regions using rc_{ij} **(section 2.2.2.2)**, adaptive mutations in enzymes were shown to maintain the balance between global (structural) stability, in favor of overall thermostability, and local flexibility, in favor of activity, at appropriate enzyme working temperatures; this important information provides guidelines for what (not) to mutate in prospective studies[21].

Extending these study to series of protein variants, Rathi *et al.*[112] studied the relationship between structural rigidity and thermostability of citrate synthase (CS) from five different species with T_{og} ranging from 37°C to 100°C. CNA was applied to conformational ensembles generated by MD simulations **(section 2.2.2)**. The authors obtained a good correlation (R^2 =

0.88) between predicted T_p and experimental T_{og}. This finding validates that CNA is able to quantitatively discriminate between less and more thermostable proteins even within a series of orthologs. Furthermore, from a local point of view, the study revealed that structural weak spots predominantly occur at sequence positions with a high mutation ratio. Dick *et al.*[131] applied CNA to study the thermal adaptation of 2-deoxy-D-ribose-5-phosphate aldolase (DERA) originating from psychrophilic to hyperthermophilic organisms (T_{og} = 8 – 100°C). The comparison between predicted T_p and experimental T_{og} revealed a very good correlation (R^2 = 0.97). Interestingly, the authors identified, and validated by experiment, that interface stability contributes to thermostability in the dimeric DERA structures from (hyper)thermophilic organisms. This may be exploited as a design principle when engineering thermostability in multimeric proteins.

2.2.2.3.2 Prospective application to improve protein thermostability

With the aim to further develop CNA for prospective studies on improving thermostability, Rathi *et al.*[127] analyzed the thermodynamic stability of a set of 16 variants of *Bs*LipA. Three results stood out from this analysis. First, (relative) thermodynamic stability was successfully predicted for variants that differ by only 3–12 mutations from the wild type structure of *Bs*LipA (wt*Bs*LipA). Second, a measure for the similarity/dissimilarity of constraint dilution pathways of variants was introduced for explaining false thermostability predictions. Third, $\widetilde{rc}_{ij,neighbor}$ was introduced as a new local measure for predicting thermodynamic stability **(section 2.2.2.2.2)**. Additionally, the recently developed ENT^{FNC} approach[114] **(section 2.2.2)** was used for robust rigidity analysis, which makes it unnecessary to perform computationally demanding MD simulations for each variant.

In a subsequent prospective study, Rathi *et al.*[48] described a strategy to predict AA substitutions optimal for thermostability improvement; the predictions were experimentally validated. The strategy combines a structural ensemble-based weak spot prediction of wt*Bs*LipA by CNA, filtering of *weak spots* according to sequence conservation, computational SSM, assessment of variant structures with respect to their structural quality, and screening of the variants for increased structural rigidity by ENT^{FNC}-based CNA **(section 2.2.2)**. The strategy was applied to predict single-point variants of *Bs*LipA and yielded a success rate of 25% (60% when mutations from small-to-large residues and those in the active site were excluded) with respect to experimentally validated mutations that lead to increased

thermostability. Notably, an increase in thermostability by 6.6 °C compared to wt*Bs*LipA due to a single mutation was found.

As the prospective studies from Rathi *et al.*[48, 127] show that *Bs*LipA is suitable as model enzyme with respect to improving protein's thermostability based on CNA, I used *Bs*LipA for retrospective studies in **PUBLICATIONS II**[44] and **IV**[35].

2.3 Bacterial lipolytic enzymes as model enzymes

This section focuses on bacterial lipolytic enzymes that are used as model enzymes in **PUBLICATIONS II-IV**[35, 44, 45]. Based on their classification **(section 2.3.1)**, structure **(section 2.3.2)**, and industrial applications **(section 2.3.3)** insights will be provided into why they stand out in comparison to other enzymes. In particular, *Bs*LipA **(section 2.3.4)**, the model enzyme used in **PUBLICATIONS II**[44] and **IV**[35], will be described in more detail.

2.3.1 Classification of bacterial lipolytic enzymes

According to the Enzyme Commission (EC) bacterial lipolytic enzymes belong to hydrolases (EC 3) **(Figure 9A)** that irreversibly catalyze the cleavage of chemical bonds by addition of water under physiological conditions[132]. As the homeostasis of biomolecules, e.g., polysaccharides, DNA, proteins, and lipids, is essential for every living organism, hydrolases are ubiquitous in all three domains of life[132]. Bacterial lipolytic enzymes include carboxylesterases (EC 3.1.1.1; abbreviated as esterases/ESTs) and triacylglycerol hydrolases (EC 3.1.1.3; abbreviated as 'true' lipases/LIPs) **(Figure 9A)**, both of which I studied extensively in **PUBLICATIONS II-IV**[35, 44, 45]. ESTs hydrolyze solutions of water-soluble short acyl chain esters with < 10 carbon atoms and are mostly inactive against water-insoluble long chain triacylglycerols with ≥ 10 carbon atoms, which, in turn, are specifically hydrolyzed by LIPs **(Figure 9B)**[133-137]. Besides hydrolysis, other common reaction types are (trans/inter)esterification, alcoholysis, acidolysis, and aminolysis[138].

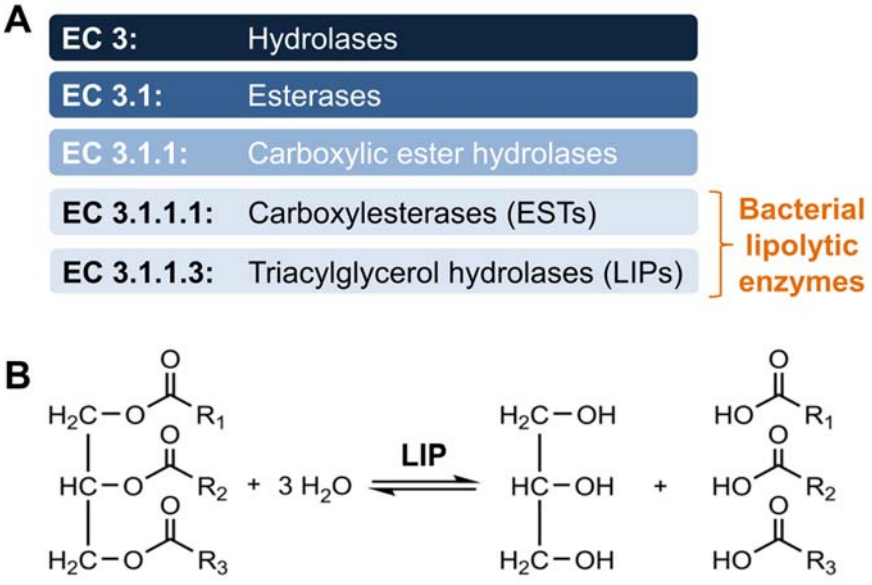


Figure 9: Classification of bacterial lipolytic enzymes according to the Enzyme Commission (EC) and lipase-catalyzed hydrolysis and esterification of triacylglycerol. (A) Hydrolases (EC 3) include carboxylesterases (EC 3.1.1.1; abbreviated as esterases/ESTs) and triacylglycerol hydrolases (EC 3.1.1.3;

abbreviated as 'true' lipases/LIPs). Together they are called 'bacterial lipolytic enzymes'. **(B)** LIPs hydrolyze triacylglycerol to form glycerol and long-chain fatty acids. The reverse reaction can also be carried out by esterification. The hydrocarbon chains are represented as R_1-R_3. Figure taken and adapted from Jaeger *et al.*[137].

Originally, LIPs were distinguished from ESTs based on kinetic terms of the phenomenon of interfacial activation at oil-water interfaces[139]. This phenomenon describes the activation of LIPs at high substrate concentrations beyond the critical micelle concentration (CMC). Hence, in contrast to ESTs, as the catalytic reaction of LIPs is not taking place in a homogenous phase, the classical Michaelis-Menten kinetic cannot be applied for LIPs[139, 140]. Instead, reaction kinetics of some LIPs follow sigmoid curves[140]. By determining the first three-dimensional structures of the fungal lipase from *Rhizomucor miehi*[141] and the human pancreas lipase[142], a flexible, amphipathic active site-covering α-helix, the so-called 'lid', was discovered, and a molecular explanation for interfacial activation was found. In short, upon interaction with the oil-water interface, the lid attains an 'open' conformation by structural changes resulting in the displacement of the lid from the active site[132, 134, 143]. Finally, the hydrophobic surface area surrounding the active site increases and the substrate can freely diffuse into the active site[132, 144]. However, due to the discovery of LIPs that show no correlation between their activity and neither interfacial activation nor the presence of a lid, both criteria were not able to appropriately distinguish ESTs and LIPs[132, 145-148]. Such exception is *Bs*LipA, the model enzyme used in **PUBLICATIONS II**[44] and **IV**[35], that does not possess a lid, and, hence, shows no interfacial activation[148, 149] **(section 2.3.4.2)**.

Due to the considerable increase of structural knowledge of bacterial lipolytic enzymes through the elucidation of many gene sequences and the resolution of numerous crystal structures **(section 2.3.2)**, today's most commonly used classification is based on phylogenetic criteria, conserved sequence motifs, and biological functions[132, 150-152]. Initially, Arpigny and Jaeger[150] classified 53 known bacterial lipolytic enzymes into eight families ($F_{EST/LIP}$), F_{I} to F_{VIII}. Later, numerous novel enzymes were added to these families and the classification was extended by eleven families, F_{IX} to F_{XIX}[132, 151, 152]. This classification simplifies the assignment of newly discovered bacterial lipolytic enzymes to the respective family[132]. Furthermore, biochemical properties of some bacterial lipolytic enzymes were correlated with the nature of the often extremophilic microorganism from which the respective enzyme was isolated[132]. This allows the identification of $F_{EST/LIP}$ with novel biocatalysts for industrial applications **(section 2.3.3)**. In addition, the classification enables us to predict important structural features, e.g., the identification of active sites, and secretion mechanism.

The model enzyme *Bs*LipA used in **PUBLICATIONS II**[44] and **IV**[35] was found in the largest $F_{EST/LIP}$, F_I, combining 'true lipases' in eight subfamilies, $F_{I.1}$ to $F_{I.8}$[132, 149-152]. *Bs*LipA belongs to $F_{I.4}$, the subfamily representing the smallest triacylglycerol LIPs with molecular weights (MW) of about 20 kDa[149, 150]. Especially in $F_{I.4}$, several LIPs of the Gram-positive genus *Bacillus*, e.g., *B. licheniformis*, *B. subtilis*, and *B. pumilis*, were identified[132]. In comparison to the conserved pentapeptide sequence Gly-X-Ser-X-Gly, where X denotes any AA, LIPs of $F_{I.4}$ contain an Ala at the first position. Furthermore, these LIPs reach the maximum activity at pH 10.0-11.5[153]. *Bs*LipA will be described later in more detail **(section 2.3.4)**.

Furthermore, the large-scale data set used in **PUBLICATION III**[45] contains ESTs that are mainly assigned to F_{IV}, the hormone-sensitive lipase (HSL) family. F_{IV} consists of several ESTs from distantly related prokaryotes including psychrophilic to thermophilic bacteria[132]. Most ESTs of F_{IV} show a striking AA sequence similarity to the mammalian HSL[154].

2.3.2 Structural insights into bacterial lipolytic enzymes

The importance of lipolytic enzymes can be seen by the collection of 4257 LIPs and 3121 ESTs in BRENDA (BRaunschweig ENzyme DAtabase)[155, 156]. Considering that among them only 350 LIPs and 273 ESTs are linked to primary literature shows that the majority of lipolytic enzymes have not been experimentally studied yet[132]. The analysis of lipolytic enzymes into taxonomic groups revealed that they are conserved among all three domains of life and mostly originate from microorganism[132].

Most of the bacterial lipolytic enzymes have a canonical α/β-hydrolase fold **(Figure 10)** with the conserved pentapeptide sequence Gly-X-Ser-X-Gly, where X denotes any AA[132, 157]. Moreover, a second large structural family of bacterial lipolytic enzymes shows a canonical α/β/α-hydrolase fold with a conserved active site motif Gly-Asp-Ser-Leu and only few bacterial lipolytic enzymes with a β-lactamase-like fold were found[132, 158, 159]. This section focuses on the canonical α/β-hydrolase fold because most of the bacterial lipolytic enzymes in **PUBLICATIONS II-IV**[35, 44, 45] have this fold.

The canonical α/β – hydrolase fold consists of a central β-sheet with eight β-strands (β1-β8), flanked by six α-helixes (αA-αF)[157, 160] **(Figure 10)**.

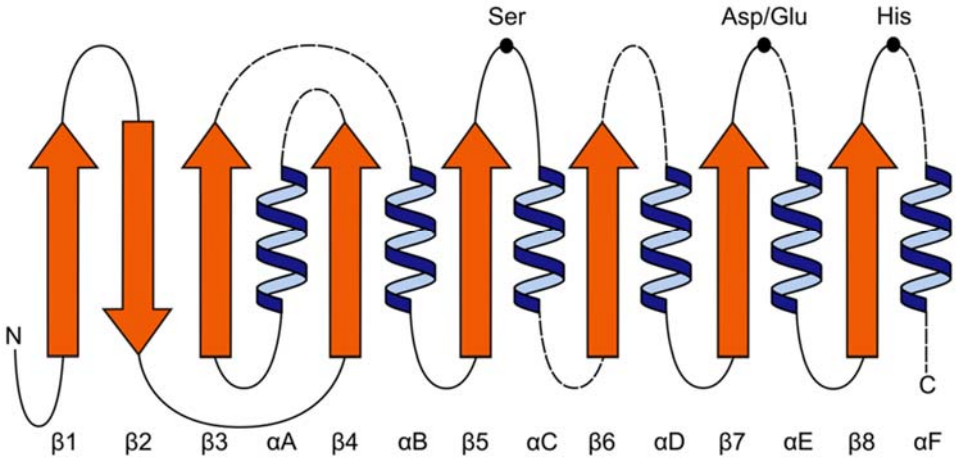


Figure 10: Schematic drawing of the α/β-hydrolase fold. Secondary structure topology of bacterial lipolytic enzymes showing an α/β-hydrolase fold with α-helices colored in dark blue (αA-αF) and β-strands (β1- β8) colored in orange. Broken lines indicate loops with variable lengths. The catalytic triad of Ser, Asp/Glu and His are shown as dots. Figure taken and adapted from Ollis *et al.*[157].

The β-sheet shows a parallel orientation, with the exception of the antiparallel orientated β2-strand. As the globular scaffold of this folding pattern is characterized by an extraordinary plasticity structural elements and even domains, e.g., the lid or cap, can be inserted into the loops connecting β-strands and α-helixes without disturbing the fold itself[132, 161, 162]. Besides the folding pattern, the active site is conserved, formed by a catalytic triad consisting of His, Ser and Asp/Glu[157, 163-165]. The nucleophilic Ser is located at the C-terminus of the β5-strand and part of the conserved pentapeptide sequence Gly-X-Ser-X-Gly, where X denotes any AA. This highly conserved pentapeptide forms a very sharp γ-turn called the 'nucleophilic elbow'[157]. Therefore, the nucleophilic Ser adopts energetically unfavorable backbone dihedral angles that lead to a surface-exposed position of the catalytic residue. The acidic residue Asp/Glu and His are situated in loop regions after the β7- and β8-strand[157]. The catalytic mechanism of lipolytic enzymes is essentially the same and comprises two steps based on the catalytic triad[157, 166]. Although most of the bacterial lipolytic enzymes show a canonical α/β-hydrolase fold, identifying the catalytic triad is not trivial. In **PUBLICATION IV**[35], we used structural knowledge together with the abovementioned classification **(section 2.3.1)** to unambiguously identify the active sites of the investigated bacterial lipolytic enzymes.

2.3.3 Industrial applications of bacterial lipolytic enzymes

Bacterial lipolytic enzymes constitute one of the most important and widely used classes of biocatalysts in the global industrial enzymes market. They are well established in many industrial applications for daily products, such as flavor development in the food industry, pitch control in the paper and pulp industry, as detergent additives in the laundry industry, and

for developing antibiotics as well as anti-inflammatory drugs in the pharmaceutical industry [135, 137, 166-169] **(Figure 11)**.

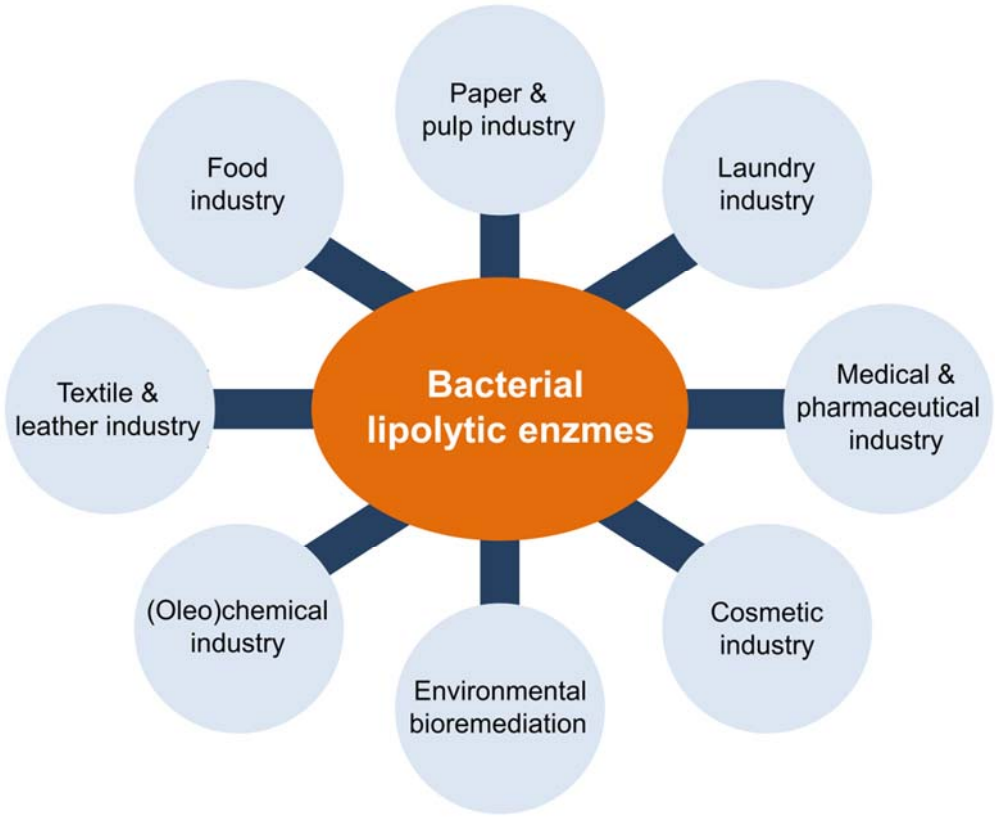


Figure 11: Industrial applications of bacterial lipolytic enzymes. Bacterial lipolytic enzymes are well established in many industrial applications for daily products.

The increasing demand for bacterial lipolytic enzymes is due to the fact that they are widely distributed in nature within microbial communities (at least one lipolytic enzyme is found in each bacterial genome) [34, 35, 44, 152, 170, 171]. They have been extensively examined with state-of-the-art (meta)genomics techniques and investigated by functional screenings compared to many other enzyme classes, and they exhibit high regio-, enantio-, and stereo-selectivity[34, 35, 44, 152, 170, 171]. In addition, most of the bacterial lipolytic enzymes do not require chaperons or cofactors and possess outstanding properties in terms of stability, promiscuity, reactivity, and scalability[34, 35, 44, 152, 170, 171]. Indeed, bacterial lipolytic enzymes are stable under harsh conditions, e.g., high temperatures, broad pH ranges, and the presence of detergents or ionic solvents[44, 172-174]. Another reason for the increasing demand of bacterial lipolytic enzymes in industrial applications is their substrate promiscuity. Their broad substrate spectra means that the production of multiple enzymes, which are specific to only a subset of substrates, is unneccessary[34]. Indeed, the market is dominated by highly versatile commercially available preparations such as the promiscuous Novozym 435 (N435), an immobilized preparation of lipase B from *Candida antarctica* (CalB), supplied by Novozymes[170, 175]. As the scope of their

catalyzed reactions in industrial applications is enormous, their heterologous and homologous production in large-scale fermentation processes becomes more and more attractive[35, 176]. Therefore, expression hosts with highly efficient secretion systems and high product yields are required. One example is *B. subtilis* that produces and secretes proteins in amounts of up to 25 g/l under optimal conditions[35, 176, 177] **(section 2.3.4.1)**. To achieve even more efficient secretion systems and high product yields for industrial applications, comprehensive optimization strategies at different stages of protein production and secretion have been developed[35, 178, 179].

2.3.4 *Bacillus subtilis* lipase A as model enzyme

2.3.4.1 The expression host *Bacillus subtilis*

The Gram-positive, aerobic, and spore-forming soil bacterium *B. subtilis* is one of the most important expression hosts for the production of homologous and heterologous proteins, especially in large-scale fermentation processes[35]. The characteristics of *B. subtilis* have been intensely studied over many years and, as a consequence, it was established as 'microbial cell factory'[176-178]. This is due to its known genome sequence[180] followed up by genome wide gene function analysis studies[181], its adaptability to continuously changing environments[182-184], its consideration as generally recognized as safe (GRAS) organism by the Food and Drug Administration (FDA), and its highly efficient secretion system with product yields of up to 25 g/l[176, 177]. In contrast to the well-known Gram-negative bacterium *Escherichia coli* (*E. coli*), *B. subtilis* lacks an outer cell membrane (OM), which contains lipopolysaccharides (LPS) representing endotoxins and are pyrogenic in humans and other mammals[185]. Additionally, *E. coli* is found in the human intestinal flora already in infants[186-188]. Moreover, in contrast to *E. coli*, *B. subtilis* secretes proteins directly into the extracellular medium[189]. From this follows that secreted proteins are naturally separated from cell components,simplifying downstream processing and enzyme production as well as preventing the formation of inclusion bodies[185, 190].

The majority of secretory proteins in *B. subtilis* are targeted to the Sec translocon and translocated via the cotranslational Sec-SRP pathway[191]. Alternatively, proteins can be secreted via the posttranslational Sec-SRP pathway, the twin-arginine translocation (Tat) pathway[192, 193], and several ATP-binding cassette (ABC) pathways[193, 194]. *Bs*LipA used as model enzyme in **PUBLICATIONS II**[44] and **IV**[35] follows the cotranslational Sec-SRP pathway[35].

2.3.4.2 Structural insights into *Bacillus subtilis* lipase A

With a MW of 19.34 kDa and 181 AAs *Bs*LipA, the model enzyme used in **PUBLICATIONS II**[44] and **IV**[35], is one of the smallest known 'true' LIPs[149] **(section 2.3.1)**. The characteristic folding pattern of *Bs*LipA is called minimal α/β-hydrolase fold[149]. In comparison to the common α/β-hydrolase fold **(section 2.3.2)** *Bs*LipA has no β1- and β2-strand and the αD-helix is replaced by a 3_{10}-helix **(Figure 12)**. With the help of a multiple sequence alignment of various microbial lipases, the residues of the catalytic triad were identified as Ser77, Asp133 and His156[149]. The first residue of the common lipase consensus sequence Gly-X-Ser-X-Gly, where X denotes any AA, is replaced by Ala75[149, 150]. Backbone amide groups of Ile12 and Met78 form the oxyanion hole that stabilizes the negatively charged transition state[149]. Like several other 'true' LIPs, e.g. LIPs from *Pseudomonas aeruginosa*[195] and *Pseudomonas glumae*[145], the active site of *Bs*LipA is not covered by a lid and, therefore, *Bs*LipA does not show interfacial activation at oil-water interfaces[149].

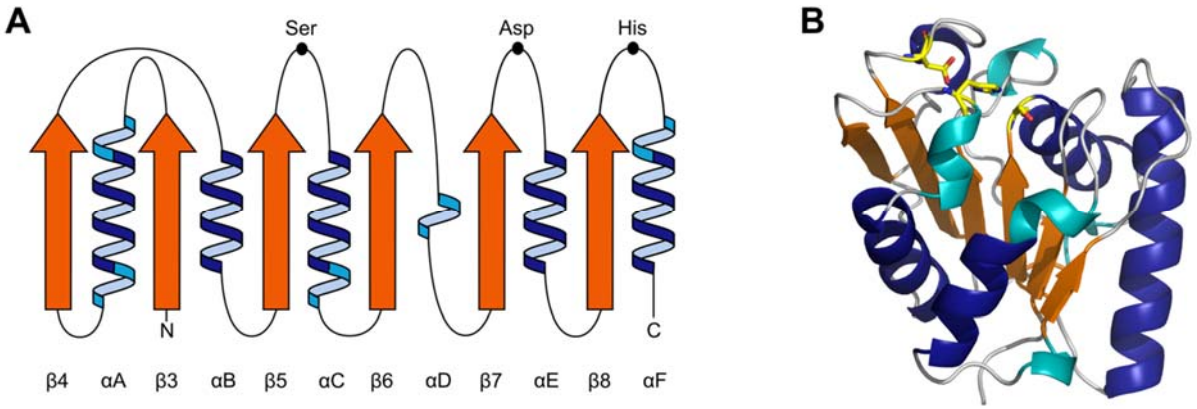


Figure 12: Minimal α/β-hydrolase fold of *Bs*LipA (PDB code: 1ISP). (A) Secondary structure topology and **(B)** three-dimensional cartoon-representation of *Bs*LipA with α-helices colored in dark blue, 3_{10}-helices colored in light blue, and β-strands colored in orange. The catalytic triad of *Bs*LipA consists of Ser77, Asp133, and His156 shown as **(A)** dots and **(B)** stick representation. Figure taken and adapted from van Pouderoyen *et al.*[149].

3 SCOPE OF THE THESIS

Nowadays, enzymes are becoming ever more ubiquitous in our daily lives because of their diverse applications such as in the food, detergent, and medical or pharmaceutical industries[4]. However, they do not always meet the required demands of industrial applications in terms of harsh environments, such as high temperatures or the presence of solvents and detergents[32, 33]. In addition, to make industrial applications more efficient, enzymes with a broad substrate spectrum and high product yields are preferred[34, 35]. Modern enzyme technology offers an increasing potential of a wide range of interdisciplinary processes for designing novel tailor-made enzymes according to human purposes[2]. Especially, protein engineering has emerged as a useful tool for developing novel tailor-made enzymes with improved properties **(section 2.1)**. However, most common are *knowledge-driven strategies* **(section 2.1.3)**, where the "knowledge" from information about the protein structure and / or sequence as well as computational techniques is combined with experiments[36-39]. However, as there is a lack of available experimental large-scale data measured in a uniform way the development and validation of algorithms for knowledge-driven strategies remain often unsatisfactory[40-43].

To address this issue, here, for the first time, I rationalized the impact of enzyme flexibility and rigidity on

- **I.** protein thermostability and / or detergent tolerance **(section 5, PUBLICATION II[44])**,
- **II.** substrate promiscuity **(section 6, PUBLICATION III[45])**,
- **III.** and expression **(section 7, PUBLICATION IV[35])**

using our in-house Constraint Network Analysis (CNA) software **(section 2.2.2)** at large-scale for biotechnologically highly relevant bacterial lipolytic enzymes **(section 2.3)**. This was done with the aim to define the scope and limitations of biomolecular flexibility predictions in knowledge-driven strategies for protein engineering.

The three tasks are related to increasing complexity in that, in the first, the solvent impact on protein thermostability is investigated, in the second, the impact of molecular recognition in the context of protein-substrate binding is scrutinized, and, in the third, the impact of protein production and secretion in a cellular context is analyzed.

4 PUBLICATION I

Rigidity theory for biomolecules: concepts, software, and applications

Hermans, S.M.A., Pfleger, C., Nutschel, C., Hanke, C.A., Gohlke, H.

WIREs Comput Mol Sci. 2017, 7, e1311.

Review, see pages 56-86 (Contribution: 20 %).

This publication was used to explain the basis of rigidity theory **(section 2.2)**.

5 PUBLICATION II

Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for *Bacillus subtilis* lipase A

Nutschel, C., Fulton, A., Zimmermann, O., Schwaneberg, U., Jaeger, K.-E., Gohlke, H.

J Chem Inf Model. 2020, 60, 3, 1568-1584.

Original publication, see pages 87-131 (Contribution: 60 %).

5.1 Background

Improving a protein's (thermo-)stability[21, 22, 48, 112, 120, 127, 131, 196] or tolerance against solvents[174, 197-203] and detergents[173, 204, 205] has become of utmost importance in protein engineering **(section 2.1)**. There are three general approaches for protein engineering: Rational design, directed evolution and knowledge-driven strategies **(section 2.1.3)**. Recent developments have tended towards knowledge-driven strategies, where available knowledge about the protein is used to identify substitution sites with a high potential to yield protein variants with improved stability and, subsequently, substitutions are engineered by mutagenesis studies[36, 41]. However, the development and validation of algorithms for knowledge-driven strategies has been hampered by the lack of availability of large-scale data measured in a uniform way and being unbiased with respect to substitution types and locations[40-43].

Here, with the objective to implement new guidelines for time- and cost-efficient protein engineering following a knowledge-driven strategy based on CNA[46] **(section 2.2.2)**, we scrutinized the impact of substitution sites on two types of protein stability for one protein at very large-scale. To do so, I systematically analyzed a complete experimental SSM library of the model enzyme *Bs*LipA **(section 2.3.4)**, which was evaluated as to thermostability (T_{50}) and detergent tolerance (D). Considering the screening results of the SSM library is important in view of the challenges of multi-dimensional property optimization of modern biocatalysts **(section 2.1)**. The measured T_{50} and D values provide valuable reference data for future analyses because, in contrast to other data sources[40-43], the different types of protein stability

were measured under respectively uniform conditions, such that there is no bias towards any particular substitution type or site. We set out to identify consistently defined *hot spot* classes for evaluating the performance of CNA.

5.2 Results and Discussion

The *Bs*LipA SSM library contained T_{50} as well as D data towards four detergents for all 3439 theoretically possible single variants (181 substitution sites of *Bs*LipA x 19 naturally occurring AAs). Across the SSM library, the likelihoods to find variants with significantly increased T_{50} (~12%) or D towards one detergent (~14%) are almost identical and small. Exemplarily, the distribution of T_{50} changes in *Bs*LipA variants is shown below **(Figure 13)**.

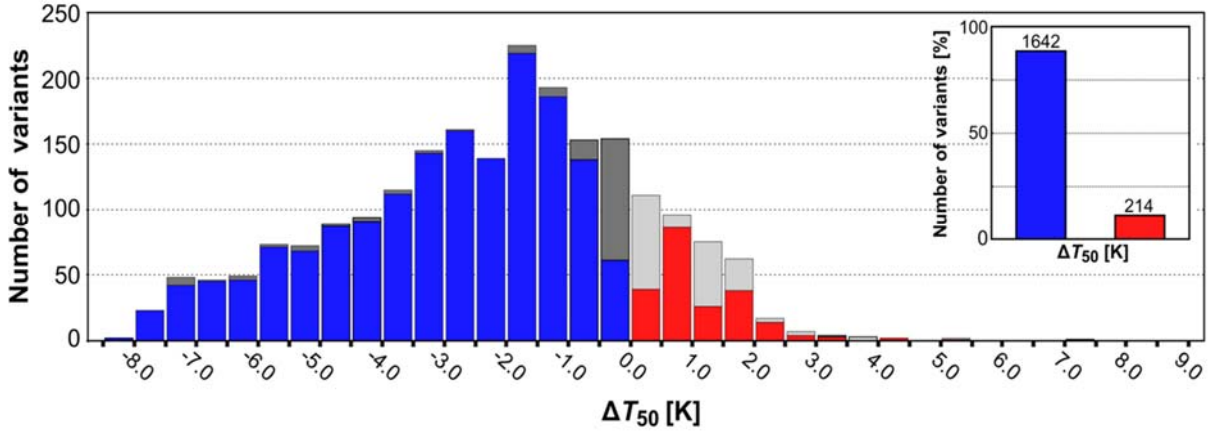


Figure 13: Distribution of *Bs*LipA variants' changes in T_{50}. Distribution of *Bs*LipA variants' changes in T_{50} (ΔT_{50}) compared to wt*Bs*LipA (ΔT_{50} = 0). Variants with ΔT_{50} lower than the experimental uncertainty (standard deviation σ_T for the respective variant) were excluded from further analyses (grey). The insets show the numbers of variants, which cause a significant in- or decrease in T_{50}. Figure was taken and adapted from **PUBLICATION II**[44].

The finding that the overwhelming number of single AA substitutions introduced by *random mutagenesis* causes a destabilizing effect is in agreement with previous studies[41, 206-209]. The identified largest increases in T_{50} of 7.7 K and D of 2.4 demonstrate that considerable improvements of protein stability can already be achieved by single AA substitutions. Hence, beyond the single T_{50} and D data, due to the completeness of our library and the model character of our protein, our results also constitute unbiased reference data as to what efficiency can be expected for a protein system when optimizing thermostability or detergent tolerance by *random mutagenesis*.

In the context of knowledge-driven protein engineering, I identified substitution sites for which variants yield significantly increased T_{50} or / and D. At most, and without considering the magnitude of the increase, only about one third or below of all *Bs*LipA residues constitute

such favorable substitution sites if T_{50} and D are considered separately, demonstrating that the location of a residue within a protein structure matters with respect to a substitution effect. In addition, I revealed for such substitution sites a significant and fair correlation between the frequency of T_{50} or / and D-increasing substitutions and the magnitude of the maximum effect. Together, these results show that addressing all substitution sites in an unbiased manner by *random mutagenesis* results in a considerable experimental effort coupled to low efficiency. In turn, identifying *a priori* substitution sites with a high likelihood for significantly increased T_{50} or D will also be beneficial with respect to the magnitude of effects that can be achieved there by substitutions.

This conclusion also holds if more than one type of protein stability is considered at a time. As such, I showed that at eleven substitution sites a ~4.6-fold higher likelihood to find for each detergent variants with significantly increased D compared to *random mutagenesis* is found. Additionally, seven substitution sites yield a ~3.4-fold higher likelihood to find significantly increased T_{50} and a ~4.7-fold higher likelihood to obtain for each detergent variants with significantly increased D compared to *random mutagenesis*. Hence, approaches that can identify substitution sites with a high likelihood for significantly increased T_{50} should also be beneficial for identifying substitution sites with a high likelihood for significantly increased D, or *vice versa*. This is an important finding for practical applications as many more algorithms have been developed to preferably address thermostability rather than detergent tolerance.

As another set of reference data, I defined *hot spot* classes from the previously identified substitution sites to provide benchmark data for evaluating the performance of CNA **(section 2.2.2)**. The first five classes follow the strict criterion that only the six substitution sites with the respective highest maximum effects of T_{50} (abbreviated as $\Delta T_{50;\,max}$) or D (abbreviated as ΔD_{max}) are considered **(Figure 14A)**. Accordingly, all combinations of the 20 proteinogenic AAs at such sites could still be experimentally tested[32, 39, 49, 71]. The intersections between the classes comprising the substitution sites with the broadest impact on ΔD_{max}, or $\Delta T_{50;\,max}$ and ΔD_{max}, are empty **(Figure 14A)**. Thus, I defined two additional classes with the somewhat relaxed criterion that the comprised substitution sites show significantly increased D towards each detergent, or significantly increased T_{50} and D towards each detergent, regardless of the magnitude of the single effect **(Figure 14B)**.

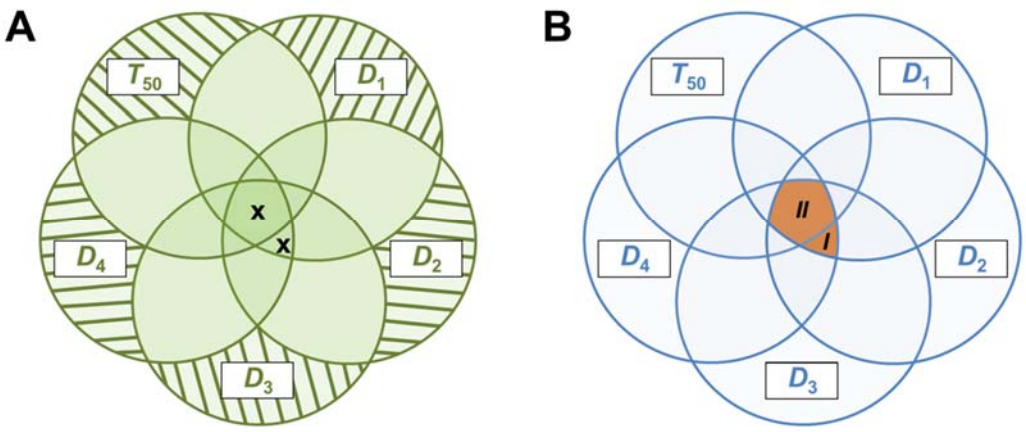


Figure 14: Overview of *hot spot* classes. (A) Five *hot spot* classes follow the strict criterion that only the six substitution sites with the respective highest maximum effects of T_{50} (abbreviated as $\Delta T_{50;\,max}$) or D (abbreviated as ΔD_{max}) are considered (shaded areas). The intersections comprising the substitution sites with the broadest impact on ΔD_{max}, or $\Delta T_{50;\,max}$ and ΔD_{max}, are empty (areas with crosses). **(B)** Two *hot spot* classes with substitution sites showing significantly increased D towards each detergent (orange area numbered as I), or significantly increased T_{50} and D towards each detergent (orange area numbered as II), regardless of the magnitude of the single effect.

I used the complete, unbiased, and uniformly generated T_{50} and D data to probe if universal rules for protein engineering can be established. I thereby focused on using "one-dimensional" descriptors in terms of location in secondary structure elements, degree of burial, physicochemical properties, and conservation degree of substituted AA. Notably, considering my descriptors, many (up to 98 substitution sites) predicted *hot spots* result, which would require considerable experimental efforts particularly if beneficial substitutions need to be accumulated to reach a desired effect. This finding demonstrates on a single protein level that, with the use of these descriptors, no universal and sufficiently discriminating rule(s) can be identified, a finding that is mirrored in other studies across protein families[210, 211] and with respect to low successes in assessing thermostabilities[212]. Still, if a higher number of predicted *hot spots* is acceptable, partially solvent-exposed residues are good *hot spot* candidates, whereas loop positions show mostly destabilizing effects. In addition, *hot spots* were preferentially found at both non-conserved and semi-conserved position. This finding may help refining future consensus concepts where multiple sequence alignments are used to preferentially substitute non-consensus residues by consensus ones.

Finally, I made use of the reference data to unequivocally benchmark CNA with respect to predicting *hot spots* as structural *weak spots* of the protein. With this respect, a constraint dilution simulation of wt*Bs*LipA was carried out with CNA on ENT generated from MD simulations **(section 2.2.2)** to predict major phase transitions at which the network switches from overall rigid to flexible states **(Figure 15)**.

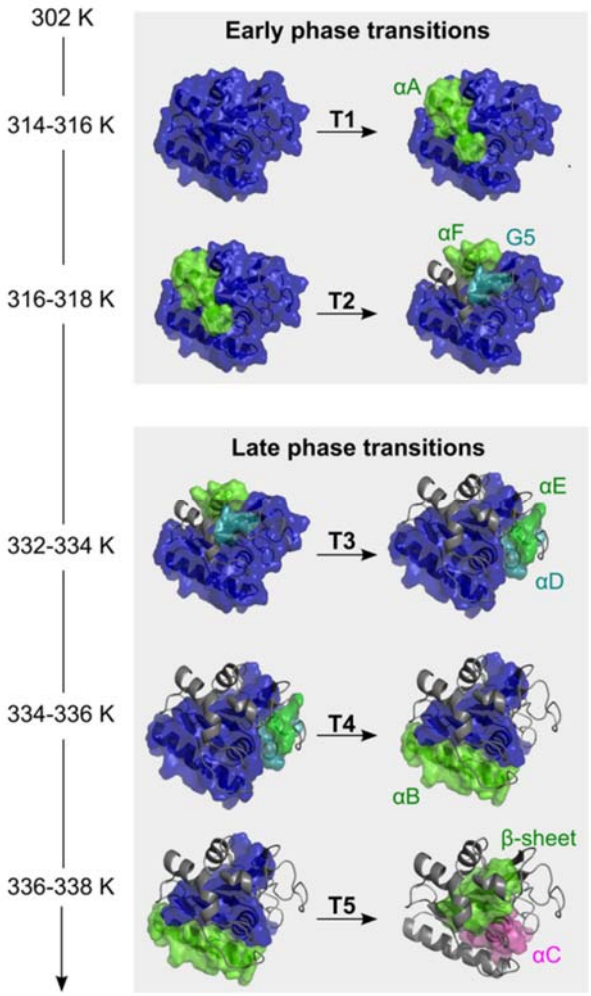


Figure 15: Prediction of the constraint dilution pathway of wt*Bs*LipA. Constraint dilution pathway of wt*Bs*LipA (PDB ID: 1ISP) showing the early (T1 – T2) and late (T3 – T5) phase transitions. CNA was carried out on ENT of wt*Bs*LipA generated by MD simulations. Rigid clusters are represented as uniformly colored blue, green, magenta, and cyan bodies in the descending order of their sizes. Figure was taken and adapted from **PUBLICATION II**[44].

From the constraint dilution pathway of wt*Bs*LipA, five major phase transitions, T1 – T5, were predicted based on the *global* index H_{type2} **(section 2.2.2.2.1) (Figure 15)**. In addition to using H_{type2}, we also characterized the hierarchy of rigid and flexible regions of wt*Bs*LipA at a *local* level by computing $rc_{ij,neighbor}$ **(section 2.2.2.2.2)**. $rc_{ij,neighbor}$ demonstrates that the rigid contacts between neighboring residues are stronger at the *N*-terminus than at the *C*-terminus along the contraint dilution simulation, i.e., the *C*-terminus of wt*Bs*LipA starts to unfold first. We confirmed the unfolding pathway of wt*Bs*LipA predicted by CNA with the independent Markov Chain Monte Carlo (MCMC)-based Protein Folding and Aggregation Simulator (ProFASi) approach[213, 214].

Finally, from a practical point of view, it is relevant that CNA predicted only ten *weak spots* **(Figure 16)**, allowing to focus subsequent substitution efforts on only ~6% of the protein

residues. Furthermore, the gain in precision over random classification is between ~3 and ~9, depending on the *hot spot* class. These results indicate that applying CNA-based *weak spot* predictions before attempting experimental engineering is beneficial, in particular if the number of substitution sites that can be dealt with in experiment is low.

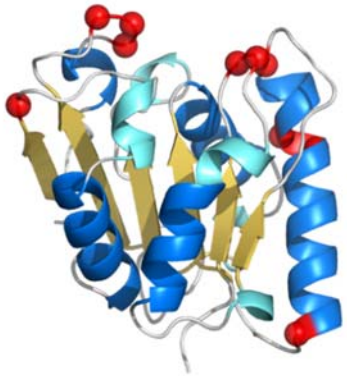


Figure 16: Localization of CNA-predicted *weak spots* of wt*Bs*LipA. Ten *weak spots* were predicted by CNA on ENT of wt*Bs*LipA (PDB ID: 1ISP) generated from MD simulations (red spheres). Figure was taken and adapted from **PUBLICATION II**[44].

5.3 Conclusion and Significance

In this study, for the first time, we performed a systematic large-scale analysis of a complete experimental SSM library of *Bs*LipA to scrutinize the impact of substitution sites on two types of protein stability with CNA.

The principle results of this study are:

- The SSM library provides systematic and unbiased reference data at unprecedented scale for engineering *Bs*LipA towards improved T_{50} or / and D.
- The identification of consistently defined *hot spot* types enables the evaluation of the performance of knowledge-driven strategies.
- CNA yields *hot spot* predictions with an up to 9-fold gain in precision over *random classification*.

The results suggest that knowledge-driven strategies based on CNA could be used prior to experiments when seeking to optimize enzymes' thermostability and detergent tolerance.

6 PUBLICATION III

Promiscuous esterases counterintuitively are less flexible than specific ones

Nutschel, C., Coscolín, C., Mulnaes, D., David, B., Ferrer, M., Jaeger K.-E., Gohlke, H.

J Chem Inf Model. 2020, DOI: 10.1021/acs.jcim.1c00152.

Original publication, see pages 132-200 (Contribution: 60 %).

6.1 Background

The universe of promiscuous activities available in nature has been suggested to be enormous[215, 216]. Understanding mechanisms of promiscuity thus has become increasingly important both from a fundamental and an application point of view[217, 218]. As to enzyme structural dynamics, more promiscuous enzymes generally have been recognized to also be more flexible[19523-25, 219]. However, examples for the opposite have received much less attention, although conformational changes may have been selected in evolution for their ability to enhance recognition specificity[20].

In this study, we exploit previously described comprehensive experimental information on the substrate promiscuity ($P_{EST/LIP}$) of 147 ESTs/LIPs tested against a customized library of dissimilar esters[34]. Here, $P_{EST/LIP}$ means that an EST/LIP carries out its typical catalytic function on non-canonical substrates, in that experimental conditions had been kept constant for the assessment of the different enzyme/ester combinations. I used computationally efficient rigidity analyses based on CNA **(section 2.2.2)** to understand the structural origin of and to predict $P_{EST/LIP}$.

6.2 Results and Discussion

The present study builds on one of the still few experimental large-scale datasets on enzyme promiscuity generated by Ferrer *et al.*[34]. The authors experimentally investigated $P_{EST/LIP}$ of 147 ESTs/LIPs (termed *experimental data set*) against 96 esters. Additionally, they ranked (classified) $P_{EST/LIP}$ of 96 ESTs/LIPs (termed *volume data set*) based on a newly introduced structural parameter, the active site effective volume (Vol_{eff}), which will be used here as a

reference to compare the power of $P_{EST/LIP}$ predictions based on CNA **(section 2.2.2)**. As our computational approach involves extensive MD simulations for generating large conformational ensembles, I selected 35 ESTs/LIPs from the *volume data set* (termed *flexibility data set*) based on the following criteria **(Figure 17)**: I.) The data set contains ESTs/LIPs with known and unknown crystal structures. That way, we probe to what extent the source of structural information influences the outcome of our results. II.) The chosen ESTs/LIPs of the data set show high diversities as to $P_{EST/LIP}$ and association to ESTs/LIPs families ($F_{EST/LIP}$, as defined by Arpigny and Jaeger[150]), similar to those found for the *volume data set*. III) Only ESTs/LIPs with AA sequence identities ≥ 25% in comparison to any existing crystal structure were considered in order to ensure a sufficient quality of generated comparative models. Finally, in order to uniformly depict the results across the present study, six EHs were selected as representatives of the *flexibility data set* based on $P_{EST/LIP}$ (termed *representative data set*): ESTs/LIPs with the lowest (EST/LIP115) or highest $P_{EST/LIP}$ (EST/LIP001) and known crystal structures, ESTs/LIPs with the lowest (EST/LIP127) or the highest $P_{EST/LIP}$ (EST/LIP005) and unknown crystal structures, and commercial ESTs/LIPs with the lowest (CalA) or highest $P_{EST/LIP}$ (CalB).

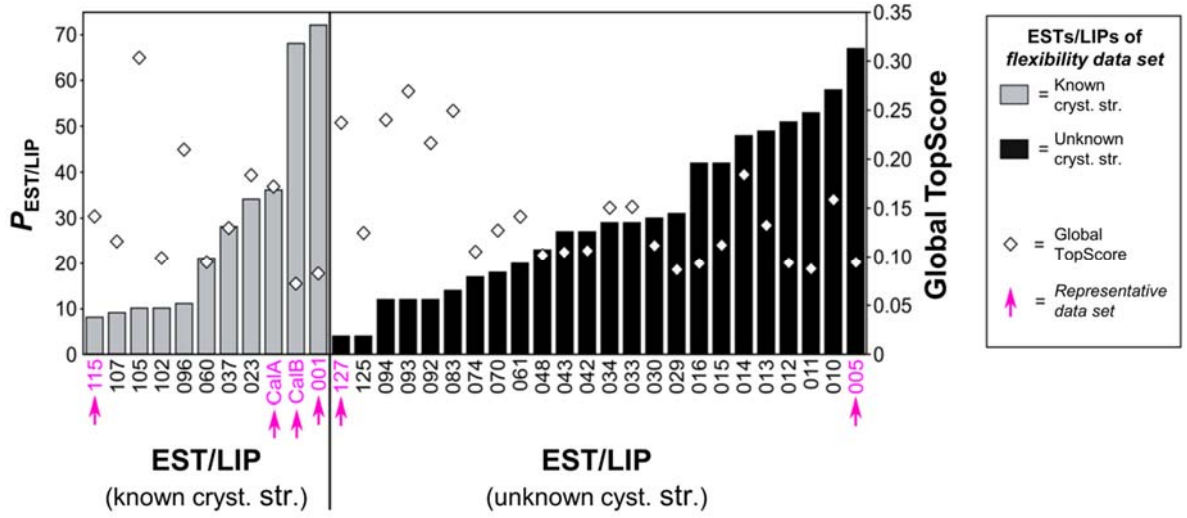


Figure 17: Comparative modeling of ESTs/LIPs. Based on sequence data provided by a large-scale study from Ferrer *et al*[34], comparative models were generated for 35 ESTs/LIPs with known (left, 11 ESTs/LIPs) and unknown (right, 24 ESTs/LIPs) crystal structures using TopModel[220]. These ESTs/LIPs constitute the *flexibility data set*. The ESTs/LIPs vary in $P_{EST/LIP}$ (left ordinate, bars) and global TopScores[221] (right ordinate, diamonds). Six ESTs/LIPs were selected as representatives of the *flexibility data set* (termed *representative data set*) as indicated by magenta arrows. Figure was taken and adapted from **PUBLICATION III**[45].

Comparative models of the *flexibility data set* were generated using our in-house structure prediction meta-tool TopModel[220]. TopModel uses multiple state-of-the-art threading and sequence/structure alignment tools to generate a large ensemble of models from different

pairwise and multiple alignments of the top five highest ranked template structures. The quality of the comparative models of the *flexibility data set* was assessed with TopScore[221], a meta Model Quality Assessment Program (meta-MQAP). TopScore uses deep neural networks (DNN) to combine scores from 15 different primary MQAP to predict accurate residue-wise and whole-protein error estimates. The models showed both an overall and residue-wise good structural quality. Additionally, we validated that catalytically active residues (CARs) in all models are accessible for substrates according to CAVER results

Previous studies indicated that enzyme flexibility influences the substrate promiscuity of enzymes[23-25, 219]. In order to investigate if the global flexibility of the EHs influences $P_{EST/LIP}$, I applied CNA to the *flexibility data set* and predicted T_p, the phase transition temperature previously applied as a measure of structural stability of a protein **(section 2.2.2.3)**. A good and significant correlation between T_p and $P_{EST/LIP}$ was found for the *flexibility data set* (R^2 = 0.60, p = 5.4*10^{-8}) **(Figure 18)**. These findings demonstrate that promiscuous ESTs/LIPs are globally less flexible.

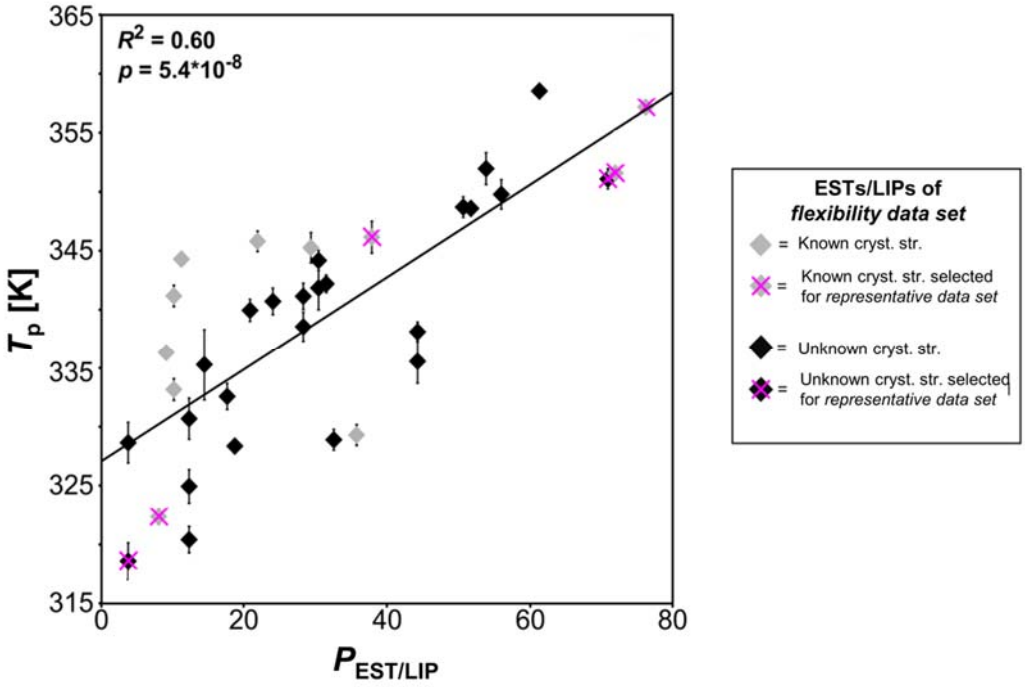


Figure 18: Correlation of T_p *versus* $P_{EST/LIP}$. (A) Correlation between predicted T_p based on the global index H_{type2} and P_{EH} for the *flexibility data set*. Data points colored grey (black) represent comparative models of ESTs/LIPs with (un)known crystal structures. The *representative data set* is indicated by magenta crosses. Error bars show the SEM over five independent MD simulations of 1 µs length each. Figure was taken and adapted from **PUBLICATION III**[45].

The good correlation of $P_{EST/LIP}$ and T_p encouraged us to investigate if local flexibility characteristics of CARs will provide an even better predictor of $P_{EST/LIP}$. With this respect, I

thus computed a parameter called $Flex_{CAR}$ for the *flexibility data set*. This parameter quantifies the stability of rigid contacts between CARs and other residues that are at most 5 Å apart from each other, based on the *local* index $rc_{ij,neighbor}$ **(section 2.2.2.2.2)**. A good and significant correlation between $Flex_{CAR}$ and $P_{EST/LIP}$ was found for the *flexibility data set* (R^2 = 0.51, $p = 1.7*10^{-6}$). Hence, promiscuous ESTs/LIPs tend to have less flexible CARs. Mobility characteristics computed directly from MD trajectories show the same trends, although the correlation with $P_{EST/LIP}$ is insignificant. Throughout our study, we probed for the consistency of our analyses between subsets of ESTs/LIPs for which either crystal structures are known or not; we only found quantitative differences, but no qualitative ones. One of the reasons is likely that CNA was carried out on ENT generated by multiple and μs-long MD simulations, which markedly increases the robustness of the results **(section 2.2.2)**.

Previous studies indicated that thermodynamically more thermostable proteins frequently have a higher structural stability. We used experimental melting temperatures of ESTs/LIPs determined by CD spectroscopy as indicators for enzyme flexibility. This experimental data led to the same conclusion with respect to $P_{EST/LIP}$ as the one drawn from the computed flexibility predictions, i.e., promiscuous ESTs/LIPs are not only globally less flexible but also more thermostable. Overall, these consistent and robust findings indicate that when applying this workflow to novel ESTs/LIPs, it should be possible to discover enzymes with 'sufficient' substrate promiscuity to serve as a starting point for further exploration in biotechnology and synthetic organic chemistry. In that respect, the flexibility characteristics of ESTs/LIPs analyzed here have a notably stronger predictive power than Vol_{eff} introduced earlier.

The finding that promiscuous ESTs/LIPs are significantly globally *less* flexible and have *less* flexible CARss than specific ESTs/LIPs is in stark contrast to the general view of the role of structural flexibility for promiscuity[23-25, 219]. It has been recognized that conformational changes may not always be necessary for promiscuity if a variety of substrates can be bound by partial recognition or the presence of multiple binding sites[218]. However, these cases do not seem to be relevant reasons for $P_{EST/LIP}$ because partial recognition often is associated with catalytic inefficiency[222], which is contrary to our observation that promiscuous ESTs/LIPs have a significantly increased specific activity. In addition, the presence of multiple binding sites for $P_{EST/LIP}$ is controverted by the finding that promiscuous ESTs/LIPs have large Vol_{eff}, i.e., large pockets with few subpockets. Inversely, our findings of rigid promiscuous ESTs/LIPs may be consistent with the idea that multiple ligands can be accommodated in a single site by exploiting diverse interacting residues.

Our results as to *specific but flexible* ESTs/LIPs may be reconciled with a model according to which conformational changes may have been selected in EST/LIP evolution for their ability to enhance specificity in recognition, resulting in what has been termed conformational proofreading[20]. In the case of specific ESTs/LIPs, flexibility may help to overcome a structural mismatch between the enzyme and its substrate existing when both are in their ground states, that way enhancing recognition specificity. This view is corroborated by our finding that specific ESTs/LIPs prefer to hydrolyze large and flexible substrates: Larger substrates can form more interactions with the enzyme, that way helping to overcome the deformation energy required by the enzyme to optimizing the correct binding probability over the incorrect one; flexible substrates can tolerate higher strains and thus can be expected to participate in more binding events[223, 224].

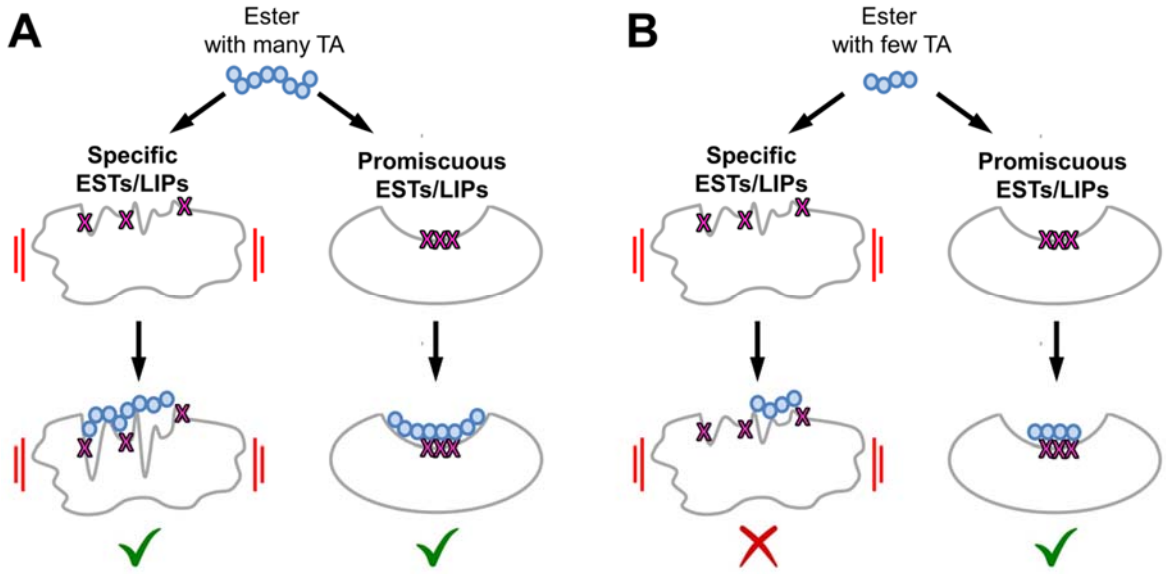


Figure 19: Mechanistic model of EST/LIP flexibility, ligand size and conformational dynamics affecting P_{EH}. Impact of esters with **(A)** many or **(B)** few TA on specific, and hence more flexible **(left)**, and promiscuous, and hence more rigid **(right)** LIPs. Ligand parts connected by TA are represented as blue circles. Specific ESTs/LIPs and large ligands with many TA can mutually adapt **(panel A, left)**, and promiscuous EST/LIP can bind large ligands **(panel A, right)** and small ligands **(panel B, right)** exploiting different interaction partners. Small (and/or rigid) ligands are not able to lead to a structural adaptation of specific ESTs/LIPs **(panel B, left)**, though, resulting in conformational proofreading. The red bars indicate the flexibility of the ESTs/LIPs. A green tick (red cross) indicates that ester cleavage is (not) catalyzed. Figure was taken and adapted from **PUBLICATION III**[45].

6.3 Conclusion and Significance

In this study, we exploit previously described comprehensive experimental information on $P_{EST/LIP}$ of 147 ESTs/LIPs tested against 96 esters together with computationally efficient rigidity analyses based on CNA to understand the structural origin of and predict $P_{EST/LIP}$.

The principle results of this study are:

- Promiscuous ESTs/LIPs are significantly globally less flexible, have less flexible CARs than specific ones, are significantly more thermostable, and have a significantly increased specific activity.
- Specific ESTs/LIPs prefer to hydrolyze large and flexible esters.

These results may be reconciled with a model according to which multiple ligands can be accommodated in a single site of promiscuous ESTs/LIPs by exploiting diverse interacting residues, whereas structural flexibility in the case of specific ESTs/LIPs serves for conformational proofreading. Our results furthermore signify that EST/LIP sequence space, charted, e.g., by (meta)genomics studies, can be screened by rigidity analyses based on CNA for promiscuous ESTs/LIPs that may serve as starting points for further exploration in biotechnology and synthetic chemistry.

7 PUBLICATION IV

Contribution of single amino acid and codon substitutions to the production and secretion of a lipase by *Bacillus subtilis*

Skoczinski, P., Volkenborn, K., Fulton, A., Bhadauriya, A., Nutschel, C., Gohlke, H., Knapp, A., Jaeger, K.-E.

Microb Cell Fact. 2017, 16, 160.

Original publication, see page 201-230 (Contribution: 10 %).

7.1 Background

Due to the fact that *B. subtilis* produces and secretes proteins in amounts of up to 20 g/l under optimal conditions, it has been intensively studied and optimized as a protein production host, establishing it as a *microbial cell factory*[176-178, 185] **(section 2.3.4.1)**. However, protein production can be challenging if transcription and cotranslational secretion are negatively affected, or the target protein is degraded by extracellular proteases[178, 225]. Here, we aim to elucidate the influence of a target protein on its own extracellular activity and amount by a systematic analysis of the homologous model enzyme *Bs*LipA **(section 2.3.4).** Therefore, a nearly complete SSM library of *Bs*LipA was generated and about 30000 clones were qualitatively as well as quantitatively screened with respect to extracellular activity and amount. Variants with beneficial effects were sequenced and analyzed with respect to *B. subtilis* growth behavior, extracellular activity and amount as well as *lipA* transcription. In order to determine to what extent an increase in (thermo)stability could contribute to an increased extracellular amount, I predicted differences in the thermodynamic thermostability of variants with respect to wt*Bs*LipA by constraint dilution simulations using CNA[46] **(section 2.2.2)**.

7.2 Results and Discussion

In total, 155 AA residues of *Bs*LipA with a conservation < 95% were used to generate a nearly complete SSM library resulting in about 30,000 clones **(Figure 20A)**. To identify variants with increased extracellular activity or amount, a two-step screening procedure was applied to the SSM library **(Figure 20A)**. In the first step, the about 30,000 clones were

analyzed towards increased extracellular activity by a lipase activity assay in the culture supernatants. 175 clones were sequenced and 80 variants showed an increase in extracellular activity from 1.2- to 3.4-fold in comparison to wt*Bs*LipA **(Figure 20B)**. In the second step, the culture supernatants of these variants were analyzed as nine biological replicates. Extracellular activity was determined by a lipase activity assay and extracellular amount was quantified by an enzyme-linked immunosorbent assay (ELISA). 38 variants showed an increased or similar extracellular activity and an increased extracellular amount compared to wt*Bs*LipA. Their extracellular amount ranged from 1.3-fold to 3.8-fold higher than that of wt*Bs*LipA **(Figure 20C)**.

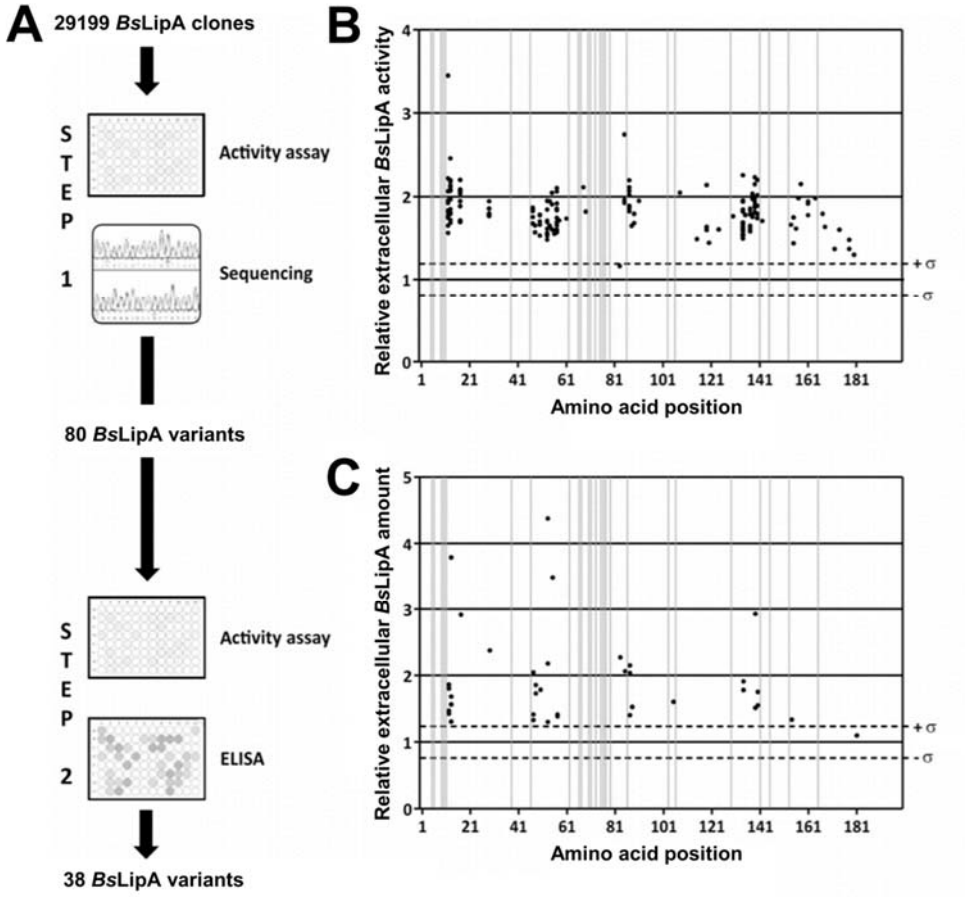


Figure 20: Identification of *Bs*LipA variants with increased extracellular activity or amount. (A) Schematic representation of the two-step screening procedure. In the first step, 29199 clones were analyzed for increased extracellular activity by a lipase activity assay in the culture supernatant. 175 clones were sequenced and 80 variants identified with increased extracellular activity. In a second step, culture supernatants of these variants were analyzed as nine biological replicates. Extracellular activity was determined by a lipase activity assay and extracellular amount was quantified by an enzyme-linked immunosorbent assay (ELISA). **(B)** 80 variants with increased extracellular activity. The relative extracellular activity of the variants is plotted against the substituted AA position. **(C)** 34 variants with increased extracellular amount. The relative extracellular amount is plotted against the substituted AA position. Each black dot represents one variant, and the grey bars mark the highly conserved AA positions (≥ 95%). Values for wt*Bs*LipA, which were **(B)** 0.57 ± 0.12 U/ml and **(C)** 3.7 ± 0.6 µg/ml, respectively, were set to 1 and the grey horizontal dotted lines mark the standard deviation (σ). Figure taken and adapted from **PUBLICATION IV**[35].

Next, we produced these variants by cultivating *B. subtilis* clones in a microfermentation system linked to online biomass measurement and analyzed their extracellular activity and amount as well as *lipA* transcription. Furthermore, online biomass measurements were performed to exclude differences in growth of variant-producing *B. subtilis* clones, which was, however, not observed.

We identified six variants with an up to 2.4-fold increase in extracellular activity **(Figure 21A)** and 21 variants with an up to 2.3-fold increase in extracellular amount in comparison to wt*Bs*LipA **(Figure 21B)**. In addition to single AA substitutions increasing extracellular activity and amount, several codon-related effects were observed. For example, the variants I12L$_{CTG}$, I12V$_{GTG}$, and G13T$_{ACC}$ showed an increase in extracellular activity, whereas identical AA substitutions encoded by different codons either showed no effect on extracellular activity and amount (I12L$_{TTG}$ and I12V$_{GTC}$) or resulted in increased extracellular amount (G13$_{ACG}$) **(Figure 21A)**. Another example is that variant I87I with a silent mutation showed a 2.4-fold increase in extracellular activity but also a 3.6-fold significant change in *lipA* transcript level **(Figure 21A)**.

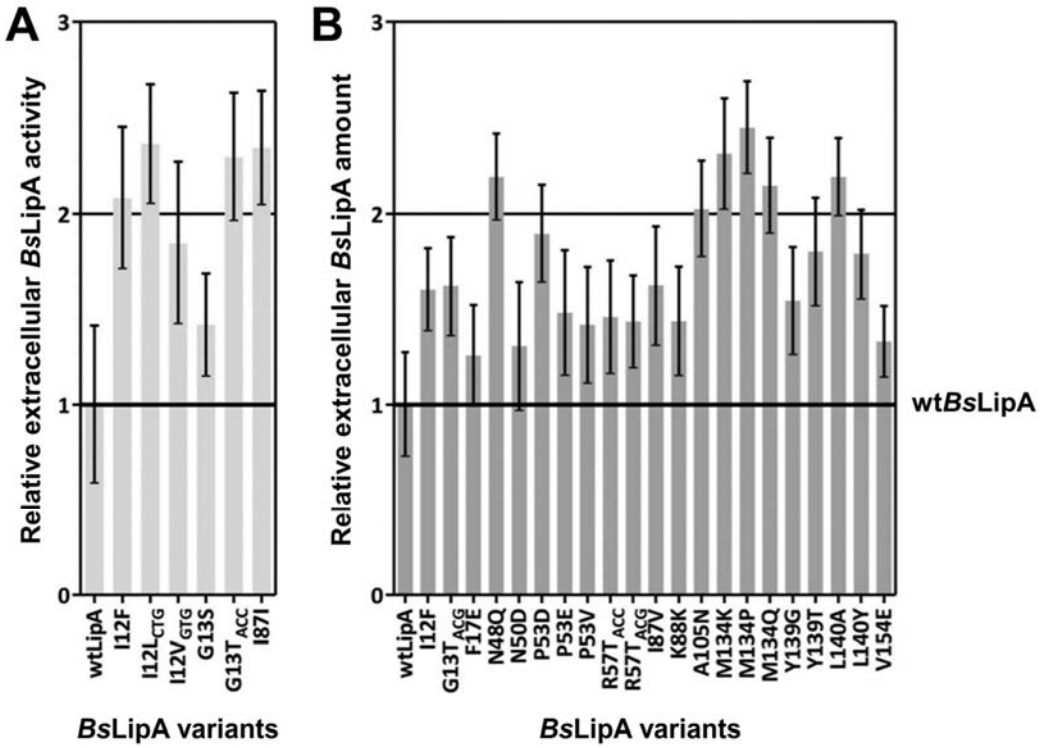


Figure 21: *Bs*LipA variants showing increased extracellular activity or amount. (A) Six variants with increased extracellular activity and **(B)** 21 variants with increased extracellular amount. Variants were produced by cultivating *B. subtilis* clones in a microfermentation system linked to online biomass measurement. **(A)** Extracellular activity was determined by a lipase activity assay. The relative extracellular amount is plotted against the respective variant. **(B)** Extracellular amount was quantified by an enzyme-linked immunosorbent

assay (ELISA). The relative extracellular amount is plotted against the respective variant. Respective wt*Bs*LipA values were set to 1 (thick black line). Figure taken and adapted from **PUBLICATION IV**[35].

Seven variants with increased extracellular amount have AA substitutions located either in the αB-helix (N50D, P53D, P53E, P53V, R57T$_{ACC}$, R57T$_{ACG}$), or carry a substitution to glutamine at position 134 (M134Q**) (Figure 21C)**. Position 134 is known to contribute to thermostability[226], and the αB-helix also plays an important role in tolerance towards detergents[173] and ionic liquids[174]. Therefore, it is possible that the increased extracellular amount of these variants is not due to a more efficient secretion, but due to an increased stability in the culture supernatant.

In order to determine to what extent an increase in (thermo)stability could contribute to an increased extracellular amount I predicted thermodynamic thermostabilities of the six variants N50D, P53D, P53E, P53V, R57T, and M134Q by constraint dilution simulations using CNA[46] **(section 2.2.2)**. Here, as done previously for *Bs*LipA[127], the thermodynamic thermostabilities of the variants were compared to wt*Bs*LipA in terms of a local index, the median neighbor stability map $\tilde{rc}_{ij,\ neighbor}$ **(section 2.2.2.2.2)**. $\tilde{rc}_{ij,\ neighbor}$ has been shown to be related to the experimental melting temperature (T_m) and to be robust if variants follow different constraint dilution pathways[127] **(section 2.2.2.3.1)**. While for three variants, i.e., P53D, P53E, P53V, marginal changes in the predicted thermostability compared to wt*Bs*LipA were found, a pronounced decrease in the thermostability was predicted for the other three variants, i.e., N50D, R57T, M134Q **(Table 1)**. The magnitude of this decrease is in the same ballpark as the magnitude of the median increase in the T_m found for 93 cases of engineered proteins, most of which contain more than one substitution[227]. Thus, the results of the CNA analyses do not support the hypothesis that increased thermodynamic thermostability of the six variants led to an increased extracellular amount in the culture supernatant. However, it should be noted that CNA does not consider time-dependency of processes; hence, our analyses do not rule out an increase in kinetic thermostability as a cause for higher extracellular amount.

Table 1: Predicted thermodynamic thermostabilities of wt*Bs*LipA and *Bs*LipA variants using CNA.

***Bs*LipA variants**	$\tilde{rc}_{ij,\ neighbor}$ [K][a]	$\Delta\tilde{rc}_{ij,\ neighbor}$ [K][b]
wt*Bs*LipA	316.1	/
N50D	312.1	-4.0
P53D	316.2	0.1
P53E	315.8	-0.3
P53V	315.8	-0.3

R57T	314.9	-1.2
M134Q	314.7	-1.4

[a] The $\widetilde{rc}_{ij,\ neighbor}$ values were converted to a temperature scale according to **Eq. 4 (section 2.2.2.1)**.
[b] Difference of $\widetilde{rc}_{ij,\ neighbor}$ values of *Bs*LipA variants and wt*Bs*LipA, respectively.

Finally, in order to answer the question whether a synergistic effect can be achieved by combining single AA substitutions that themselves have led to increased extracellular activity or amount, we chose single AA substitutions with beneficial effects. Combination of beneficial single AA substitutions revealed an additive effect solely at the level of extracellular amount of *Bs*LipA. Similar additive effects were already described for AA substitutions improving thermostability, where 12 amino acid substitutions were introduced by several rounds of in vitro evolution resulting in an increase of the LipA temperature optimum by ~ 30 °C[228]. However, extracellular activity and amount of *Bs*LipA could not be increased simultaneously.

7.3 Conclusion and Significance

In this study, for the first time, we performed a systematic large-scale analysis of a nearly complete experimental SSM library of *Bs*LipA towards the contribution of single AA and codon substitutions to the production and secretion with CNA.

The principle results of this study are:

- Out of ~30,000 clones 26 variants were identified showing an up to twofold increase in either extracellular activity or amount of *Bs*LipA.
- Single AA and codon substitutions did not substantially affect *B. subtilis* growth.
- Single AA and codon substitutions affect extracellular activity and amount of *Bs*LipA as well as *lipA* transcription.
- The CNA analyses did not support the hypothesis that increased thermodynamic thermostability led to an increased extracellular amount of *Bs*LipA.
- Combination of beneficial single AA substitutions revealed an additive effect solely at the level of extracellular amount of *Bs*LipA. However, extracellular activity and amount of *Bs*LipA could not be increased simultaneously.

The results signify that the optimization of the expression system is not sufficient for efficient protein production in *B. subtilis*. The sequence of the target protein should also be considered as an optimization target for successful protein production. Our results further suggest that

variants with improved properties might be identified much faster and easier if mutagenesis is prioritized towards elements that contribute to enzymatic activity or structural integrity.

8 SUMMARY AND PERSPECTIVES

In the present work, I rationalized the impact of enzyme flexibility and rigidity on protein thermostability and / or detergent tolerance **(PUBLICATION II)**[44], substrate promiscuity **(PUBLICATION III)**[45], and expression **(PUBLICATION IV)**[35] using our in-house Constraint Network Analysis (CNA) software[46, 72] at large-scale for biotechnologically highly relevant bacterial lipolytic enzymes (esterases/ESTs and lipases/LIPs). This was done with the aim to define the scope and limitations of biomolecular flexibility predictions in knowledge-driven strategies for protein engineering.

In **PUBLICATION II**[44] I performed a systematic large-scale analysis of a complete experimental site saturation mutagenesis (SSM) library of the model enzyme *Bacillus subtilis* lipase A (*Bs*LipA) to scrutinize the impact of substitution sites on two types of protein stability, thermostability (T_{50}) and detergent tolerance (D), with CNA. The results provide systematic and unbiased reference data at unprecedented scale for *Bs*LipA, identify consistently defined *hot spot* types for evaluating the performance of CNA, and show that CNA-based *hot spot* predictions yield an up to 9-fold gain in precision over random classification. Hence, CNA can be used prior to experiments when seeking to optimize enzymes' thermostability and detergent tolerance. In future studies, the study should be extended to other types of protein stability, such as tolerance against ionic liquids. Experimental data at large scale that can provide the basis for such investigations has been published recently[174].

In **PUBLICATION III**[45] I exploit comprehensive experimental information on the substrate promiscuity ($P_{EST/LIP}$) of 147 ESTs/LIPs tested against a customized library of dissimilar esters[34]. I used CNA to understand the structural origin of and to predict $P_{EST/LIP}$. Unexpectedly, our data reveal that promiscuous ESTs/LIPs, in contrast to specific ones, are significantly globally less flexible and have less flexible catalytically active residues, are significantly more thermostable, and have a significantly increased specific activity. Furthermore, specific ESTs/LIPs prefer to hydrolyze large and flexible esters. These results may be reconciled with a model according to which multiple ligands can be accommodated in a single site of promiscuous ESTs/LIPs by exploiting diverse interacting residues, whereas structural flexibility in the case of specific ESTs/LIPs serves for conformational proofreading. Our results furthermore signify that EST/LIP sequence space, charted, e.g., by (meta)genomics studies, can be screened by rigidity analyses based on CNA for promiscuous

ESTs/LIPs that may serve as starting points for further exploration in biotechnology and synthetic chemistry. This knowledge can now be used to characterize prospectively further ESTs/LIPs of industrial/commercial relevance with respect to $P_{EST/LIP}$. Furthermore, the unexpected relationship of flexibility and $P_{EST/LIP}$ warrants further experimental validation by methods that are capable to resolve structural dynamics, such as NMR or FRET experiments.

In **PUBLICATION IV**[35] I analyzed parts of a nearly complete experimental SSM library of *Bs*LipA towards the contribution of single AA and codon substitutions to the production and secretion with CNA. The results suggest that single AA and codon substitutions affect extracellular activity and amount of *Bs*LipA as well as *lipA* transcription. Combination of beneficial single AA substitutions revealed an additive effect solely at the level of extracellular amount of *Bs*LipA. The CNA analyses did not support the hypothesis that increased thermodynamic thermostability led to an increased extracellular amount of *Bs*LipA. In future studies it would be very interesting to investigate the relation between biomolecular flexibility and secretion of *Bs*LipA with CNA.

To sum up, nowadays computational techniques used for knowledge-driven strategies emerged as useful tools in protein engineering with respect to save resources, e.g. working effort, time, and costs.

ACKNOWLEDGEMENT

First, I thank Prof. Dr. Karl-Erich Jaeger for giving me the opportunity to carry out my research under his supervision. I am thankful for his continuous interest in my projects and the valuable discussions during the time.

I also thank Prof. Dr. Birgit Strodel for agreeing to act as a second supervisor.

I am grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich Heine University Düsseldorf. Furthermore, I gratefully acknowledge the computing time granted by the John von Neumann Institute for Computing (NIC) and provided on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC).

I thank Prof. Dr. Holger Gohlke for in-depth discussions and guidance with respect to the computational aspects of my work.

Herewith, I thank Prof. Dr. Karl-Erich Jaeger and Prof. Dr. Ulrich Schwaneberg for providing me the experimental thermostability and detergent tolerance data of the site saturation mutagenesis library of the *Bacillus subtilis* lipase A (*Bs*LipA).

Thanks to Susanne Hermans, Dr. Christopher Pfleger, and Dr. Christian Hanke for the successful cooperation to review fundamental concepts in rigidity theory for biomolecules.

Thanks also to Dr. Alexander Fulton and Dr. Olav Zimmermann for the productive cooperation to systematically scrutinize the impact of substitution sites on thermostability and detergent tolerance for *Bs*LipA.

Moreover, thank you to Dr. Manual Ferrer for the fruitful cooperation to investigate the substrate promiscuity of esterases.

I also thank Dr. Pia Skoczinski for the successful cooperation to investigate the contribution of single amino acid and codon substitutions to the production and secretion of *Bs*LipA.

In addition, I thank Dr. Christopher Pfleger for giving me access to CNA and the fruitful discussions.

Thanks to Daniel Mulnaes for giving me access to TopModel and TopScore, as well as for the profound discussions.

I am grateful to Dr. Christoph Gertzen and Dr. Benoit David for critically reading my thesis.

Thanks to Dr. Benoit David, Daniel Becker, and in particular Birte Schmitz for the cordially working atmosphere in our office.

Moreover, I thank all employees of the Jülich Supercomputing Centre (JSC), the John von Neumann Institute for Computing (NIC), the Computational Biophysical Chemistry group (CBClab), the Institute of Biological Information Processing (IBI-7), the Institute of Molecular Enzyme Technology *(*IMET*),* the Institute for Pharmaceutical and Medicinal Chemistry, and the Computational Pharmaceutical Chemistry group (CPClab) for the friendly working atmosphere and the many helpful suggestions during my PhD thesis.

Last but not least, special thanks to my family, who were always there for me.

REPRINT PERMISSIONS

PUBLICATION I (pages 56-86), Figure 4 (page 10), Figure 5 (page 12), Figure 6 (page 13), Figure 7 (page 15), Figure 8 (page 16)

Reprinted (adapted) with permission from **"Rigidity theory for biomolecules: Concepts, software, and applications"**, Hermans, S.M.A., Pfleger, C., Nutschel, C., Hanke, C.A., Gohlke, H., ***WIREs Comput Mol Sci***. 2017, 7, e1311, Copyright (2020) John Wiley & Sons.

PUBLICATION II (pages 87-131), Figure 13 (page 34), Figure 15 (page 37), and Figure 16 (page 38)

Reprinted (adapted) with permission from **"Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for Bacillus subtilis lipase A"**, Nutschel, C., Fulton, A., Zimmermann, O., Schwaneberg, U., Jaeger, K.-E., Gohlke, H., ***J Chem Inf Model.*** 2020, 60, 3, 1568-1584, Copyright (2020) American Chemical Society.

PUBLICATION III (pages 132-200), Figure 17 (page 40), Figure 18 (page 41), and Figure 19 (page 43)

Reprinted (adapted) with permission from **"Promiscuous esterases counterintuitively are less flexible than specific ones"**, Nutschel, C., Coscolín, C., Mulnaes, D., David, B., Ferrer, M., Jaeger K.-E., Gohlke, H., ***J Chem Inf Model.*** 2020, DOI: 10.1021/ acs.jcim.1c00152.

PUBLICATION IV (pages 201-230), Figure 20 (page 46), Figure 21 (page 47), and Table 1 (page 48)

Reprinted (adapted) with permission from **"Contribution of single amino acid and codon substitutions to the production and secretion of a lipase by *Bacillus subtilis"***, Skoczinski, P., Volkenborn, K., Fulton, A., Bhadauriya, A., Nutschel, C., Gohlke, H., Knapp, A., Jaeger, K.-E., ***Microb Cell Fact.*** 2017, 16, 160.

ORIGINAL PUBLICATION I

Rigidity theory for biomolecules: concepts, software, and applications

Hermans, S.M.A., Pfleger, C., Nutschel, C., Hanke, C.A., Gohlke, H.

WIREs Comput Mol Sci. 2017, 7, e1311.

https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1311

Advanced Review

Rigidity theory for biomolecules: concepts, software, and applications

Susanne M.A. Hermans,[†] Christopher Pfleger,[†] Christina Nutschel, Christian A. Hanke and Holger Gohlke*

The mechanical heterogeneity of biomolecular structures is intimately linked to their diverse biological functions. Applying rigidity theory to biomolecules identifies this heterogeneous composition of flexible and rigid regions, which can aid in the understanding of biomolecular stability and long-ranged information transfer through biomolecules, and yield valuable information for rational drug design and protein engineering. We review fundamental concepts in rigidity theory, ways to represent biomolecules as constraint networks, and methodological and algorithmic developments for analyzing such networks and linking the results to biomolecular function. Software packages for performing rigidity analyses on biomolecules in an efficient, automated way are described, as are rigidity analyses on biomolecules including the ribosome, viruses, or transmembrane proteins. The analyses address questions of allosteric mechanisms, mutation effects on (thermo-)stability, protein (un-)folding, and coarse-graining of biomolecules. We advocate that the application of rigidity theory to biomolecules has matured in such a way that it could be broadly applied as a computational biophysical method to scrutinize biomolecular function from a structure-based point of view and to complement approaches focused on biomolecular dynamics. We discuss possibilities to improve constraint network representations and to perform large-scale and prospective studies.

How to cite this article:
WIREs Comput Mol Sci 2017, e1311. doi: 10.1002/wcms.1311

INTRODUCTION

Biomolecules are generally marginally stable[1] and are heterogeneously composed of flexible and rigid regions.[2] Here, flexibility and rigidity denote the possibility, or impossibility, of internal motions in an object under force without giving information about directions and magnitudes of movements. The importance of the mechanical heterogeneity, which is usually highly conserved within homologs,[3] for biomolecular function cannot be overstated. For enzymes, a dual character of active sites in terms of high and low structural stability has been described,[4] reflecting optimization for ligand access,[5] binding affinity,[6] and catalytic efficiency.[7] Regulatory sites of biomolecules need to display a sufficiently low structural stability such that bound effector molecules can modify their flexibility and rigidity in order to initiate signaling.[8] As to thermal stability, proteins from thermophilic organisms are generally less flexible than their mesophilic homologs.[9] Therefore, understanding biomolecular flexibility and rigidity, and how they change due to binding of another molecule, mutations, temperature, or solvent, is instrumental both for a fundamental understanding of biomolecular function[10,11] and with respect to protein engineering and ligand design.[2,12–15]

From an experimental point of view, flexibility and rigidity characteristics of biomolecules have been

[†] These authors contributed equally to this work.

*Correspondence to: gohlke@uni-duesseldorf.de

Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Conflict of interest: The authors have declared no conflicts of interest for this article.

investigated using X-ray crystallography,[16] nuclear magnetic resonance (NMR) spectroscopy,[17] or fluorescence spectroscopy.[18] The main sources of information from these techniques reflecting flexibility characteristics are crystallographic B-factors, NMR order parameters and residual dipolar couplings, and relaxation times.[19–21] These sources report on atomic mobility, however, from which flexibility and rigidity characteristics then have to be derived.[19,22] In contrast, atomic force microscopy (AFM) allows for measuring the mechanical rigidity of biomolecules directly on a single molecule level.[23]

From a computational point of view, molecular dynamics (MD) simulations,[24] coarse-grained (CG) simulations,[25] or normal mode analysis (NMA)[26] and related analyses[27] are widely used to investigate biomolecular flexibility and rigidity. Again, the primary information these approaches yield is about atomic mobility, from which flexibility and rigidity characteristics then have to be derived.[28,29] Alternative approaches rely on a representation of the 3D structure of a biomolecule in terms of a *connectivity* network, where atoms or residues are represented as nodes and the interactions between them as edges.[30–41] In such a network, the actual lengths and angles of bonds are irrelevant for subsequent analysis. A structural hierarchy is then deduced, with atoms or residues within a subgraph having a high connectivity, thus indicating a region of higher structural stability. In contrast, atoms or residues connecting two subgraphs are less tightly connected, thus forming the flexible regions.[42–45]

Biomolecules can also be modeled as *constraint* networks, where the edges represent constraints due to covalent and noncovalent interactions that fix the distance between the nodes, thereby restricting internal motions.[46] In contrast to MD and CG simulations or NMA, where interactions between atoms are modeled by forces of varying strengths, in constraint networks a constraint is either present or not, but does not vary in strength with respect to the atoms' geometry. The constraint network can be efficiently decomposed into rigid clusters and flexible regions according to the number and spatial distribution of the remaining degrees of freedom (DOF), as described in detail below.[47] The study of network rigidity and how a network transitions from a flexible to a rigid state is known as rigidity percolation or rigidity theory.[48–50] The essential property common to all percolation type problems is that of a connected pathway; in rigidity percolation, the path consists of sites that are mutually rigid.[50] In comparison to the connectivity percolation studied in the above connectivity networks, there are two important differences.[51] First, in connectivity percolation, the propagation of a scalar property is monitored (e.g., conductivity), while in rigidity percolation the propagation of a vector (e.g., stress) is, in general, considered.[52] Second, there is an inherent long-range aspect to rigidity percolation, that is, whether a region is flexible or rigid generally depends on structural details that are far away.[50,52,53]

The study of network rigidity originated from the field of structural engineering more than 150 years ago, where it was first applied to mechanical systems (Figure 1; Box 1).[54,55] Later, it was extended to the fields of solid state physics, for addressing network glasses[56,57] and zeolithes,[58] and biophysics for investigating biomolecules.[59–62] Since the underlying idea is simple yet not trivial, computationally highly efficient, and gives insights into flexibility and rigidity characteristics of biomolecules at an atomistic level, the approach has gained much attention recently. In the following, we will describe the theory underlying this approach, current methods for modeling and analyzing constraint networks, as well as applications to biomolecules linking flexibility and function.[63] These applications include investigating large biomolecules such as the ribosome,[64] understanding allostery,[64,65] predicting thermodynamic properties,[66] assessing the structural stability of complexes,[67,68] identifying folding cores of

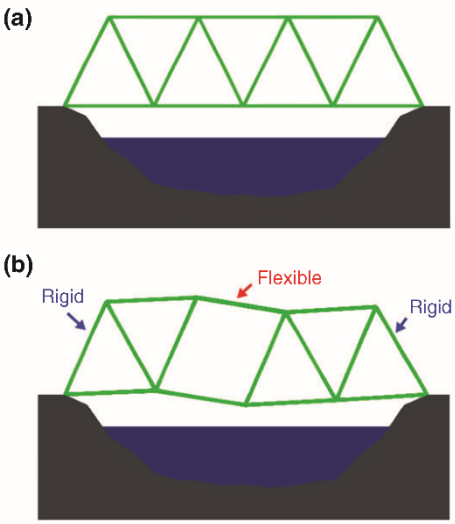


FIGURE 1 | Schematic representation of a structural engineering construction (bridge) consisting of struts (distance constraints) connected by joints. (a) In 2D, the triangle is the smallest rigid unit. Hence, if all constraints are in place, the bridge is *isostatic* or *minimally rigid*. (b) Removing one constraint divides the bridge into two rigid clusters with a flexible region in between.

proteins,[69,70] sampling of biomolecular conformational spaces,[71–74] finding putative binding sites,[15] and analyzing structural determinants of thermostability.[75,76]

BOX 1

CONSTRAINT COUNTING

The first mathematical formulation of rigidity analysis dates back to the 19th century, where James Clerk Maxwell investigated the conditions under which mechanical structures, made of joints and connecting struts, are stable or instable (Figure 1).[54] For this, Maxwell used constraint counting as a mean field approach, which circumvented any detailed local calculations, to assign the number of *independent* internal degrees of freedom (DOF), also called 'floppy modes' (F). F determines possible movements of a structure in the d-dimensional space without violating any of the constraints. For a network with N sites, lacking any constraints, F is given by Eq. (1), where the latter term denotes the global degrees of freedom.

$$F = dN - d(d+1)/2 \quad (1)$$

In a system with N_c constraints, assumed by Maxwell to be independent, each constraint removes one floppy mode, resulting in the number of floppy modes according to Maxwell (F_{mxw}, Eq. (2)).

$$F_{mxw} = dN - N_c - d(d+1)/2 \quad (2)$$

If not all constraints are independent, using Maxwell's equation will lead to an underestimation of F. This is corrected for by considering the number of redundant constraints N_r (Eq. (3)).[55]

$$F = dN - (N_c - N_r) - d(d+1)/2 \quad (3)$$

Redundant constraints introduce stress in the network and do not add to the stability of the network anymore.[46] A network region with redundant constraints is *overconstrained* or *stressed*. If a region has fewer constraints than internal DOF, it is *underconstrained* or *flexible*. If a region has as many constraints as internal DOF, the region is *isostatically* (or *minimally*) *rigid*.[77]

MODELING AND ANALYZING BIOMOLECULES AS CONSTRAINT NETWORKS

Constraint Network Representations for Proteins

Biomolecules are represented as constraint networks by transforming atoms into nodes, and covalent and noncovalent bonds into constraints in between. There are several types of constraint networks (Figure 2(a)–(d)).[56] In *bond-bending* networks, nodes are considered joints having three DOF, and constraints connect nearest-neighbor nodes to fix the distance between them. Next-nearest-neighbors are also connected to fix the angles (Figure 2(b)). This representation is also called a *molecular graph* or *molecular framework*, as it intuitively represents molecules with their strong bond and angle forces.[80,81] For propene (Figure 2(a)), with one double and one single bond between the carbon atoms, free rotation about the single bond is possible, resulting in one independent internal degree of freedom (also termed floppy mode) (Figure 2(b) and (e) top left). The molecule can be decomposed into two rigid clusters, one consisting of five atoms **a**, **b**, **c**, **d**, and **e**, and one of four atoms **f**, **g**, **h**, and **i** - (Figure 2(b)). In these networks, a double bond is modeled by placing an additional distance constraint between third-nearest-neighbors, for example, **b** and **f** (Figure 2(b) and (e) middle left), preventing dihedral rotation.[78,82] Alternatively, molecular structures are represented as *body-and-bar* networks (Figure 2(c))[61,81] and *body-bar-hinge* networks (Figure 2(d)),[79,81] where atoms are considered as rigid bodies having six DOF, which are connected by bars. Two rigid bodies have in total 12 DOF. Disregarding the six global DOF, six bars are needed to lock in the internal DOF and, hence, to model double and peptide bonds (Figure 2(e) middle right). A single bond is modeled with five constraints, leaving one DOF for the dihedral rotation (Figure 2(e) top right).

Stronger noncovalent interactions, such as hydrogen bonds (including salt bridges) and hydrophobic interactions, are essential for the stability of biomolecules and, thus, require accurate modeling in the constraint network. In contrast, weaker interactions such as van der Waals or electrostatic forces are not included in the network. In all network types, modeling of different interaction strengths is possible by including a differential number of constraints/bars.[61,78] In *bond-bending* networks, hydrogen bonds have been modeled using three distance constraints, removing three DOF as does a covalent bond (Figure 2(e) top left), that way representing the geometric restriction due to hydrogen bonds.[60] Hydrophobic interactions have been modeled in

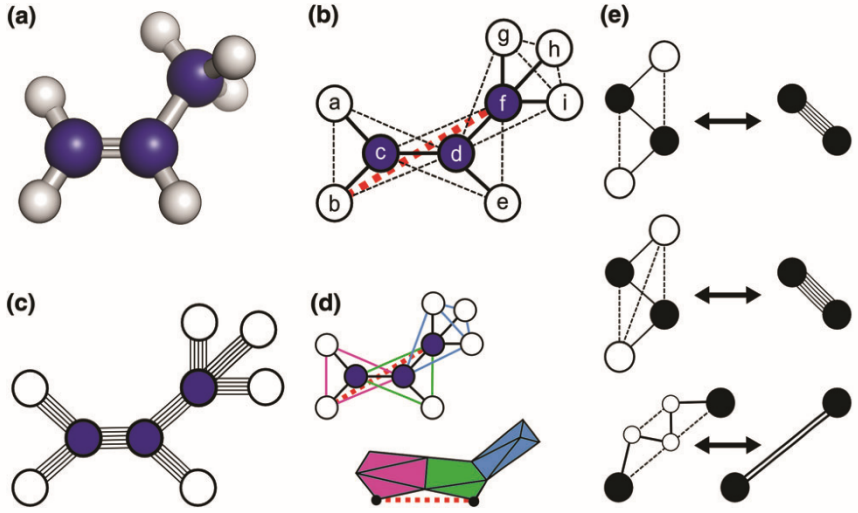


FIGURE 2 | Constraint network representations. (a) Ball-and-stick representation of propene, the carbon atoms are shown in blue and the hydrogen atoms in light gray. (b, c, d) Propene is represented in terms of 3D constraint networks.[78] (b) In the *bond-bending* network (also called *bar-and-joint* network or *molecular framework*) covalent bonds are modeled as distance constraints between nearest-neighbor atoms (thick lines) and angle constraints between next-nearest-neighbor atoms (dashed lines). For the double bond (**c**,**d**), there is an additional constraint (red dotted line) between third-nearest-neighbor nodes (**b**,**f**), removing the bond-rotational DOF between the two sp^2 carbons. The network represented here has a total of nine nodes, connected by eight distance constraints, eleven next-nearest-neighbor constraints, and one third-nearest-neighbor constraint. In this network, a node (atom) has three DOF, leading to a $3N - 6$ count (Eq. (1) in Box 1). With $N = 9$ nodes and a total of 20 nonredundant constraints, this network has one DOF, the rotation around the single bond. (c) In the *body-and-bar* representation, atoms are modeled as bodies with six DOF, a covalent single bond as five constraints between two bodies, and a double bond as six constraints. (d) In the *body-bar-hinge* model, all covalent bonds are replaced by hinge regions, located at the connection of two colored shapes, connected in such a way that one DOF is left. For the double bond, an additional bar (red dotted line) is added to the hinge region to lock the remaining DOF.[79] (e) The modeling of bond types is compared between the *bond-bending* network (left column) and the *body-and-bar* network (right column): The covalent bond with five constraints (top), the double bond with six constraints (middle), and the hydrophobic interaction modeled with ghost atoms in the *bond-bending* network (bottom left) and with two bars in the *body-and-bar* network (bottom right). Figure 2(e) adapted from Ref 61.

terms of three pseudoatoms and the associated constraints (Figure 2(e) bottom left), essentially removing two DOF, that way representing that hydrophobic interactions are less geometrically restrictive.[59,83] In *body-and-bar* networks, hydrogen bonds are modeled with five bars, as are covalent bonds (Figure 2 (e) top right),[61] and hydrophobic interactions with two bars (Figure 2(e) bottom right)[61,84,85] although

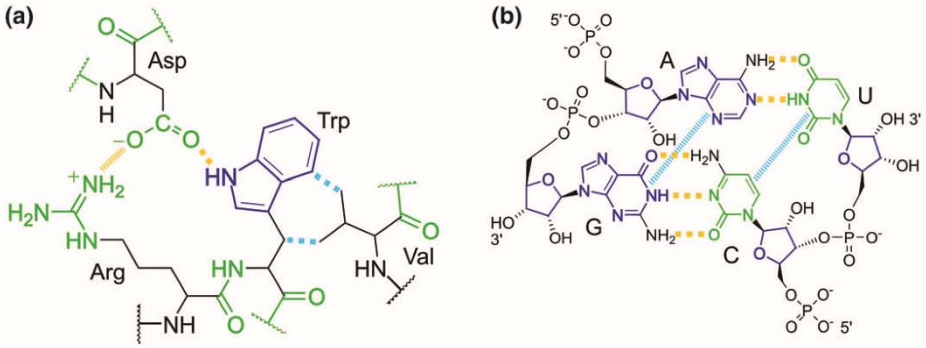


FIGURE 3 | Modeling of covalent and noncovalent interactions. For both (a) interactions within a protein and (b) RNA, the rigid clusters (green) and overconstrained regions (blue) are shown. For rigidity analysis, covalent interactions (black lines), hydrogen bonds (yellow squared dots) and salt bridges (yellow hatched lines), and hydrophobic interactions (cyan squared dots) are modeled as constraints. For RNA also base-stacking interactions (cyan hatched lines) are modeled as hydrophobic interactions.[62]

lower and higher numbers of bars have been used for hydrophobic interactions, too.[85,86]

Deciding which noncovalent interactions to include in the network is decisive for getting an accurate representation of the flexibility of the system (Figure 3).[68,87] For this, the strength of hydrogen bonds is evaluated, for example, according to Mayo's hydrogen bond potential energy (E_{HB}, Eq. (4)).[88]

$$E_{HB} = D_0 \left\{ 5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{10} \right\} f(\theta,\phi,\varphi), \qquad (4)$$

where R_0 is the equilibrium distance (2.8 Å) and R is the hydrogen bond distance between donor and acceptor. D_0 is the well-depth of the interaction. The angle term f varies depending on the hybridization state of the donor and acceptor atoms; θ is the angle of the triplet (donor, hydrogen, acceptor); ϕ is the angle of the triplet (hydrogen, acceptor, base atom bonded to the acceptor); φ is the torsion angle between the normals of two planes defined by two sp^2 centers. In the case of sp^3 hybridization, φ is not considered. Only hydrogen bonds with an energy $E_{HB} \leq E_{cut}$ are included in the constraint network.[60,82] Hydrophobic interactions are often included in the constraint network according to the criterion that the distance between carbon and/or sulfur atoms is less than the sum of their van der Waals radii (C: 1.7 Å, S: 1.8 Å) plus a distance cutoff D_{cut} = 0.25 Å.[84] Alternatively, Fox et al.[85] introduced a parameter to describe the strength of hydrophobic interactions based on the pairwise van der Waals energy derived from the Lennard-Jones potential of the AMBER parm99 force field.[89,90]

Results from rigidity analyses on biomolecules can be affected by additional factors such as water molecules, ions, small-molecule ligands, or other biomolecules. It was shown that the inclusion of structural waters in the constraint network had only a negligible effect on the protein's flexibility.[68,91] In contrast, waters that bridge protein–ligand interactions can rigidify the complex structure.[60] Bridging interactions mediated by water molecules were modeled by hydrogen bonds,[60] while interactions with structural ions were modeled as covalent bonds.[92] Effects of small-molecule ligands[15,60] and biomolecular binding partners[68] are described below.

Modification of the Constraint Network Representation for RNA Structures

In comparison to proteins, RNA structures are less globular, more elongated, and less densely packed.[93] While the structure of proteins is predominantly determined by hydrophobic interactions of amino acid side-chains in the protein core, the stability of RNA strongly depends on hydrogen bonds and base-stacking interactions.[93] Not surprisingly, the constraint network representation initially developed for proteins (see above) turned out to be not appropriate for RNA systems.[62] Fulle et al. modified the network representation for RNA structures by adapting the criteria for the inclusion of hydrophobic interactions, including a limit for the number of constraints considered between neighboring bases (Figure 3).[62] The modifications were verified by comparing predictions from rigidity analysis to mobility information derived from crystallographic B-factors of a $tRNA^{ASP}$ structure.[62] Furthermore, atomic fluctuations calculated for a structural ensemble of HIV-1 TAR RNA generated by the constrained geometric simulations tool FRODA (Framework Rigidity Optimized Dynamic Algorithm; see *Generation of Effective Constraint Networks*) were compared to the conformational variability derived from an NMR ensemble.[72] The new RNA parameterization proved more successful than the protein parameterization and another parameterization by Wang et al.[94] for the prediction of conformational variabilities of NMR ensembles of 12 RNA structures.[62] Future improvements of the RNA parameterization may consider the repulsion of negatively charged phosphate groups and sequence-dependent base-stacking. Note that the proposed parameterization may not be ideally suited for DNA molecules, due the different flexibility characteristics of RNA and DNA, for example with respect to the sugar ring and the deformability of the molecules.[62]

Constraint Counting: The Pebble Game Algorithms

For a given constraint network, Eq. (3) (see Box 1) yields F in terms of a mean field approximation.[55] In 1970, *Laman's theorem*[55] had a major impact in that it allows to determine the DOF **locally** in generic (i.e., lacking any special symmetries) 2D constraint networks by applying constraint counting to all subgraphs within the network. As such, a generic 2D network is minimally rigid if and only if the number of constraints is $2N - 3$, and every nonempty subgraph s induced by $N_s \geq 2$ sites spans at most $2N_s - 3$ constraints. Based on Laman's theorem, Hendrickson suggested an algorithm that exactly counts the number of floppy modes in a generic 2D network and, hence, is appropriate to

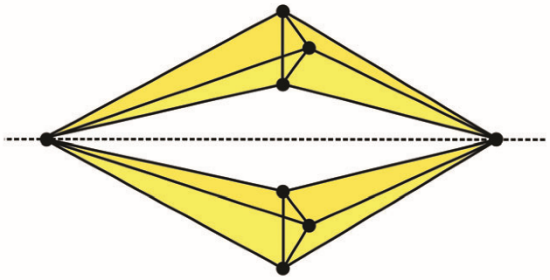


FIGURE 4 | Double banana network. Constraint counting implies that the 3D double banana network is rigid because it satisfies the $3N - 6$ counting condition considering that the nodes have three DOF. However, internal motion within this network is possible along the implied-hinge joint between the two 'banana' subgraphs (dashed line). Figure adapted from Ref 77.

decompose it into rigid regions and flexible links in between.[46] Further developments on this algorithm led to the efficient combinatorial *2D pebble game algorithm* implemented by Thorpe and Jacobs.[47]

However, this type of algorithm can fail if applied to a general 3D network such as the 'double banana' network (Figure 4).[59] This network has overall $3N - 6$ constraints, and none of the subgraphs has more than $3N_s - 6$ constraints connecting N_s sites. Applying the 3D analog of Laman's theorem would thus lead to the conclusion that this network is minimally rigid, which is wrong as there is an implied-hinge joint between the two 'banana' subgraphs. With the *molecular framework conjecture*,[81] Tay and Whiteley proposed that the constraint counting can be extended to a certain subtype of 3D networks with a molecule-like character, the *bond-bending* networks (see *Modeling and Analyzing Biomolecules as Constraint Networks*). Based on this proposition, Jacobs constructed a 3D pebble game algorithm for these networks,[77] the computational time complexity of which is, in a worst case scenario, $O(N^2)$; in practice, the algorithm runs in linear time.[82] In comparison, brute force numerical techniques can give the same result as the pebble game algorithm, but are generally unfeasible for large systems due to a computational complexity of $O(N^3)$.[82]

The pebble game algorithm for bond-bending networks has been implemented in early versions of the Floppy Inclusion and Rigid Substructure Topography (FIRST) software (see *FIRST/ProFlex*).[60] In 2004, Hespenheide et al. implemented an adapted 3D pebble game algorithm using a $6N - 6$ count[61] applied on the *body-and-bar* representation of molecules[81] (see *Modeling and Analyzing Biomolecules as Constraint Networks*). In 2008, Lee and Streinu described a family of pebble game algorithms, the (k,l)-pebble games, where k is the initial number of pebbles on each node and l is the acceptance condition, that is, the global degrees of freedom of the system (see Box 2; Figure 5).[96,97] The original 2D pebble game algorithm of Jacobs and Hendrickson[50] is a $(2,3)$-pebble game in this terminology.[96] A $(6,6)$-pebble game implemented by Fox et al.[79] for analyzing *body-bar-hinge* networks is equal to the 3D pebble game algorithm introduced by Hespenheide et al. for analyzing *body-and-bar* networks.[61] Notably, the family of (k,l)-pebble games were proven to be correct by Katoh and Tanigawa in

BOX 2

THE PEBBLE GAME ALGORITHM

For explaining the $(6,6)$-pebble game (with the $6N - 6$ counting condition), an exemplary biomolecule is modeled as a *body-and-bar* network with four nodes connected by a total of 18 constraints (Figure 5). Initially, six pebbles are placed on each node in the network, representing the six DOF in 3D (see *Modeling and Analyzing Biomolecules as Constraint Networks*). For the decomposition into rigid and flexible regions, the pebble game considers two rules for two connected nodes i and j[97]:

- Define a constraint between the nodes: if i and j have at least seven pebbles in total, place a pebble on the constraint from i to j to define the constraint in the direction of j.
- Slide a pebble: if there is a defined constraint between i and j and there is a pebble on j, reverse the direction of the constraint and move the pebble from j to i.

Accordingly, five pebbles are first placed on the constraints between **b** and **c** defining all five constraints in the same direction (1). Then, five pebbles are placed on the constraints from **c** to **d** and from **d** to **a** (2). This leaves six pebbles on **a** and one pebble on **b**, **c**, and **d**, respectively. All single pebbles are now collected on **b** (3, 4). There are now six pebbles on **a** and three pebbles on **b**; **c**, and **d** are empty. Finally, the last three constraints are defined by placing the three pebbles on the constraints between **b** and **a** (5). Now 18 pebbles are used, and all constraints are defined (6). The remaining six pebbles on **a** represent the six global DOF, demonstrating that this graph is minimally rigid.[97]

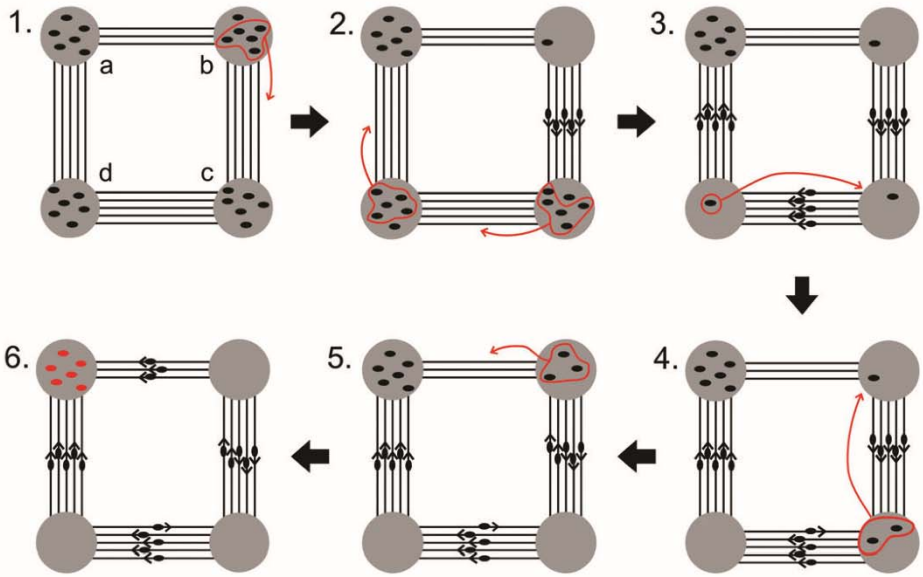


FIGURE 5 | The 3D pebble game algorithm; see Box 2 for details. Figure adapted from Ref 95.

2011,[98] almost 150 years after Maxwell's introduction of constraint counting as a mean field approach.[54] For further details on pebble game algorithms see Refs 50,78,97.

Analyzing Network States along Constraint Dilution Trajectories

By gradually removing noncovalent constraints from an initial network representation of a biomolecule, a succession of network states $\{\sigma\}$ is generated that is hereafter termed 'constraint dilution trajectory'. Analyzing such a trajectory by rigidity analysis reveals a hierarchy of rigidity that reflects the modular structure of biomolecules in terms of secondary, tertiary, and supertertiary structure.[14,69,75,83,99] In particular, constraint dilution allows simulating the loss of structural stability of a biomolecule with increasing temperature. For this, hydrogen bonds are removed from the constraint network if $E_{HB} > E_{cut,\sigma}$, where $\sigma = f(T)$ is the state of the network at temperature T (Figure 6(a)) and $E_{cut,\sigma_1} > E_{cut,\sigma_2}$ for $T_1 < T_2$.[88] Hydrophobic interactions are generally not removed along the constraint dilution trajectory because they remain constant in strength or become even stronger with increasing temperature.[100,101] Alternatively, a modified method for accounting for the temperature dependence of hydrophobic interactions has been introduced that adds more constraints to the network with increasing temperature by linearly increasing the distance cutoff D_{cut}.[102]

The hierarchy of rigidity of biomolecules leads to a percolation behavior that is often more complex than that of network glasses,[56] and multiple phase transition points can be identified along the constraint dilution trajectory at which rigid clusters decompose (Figure 6(b)).[84] The *rigidity percolation threshold* is then defined as the phase transition when the network changes from an overall rigid to an overall flexible state and thus loses its ability to transmit stress.[75]

Global and Local Indices for Characterizing Biomolecular Stability

For having maximal advantage from rigidity analysis, the results need to be linked to biologically relevant characteristics of a structure. At the macroscopic level, this is, for example, the phase transition point where a biomolecule switches from a structurally stable (largely rigid) to an unfolded (largely flexible) state; at the microscopic level, the localization and distribution of structurally weak parts may be a characteristic of interest. As links, several global and local indices were reported in the literature to depict these characteristics (see Table S1 in Ref 92 for a comprehensive overview). These indices are computed, to a varying extent, by the software packages described in section: *Software Packages for Rigidity Analysis*.

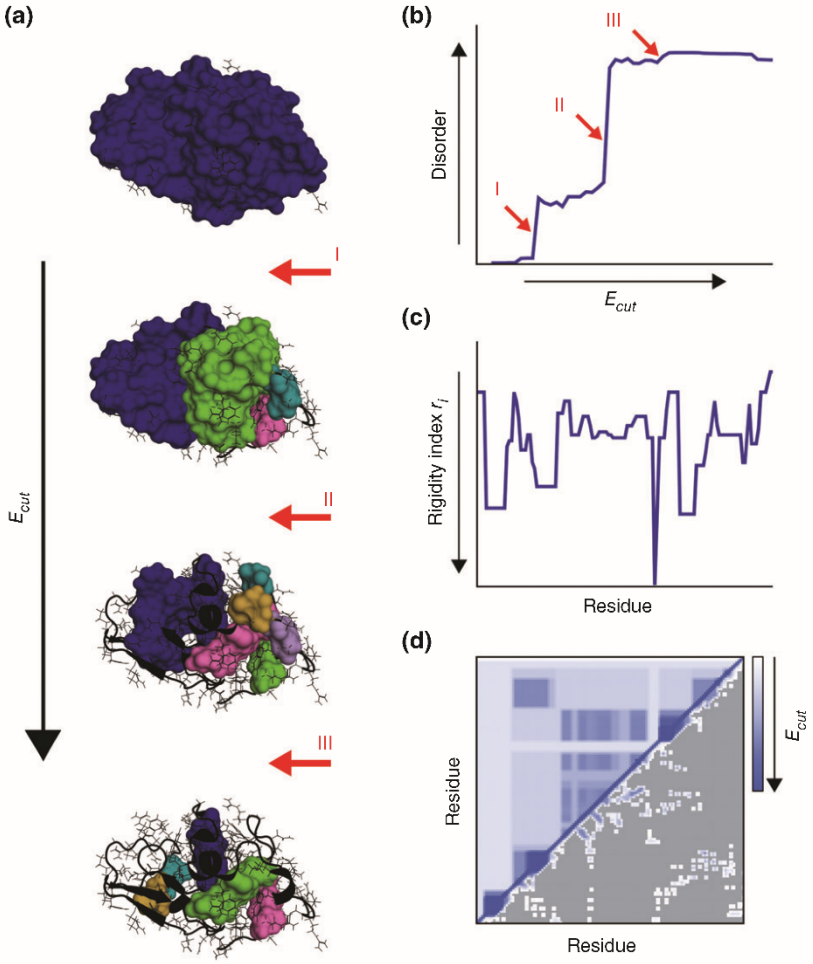


FIGURE 6 | Results of a constraint dilution simulation of hen egg white lysozyme with CNA. (a) In the constraint dilution simulation, a stepwise decrease in the cutoff energy (E_{cut}) removes hydrogen bonds from the constraint network in the order of increasing strength. The colored surfaces represent the rigid clusters, and the black lines represent the flexible regions of the protein. (b) Degree of disorder along a constraint dilution simulation as revealed form the cluster configuration entropy H.[84] The disorder is low when a single rigid cluster dominates and increases when the cluster falls apart into smaller subclusters of different sizes. (c) The rigidity index r_i characterizes the per-residue stability as it monitors when a residue i segregates from any rigid cluster during a constraint dilution simulation. A lower r_i value indicates that the residue resides in a region of higher stability. (d) Stability maps (upper triangle) and neighbor stability maps (lower triangle) represent when a 'rigid contact' between two residues of the network (both residues belong to the same rigid cluster) vanishes during the constraint dilution simulation. Gray areas in the neighbor stability map indicate that no native contact exists for that residue pair. Figure adapted from Ref 84. Note that arrows at axes labeled with E_{cut} point in the direction of more negative values.

Global flexibility indices monitor the degree of flexibility and rigidity within constraint networks at the macroscopic level. The density of internal DOF [$\Phi = F / (6N - 6)$ for a *body-and-bar* network] is a direct measure for the intrinsic flexibility of a constraint network.[92] Further indices have been derived from percolation theory and characterize the microstructure of a network, that is, properties of the set of rigid clusters generated along a constraint dilution trajectory (see *Analyzing Network States along Constraint* Dilution Trajectories).[14] They include the rigidity order parameter (P_∞),[103] which monitors the decay of the largest rigid cluster, the mean rigid cluster size (S),[104] which monitors the decay of all but the largest rigid cluster,[103,104] and the cluster configuration entropy (H), a Shannon-type entropy[105] that is a morphological descriptor of the network heterogeneity.[106] P_∞, S, and H show a noncontinuous

behavior when monitored along a constraint dilution trajectory, revealing transitions in the network rigidity when the largest rigid cluster starts to decay, stops dominating the network, and finally collapses (Figure 6(b)). That way, H was successfully applied to analyze unfolding transitions in biomolecules that are related to thermostability (see *Constraint Dilution Simulations to Investigate Protein Thermostability*).[14,75,99,102,107]

Local indices characterize the network flexibility and rigidity down to the bond level. Accordingly, indices are derived for each covalent bond in the network by monitoring the cutoff energy E_{cut} along a constraint dilution trajectory when the bond changes from rigid to flexible. By summarizing indices for several bonds, one can describe structural stability on a per-residue basis.[92] The percolation index p_i is a local analog to the rigidity order parameter P_∞ and is most suitable to monitor the percolation behavior of a biomolecule locally. The rigidity index r_i is a generalization of the percolation index p_i[92] as it monitors when a residue segregates from any rigid cluster. In a showcase example on α-lactalbumin, it has been shown that both local indices p_i and r_i are sensitive enough to detect long-range aspects of altered stability upon even small perturbations (i.e., the removal of a calcium ion) of the network topology.[92] Furthermore, this study showed that the information derived from p_i and r_i is complementary in that p_i indicates regions of the biomolecule that segregate as a whole from the largest percolating cluster and so become mobile as rigid bodies, while r_i exposes hinge regions that encompass the rigid bodies.

Another set of local indices characterizes correlations of stability between pairs of residues.[92] As such, stability maps (rc_{ij}) are 2D generalizations of the rigidity index r_i (Figure 6(c) and (d)).[14] To derive a stability map, 'rigid contacts' between residue pairs are identified. A rigid contact exists if two residues belong to the same rigid cluster. Along the constraint dilution trajectory, stability maps are then constructed by monitoring E_{cut} at which a rigid contact between two residues is lost. A contact's stability thus relates to the microscopic stability in the network and, taken together, the microscopic stabilities of all residue–residue contacts result in a stability map. The map reveals that losses of rigid contacts do not only occur between isolated pairs of residues but also in a cooperative manner. That is, parts of the biomolecule break away from the rigid cluster as a whole. The sum over all rigid contacts yields a measure for the chemical potential energy due to noncovalent bonding in the system, which has been used recently as a proxy for the melting enthalpy of a protein and correlates with a protein's melting temperature.[108] The difference in the chemical potential energy between a ground and a perturbed state of a system was used in a one-step free energy perturbation approach[109,110] to compute an approximation of the free energy associated with the change in biomolecular stability due to the removal of a ligand or the introduction of a mutation (C. Pfleger, H. Gohlke, unpublished results). The results agreed with free energies of destabilization from chemical denaturation experiments for single and double mutations in eglin c.

In some cases, similar index definitions have been introduced by different groups.[92] For example, the Distance Constraint Model (DCM) approach (see *Distance Constraint Model*)[111] computes a global index θ as the average of F over the DCM ensemble, which is related to Φ; a local index P_R as the probability whether backbone dihedral angles are rotatable over the ensemble,[76] which is related to r_i; and a cooperativity correlation plot that quantifies the correlated stability of pairs of residues in terms of rotatable dihedral backbone angles,[66,76] which is related to rc_{ij}. Thus, it is recommended to use the index notations summarized in reference[92] and displayed here in future studies to make these differences clear.

SOFTWARE PACKAGES FOR RIGIDITY ANALYSIS

Rigidity analysis can be applied to different types of biomolecules such as proteins and nucleic acids, and the investigated systems range from small proteins and RNAs to complex biomolecular assemblies such as the ribosome or viruses (see *Single-point Rigidity Analysis on RNA and Nucleic Acid–Protein Complexes*). To automate and improve the efficiency of the analysis, several software packages have been developed (Figure 7).

FIRST/ProFlex

The FIRST program, developed by Jacobs et al.,[60] was the first implementation of a pebble game algorithm together with code for generating constraint networks for proteins. For a given input structure, the number of floppy modes, a decomposition of the network into rigid clusters, and the location of overconstrained regions is provided. In its initial version, the 3D pebble game algorithm for *bond-bending* networks has been implemented (see *Constraint Counting: The Pebble Game Algorithms*). This FIRST version, extended by a hydrogen bond dilution procedure[69,83] (see *Analyzing Network States along*

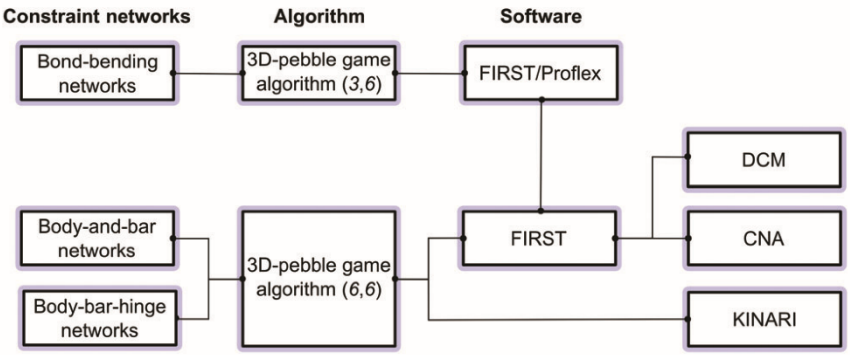


FIGURE 7 | Overview of the constraint network types, algorithms, and software packages discussed in this review.

Constraint Dilution Trajectories) and maintained in the Kuhn lab, is now available as MSU ProFlex from http://www.kuhnlab.bmb.msu.edu/software/proflex. FIRST was further developed in the Thorpe lab; this version is now based on a *body-and-bar* network representation and the (6,6)-pebble game algorithm.[59] Furthermore, the constraint network parameterization for RNA developed by Fulle et al.[62] has been included, and it has been extended by the constrained geometric simulation approach FRODA (see *Modification of the Constraint Network Representation for RNA Structures*).

Distance Constraint Model

The DCM developed by Jacobs and coworkers extends the concepts implemented in FIRST/ProFlex in that it analyses network rigidity at finite temperature applying statistical mechanics.[76,111–115] For this, constraints in the *bond-bending* network are characterized by local microscopic free energy functions, and topological rearrangements of thermally fluctuating constraints are permitted.[112,114] As noncovalent constraints, DCM models only hydrogen bonds and salt bridges, represented by three bars each, while hydrophobic contacts are neglected.[112] As a result, a partition function for the investigated system is obtained from an ensemble of constraint networks by combining microscopic free energies of individual constraints using network rigidity as an underlying long-range mechanical interaction.[112] In doing so, DCM considers that enthalpy is additive, whereas entropy is not. The nonadditivity of component entropies derives from not knowing *a priori* which constraints in the system are independent or redundant (see Box 1). In DCM, this problem is solved by recursively adding one constraint at a time to build a network, each time analyzing rigidity properties with the pebble game and determining whether a constraint is independent or redundant.[112] Since DCM works directly with free energies, it is possible to simulate the effects of temperature or pH fluctuations, as applied for c-type lysozyme[116,117] and homologous meso- and thermophilic RNAse H structures[76] (see *Single-Point Rigidity Analysis on RNA and Nucleic Acid–Protein Complexes*).[111] Generally, the DCM requires an accurate protein-specific parameterization based on *a priori* knowledge of experimentally determined heat capacity curves (C_p)[116,118]; if these were not available, C_p curves fitted to the peak of experimental melting temperatures (T_m) were used.[65,111] For DCM a minimum of three free parameters needs to be fit.[111,112]

Constraint Network Analysis

The Constraint Network Analysis (CNA) approach[84] was first introduced by Radestock and Gohlke[75] and aims at linking information from rigidity analysis derived from FIRST (see *FIRST/ProFlex*) with biomolecular structure, (thermo-)stability, and function. CNA functions as a front- and back-end to FIRST.[60] Owing to the C++–based CNA interface module pyFIRST, CNA has direct access to FIRST's data structure such that the computational efficiency of FIRST is preserved in CNA-driven computations, resulting in computing times of seconds for the rigidity analysis of a single conformation of an average-sized (250 residues) protein.[84] Going beyond the mere identification of flexible and rigid regions in a biomolecule, CNA allows for (a) performing constraint dilution simulations that consider a temperature dependence of hydrophobic tethers,[102,119] in addition to that of hydrogen bonds (see *Analyzing Network States along Constraint Dilution Trajectories*), (b) computing a comprehensive set of global and local indices for quantifying biomolecular stability (see *Global and Local Indices for Characterizing*

Biomolecular Stability), and (c) performing rigidity analysis on ensembles of network topologies (ENT). For the latter, structural ensembles and ensembles based on the concept of fuzzy noncovalent constraints (ENT^{FNC})[107] can be used (see *ENT from Fuzzy Noncovalent Constraints*). That way, information on the influence of finite temperature on constraint network representations is implicitly included without the need to derive system-specific parameters. As we[107,120] and others[91,121] observed, performing rigidity analysis on ENT instead of single networks greatly improves the robustness of the results. Furthermore, CNA can consider small-molecule ligands bound to biomolecules when constructing constraint networks.[84] In order to facilitate the processing of the highly information-rich results obtained from CNA, the VisualCNA plugin[122] for PyMOL and the CNA web server[123] have been developed. Both provide user-friendly interfaces around the CNA software for easily setting up CNA runs and analyzing results. The CNA software and VisualCNA are available under academic licenses from http://cpclab.uni-duesseldorf.de/software, and the CNA web server is accessible at http://cpclab.uni-duesseldorf.de/cna.

KINARI

KINARI is a software package for rigidity analysis of biomolecules developed by Streinu and coworkers.[79] The goal of the software is to provide a workflow for rigidity analysis that is validated, versatile, and able to analyze different biomolecules in an automated and user-friendly way.[79] KINARI was first released as a web-based front end (KINARI-Web)[79] building upon the ideas of FIRST/ProFlex,[60] where the *bond-bending* network has been replaced by the *body-bar-hinge* network (see *Modeling and Analyzing Biomolecules as Constraint Networks*, Figure 2(d)) and the (6,6)-pebble game algorithm is applied to analyze these networks. Single and double bonds, amide bonds, and disulfide bonds are identified by KINARI using the identities and coordinates of the atoms, while hydrogen bonds are determined by the HBPLUS software package.[124] A user can remove constraints associated with a bond within a certain energy range or below/above a certain energy cutoff value.[79] In 2011, KINARI was extended to KINARI-Mutagen to analyze protein rigidity changes due to the mutation of a residue to glycine (see *Constraint Dilution Simulations to Investigate Protein (Un-)folding*).[13] To further extend the scope of the analysis, Fox et al. introduced the option of studying protein–nucleic acid complexes in KINARI-Web.[85] However, here the authors used the original protein-based parameters for finding and modeling hydrophobic interactions in RNA, which may lead to overly rigid RNA structures.[62,85,86] In 2015, KINARI-2 was released to improve the curation of the biomolecular structures for analysis, with the aim to have KINARI-2 succeed on a very high percentage of the data available in the PDB, on structural ensembles as well as bioassemblies with a high degree of symmetry, and to include hydrogen bond dilution simulations.[86] KINARI-Web is accessible at http://kinari.cs.umass.edu.

ENSEMBLE-BASED APPROACHES

Initially, studies using FIRST and KINARI were performed on constraint networks derived from single input structures. However, computing flexibility and rigidity characteristics from a single structure can be challenging because rigidity analysis of biomolecules is in general sensitive to the structural information used as input.[68,91,107,121] This is because biomolecules have a soft matter-like character where noncovalent interactions frequently break and (re-)form.[125] Furthermore, they are generally marginally stable, that is, their network state is close to the rigidity percolation threshold.[1] Accordingly, a few constraints more or less can result in a network either being rigid or flexible. This sensitivity problem can be overcome by analyzing an ENT rather than a single-structure network, where the ENT can be based on a structural ensemble obtained from experimental sources, for example, crystal structure analysis[107] and NMR,[121] or molecular simulations.[68,102] This way, however, the experimental or computational burden compromises the efficiency of the rigidity analysis. Therefore, computationally more efficient alternatives have been introduced that generate ENT from a single input structure,[107,112] essentially modeling the 'flickering' of noncovalent constraints[66,107] rather than the motions of atoms.

ENT from Fuzzy Noncovalent Constraints

The ENT^{FNC} approach, available within CNA,[84] performs rigidity analysis on ENT generated from a single input structure.[107] The ENT is based on definitions of fuzzy noncovalent constraints (FNC) derived from persistency data of noncovalent interactions from MD simulations. Therefore, the approach considers thermal fluctuations of a biomolecule without actually sampling conformations. The FNC model consists of two parts related to the modeling

of hydrogen bonds and hydrophobic tethers in biomolecules. To account for the thermal fluctuations of hydrogen bonds (a) probabilities, specific for the hybridization state of donor and acceptor atoms and the secondary structure they are located in, determine the persistence of a hydrogen bond across the ENT, and (b) a Gaussian white-noise component is added to each E_{HB} in order to modulate the order with which hydrogen bonds are removed during a constraint dilution simulation. Hydrophobic tethers are modeled by a distance-dependent, Gaussian-based probability by which tethers between closer atoms are included with a higher probability in a network than those between atoms further apart. Gaussian distributions have previously been applied for modeling the strength of pairwise interactions between hydrophobic atoms.[126–128] For the training system hen egg white lysozyme, a good agreement between local flexibility and rigidity characteristics from ENT^{FNC} and MD simulations-generated ensembles was found.[107] Regarding global characteristics, convincing results were obtained when relative thermostabilities of citrate synthase and lipase A proteins were computed, both retrospectively[107,108] and prospectively.[129] Compared to an ENT based on MD simulations-generated conformations, the ENT^{FNC} approach is ~300 times more efficient for a system with ~13,000 atoms. However, as a downside, it can only mimic the flickering of noncovalent bonds starting from a single conformational state of the biomolecule such that influences due to gross conformational changes will be missed. Thus, the ENT^{FNC} approach should be most suitable for comparing biomolecular systems where major conformational changes are not expected.

ENT using Mean Field Landau Theory

Jacobs introduced DCM (see *Constraint Counting: The Pebble Game Algorithms*), which is similar in spirit to the ENT^{FNC} approach.[112] In DCM, thermal fluctuations in constraint networks are modeled by fluctuating constraints at finite temperature without having to generate atomic coordinates for each conformation. To this end, mean field probabilities of bond and torsion constraints are used to calculate the mean field Landau free energy over an ensemble of constraint networks generated from Monte Carlo sampling. Covalent interactions are treated as quenched distance constraints because they never break under physiological conditions and thus do not contribute to thermal fluctuations. In contrast, noncovalent interactions frequently break and (re-)form. Each fluctuating constraint in DCM is assigned an enthalpy and entropy contribution in order to reproduce heat capacity curves of biomolecules from experiments.[111] The sequence of how fluctuating constraints are placed is based on the assignment of entropy from strongest to weakest.[112] Constraints are recursively added one by one to the constraint network until the structure is rigid. The DCM ensemble generation procedure was about a billion times faster than MD simulations, when it was introduced in 2005.[66] However, similar to ENT^{FNC}, it can only mimic the flickering of noncovalent bonds such that influences due to gross conformational changes will be missed.

Generation of Effective Constraint Networks

The virtual pebble game (VPG) is another ensemble-based rigidity analysis approach, similar to ENT^{FNC} and DCM.[130] It uses a single input structure for which an effective constraint network is calculated from a Monte Carlo-derived ENT, that is, the possible number of constraints that can form between a pair of nodes over the ENT is replaced by the average number. The effective network is thus considered having weighted edges, where the weight of an edge quantifies its capacity to absorb DOF. The VPG is then interpreted as a flow problem on this effective network.[130] Application of the VPG on a set of 272 nonredundant protein structures yields rigidity characteristics that are comparable with ensemble-averaged results obtained with the regular pebble game.[130] However, the VPG suppresses fluctuations of network rigidity and, hence, tends to be less accurate at the rigidity percolation threshold where most of these fluctuations occur.[131] This may be a drawback when analyzing biomolecules that are marginally stable,[1] as their network states are close to the rigidity percolation threshold.

A distantly related approach was presented by Mamonova et al.,[91] where an effective network is generated based on the time-dependent behavior of noncovalent bonds in the course of short (8 nanoseconds long) MD simulations. Subsequently, a single constraint network is constructed as input for rigidity analysis, considering only the most frequent noncovalent interactions.[91] Alternatively, the lifetime of noncovalent interactions can be derived from H/D exchange data as shown by Sljoka et al.[121] Depending on their strength and lifetime from the NMR measurements, hydrogen bonds are modeled with a different number of bars ranging from 1 to 5 to improve the input information for creating the constraint network in FIRST.[121] The drawback of the

last two methods is that they either require ensemble information from either a computationally expensive MD simulation or H/D exchange NMR experiments.

APPLICATIONS

Since FIRST was released, numerous studies on the flexibility and rigidity of biomolecules have been performed. Initially, these studies were primarily done for validation; subsequently, the different approaches described above were broadly used to foster our understanding of biomolecular structural stability and function.

Single-point Rigidity Analysis on Biomolecules

In the most direct way, constraint network representations of biomolecules can be analyzed as 'single-points', that is, the constraint network is derived from a single input structure, and no constraint dilution simulation is performed. The single-point studies can be used to investigate biomolecular function or changes in biomolecular flexibility and rigidity due to ligand binding or mutations.

The accuracy of single-point analysis strongly depends on the placement of noncovalent constraints in the network representation. In particular, the accurate placement of hydrogens, which are generally not available from X-ray diffraction experiments, is important for evaluating the inclusion of hydrogen bonds in the constraint network.[60] To this end, Thorpe et al.[82] compared hydrogen positions and resulting hydrogen bonds of five different trypsin structures from neutron diffraction experiments with those resulting from hydrogens placed by the program WhatIf.[132] At a cutoff E_{cut} = −0.6 kcal mol^{-1}, which corresponds to a network state at room temperature, only 6% of the hydrogen bonds were assigned differently in both methods. Alternatively, methods such as REDUCE[133] or the H++ web server[134] have been used to prepare biomolecules for rigidity analyses.[79,87,115]

Jacobs et al. applied single-point rigidity analysis by FIRST to datasets of ligand-bound HIV protease, dihydrofolate reductase, and adenylate kinase structures.[60] The computed flexibility and rigidity characteristics captured much of the functionally important conformational flexibility observed experimentally.[60] In an extensive study, Tan and Rader applied FIRST to analyze the rigidity of a dataset of 22 HIV-1 gp120 structures.[15] By studying altered flexibility and rigidity characteristics due to strain variation, stabilizing mutations, and binding events, the authors identified stable regions in gp120 that could serve as targets for vaccine design and drug discovery.[15] Along these lines, Metz et al. showed that the single-point analysis on the protein–protein interface of interleukin-2 correctly identifies regions as flexible that are required for opening a transient pocket.[135] Recently, Raschka et al. used rigidity analysis to measure the relative interfacial rigidity of docking poses from small-molecule ligands in a set of 19 diverse protein structures.[136] The authors stressed the importance of interfacial rigidification of the native binding mode in protein–ligand complexes, which, when used as scoring method for discriminating near-native poses from decoy poses in docking experiments, performs competitively to commonly used scoring functions. Information from a static single-point analysis has also been used by Thorpe et al. to study the dynamics of HIV-1 protease by unbiased Monte Carlo sampling on flexible regions.[82] Based on this result, several sampling methods emerged for exploring a biomolecule's conformational space; these are reviewed in section: *Rigidity Analysis to Coarse-grain Biomolecules Prior to Conformational Sampling*.

The overall performance of rigidity analysis by FIRST has been demonstrated by Hespenheide et al.,[61] where the structural rigidity of the pentameric and hexameric substructures of the cowpea chlorotic mottle virus (CCMV) protein capsid was analyzed. The considerable size of the viral capsid (~280 Å diameter) and the symmetrical, repetitive structure required a novel network representation, the *body-and-bar* network, together with a more efficient 3D pebble game algorithm (see *Constraint Counting: The Pebble Game Algorithms*).[61] The rigid cluster decomposition showed that the pentameric substructure forms a large central rigid cluster, able to form a sturdy capsid to protect the CCMV. When another subunit is added, the hexamer loses its rigidity, and capsid formation is inhibited.[61]

Single-point rigidity analysis performed on single input structures may be misleading because even subtle conformational changes between input structures can have pronounced effects on the results.[87] This sensitivity problem can be overcome by single-point rigidity analyses on structural ensembles. Along these lines, Gohlke et al. generated conformational ensembles from MD trajectories of Ras, Raf, and Ras/Raf.[68] Averaging the results from rigidity analysis over the structural ensembles, the authors showed that stabilization upon Ras/Raf complex formation is not locally restricted but rather extends to regions that do not make any direct interactions with the respective binding partner. This finding manifested

the long-range aspect of rigidity percolation in biomolecules, which is also important for investigating allosteric signaling (see *Analysis of Allosteric Coupling*). In an alternative approach, Mamonova et al. computed an average constraint network, based on the persistence of noncovalent interactions along MD trajectories.[91] In the case of barnase, the predicted stability characteristics compared well with NMR experiments but showed limitations when the system underwent a conformational change, for example, upon ligand binding, as demonstrated for GluR2.[91] As a further alternative, an average constraint network can be directly generated from NMR ensembles.[121] Sljoka and Wilson showed that results obtained from a rigid cluster decomposition on such a network are in good agreement with experimental H/D exchange data.[121]

The DCM allows for sampling ensembles of constraint networks at finite temperature starting from a single input structure (see *Distance Constraint Model*).[111] DCM has been applied to study the correlated flexibility within the active site of class A,[137] B,[138] and C[139] families of β-lactamases. For all three classes the authors could show that the backbone flexibility is highly conserved across the families, while the cooperativity correlation, which indicates a residue's pairwise mechanical coupling within the structure, is, at least partially, conserved in the active site across members of the C class family.[139] Following the idea of using structural ensembles from MD simulations as input,[68] DCM has been applied to characterize the effect of stabilizing mutations within an antibody single chain Fv (scFv) fragment of the anti-LTβR antibody.[140] The study demonstrated that local mutational perturbation often leads to distant altered stability characteristics.

In order to study biomolecular thermostability, Livesay and Jacobs used DCM (see *Distance Constraint Model*) to introduce the notion of quantitative stability/flexibility relationships (QSFR) and study enthalpy-entropy compensation in homologous meso- and thermophilic RNAse H structures.[76] The authors found that the thermophilic protein is more stable than its mesophilic counterpart at any given temperature. However, the local stability profiles are markedly similar for the homologs at appropriately shifted temperatures, which is in agreement with H/D exchange experiments and the 'principle of corresponding states'. Verma et al. then used DCM to analyze melting points of human c-type lysozyme and 14 variants.[117] The DCM results showed that changes in human c-type lysozyme flexibility upon mutation are frequent, large, and long-ranged. With this retrospective study, it was demonstrated that DCM can be a viable predictor for the relative stability of protein variants. In another retrospective study, Li et al. analyzed the thermodynamic stability and flexibility characteristics of a dataset consisting of the variable domain (VL), the scFv fragments, and the fragment antigen-binding (Fab) fragments with DCM.[118] In this work, DCM was extended to analyze incomplete thermodynamic data. This development allowed high throughput QSFR studies in a large data set of antibody fragments and complexes.

Single-point Rigidity Analysis on RNA and Nucleic Acid–Protein Complexes

While most rigidity analyses are performed on proteins, the approach can also be used to study RNA structures and nucleic acid–protein complexes. Wang et al. applied rigidity analysis to the ribosome to investigate the flexibility in the ribosomal subunits.[94] To do so, the constraint definition for proteins was only slightly modified (see *Modification of the Constraint Network Representation for RNA Structures*). The authors compared FIRST and CG-based elastic network models (ENM), and observed that both methods successfully predicted the flexibility of functional key areas of the ribosome subunits. A study by Fulle et al. focused on the exit tunnel within the large ribosomal subunit, for which FIRST with an adapted RNA parameterization (see *Modification of the Constraint Network Representation for RNA Structures*) was applied.[64] The results revealed a sophisticated interplay between the static properties of the ribosomal exit tunnel and its functional role in cotranslational processes. The authors showed that considering flexibility characteristics of the antibiotics binding sites within the tunnel is required for explaining the observed binding selectivity of antibiotics.[10,64] Further applications of rigidity analysis on RNA relate to the natural coarse-graining of the structure, which is used for setting up simulations to generate conformational ensembles (see *Rigidity Analysis to Coarse-grain Biomolecules Prior to Conformational Sampling*). Prominent examples dealt with the creation of molecular-replacement search models for nucleic acids,[141] and conformational sampling of the SAM-I riboswitch aptamer domain[142] and the HIV-1 TAR RNA.[72]

Rigidity Analysis to Coarse-Grain Biomolecules Prior to Conformational Sampling

The extent of conformational changes in biomolecules ranges from fast atomic fluctuations on the

pico- to nanosecond timescale to domain movements on the micro- to millisecond timescale.[143] Despite recent major improvements, modeling large conformational transitions in biomolecules by MD simulations is still computationally costly. As a more efficient alternative, CG simulation methods have emerged, which work on systems with a reduced number of DOF. Frequently, the coarse-graining is based on a per-residue or per-secondary structure level; coarse-graining based on molecular shape is another possibility.[144,145] Alternatively, rigid regions identified by rigidity analysis within a biomolecule provide a very natural way of coarse-graining.[146] The constrained geometric simulation method FRODA[73] and its predecessor ROCK (Rigidity Optimized Conformational Kinetics)[147] explore the geometrically accessible conformational space of a CG biomolecule through diffusive motions. ROCK generates new biomolecular conformations by random movements within flexible regions and satisfying ring closure equations, whereas FRODA makes use of a more efficient algorithm where rigid regions within the biomolecule are replaced by 'ghost templates.' Overall, both approaches result in random walks on energy landscapes that are flat where bond and angle constraints are fulfilled, and infinitely high elsewhere. FRODA has been used for studying complex movements of membrane ion channels[148,149] and correlated motions between functionally relevant elements in a pigment–protein complex,[150] monitoring the intrinsic flexibility of myosin in the actin-attached and actin-detached state,[151] protein–protein docking involving multiple conformational changes,[152] identifying the opening of transient pockets in protein–protein interfaces,[135] investigating the essential dynamics of unbound and bound HIV-1 TAR RNA structures,[72] and fitting of X-ray structures to cryo-EM maps of GroEL.[153] A downside of FRODA is that generated conformational ensembles are not sampled from a thermodynamic ensemble. Accordingly, FRODA was combined with MD simulations to search for and refine native-like topologies of small globular α-, β-, and α/β-proteins.[154] CONCOORD,[155] and its successor tCONCOORD,[156] are other geometry-based approaches that generate new conformations by satisfying distance constraints derived from experimental structures of biomolecules. However, they do not apply a CG biomolecule representation, and thus, are not further discussed here.

As the FRODA approach lacks any directional guidance for sampling the biologically relevant conformational space, reaching a certain distance $n * d$ with steps of a given length d requires n^2 such steps, which limits the sampling particularly in those cases where biomolecules are very flexible. Information about directions of biomolecular motions can be derived from NMA,[157] which has been used to study large-amplitude motions in biomolecules for decades.[26,158,159] Combining directional guidance from harmonic analysis and atomistic simulations led to MD/NMA hybrid methods,[160–162] where collective motions are amplified along normal mode directions. ENM have emerged as efficient alternatives to NMA; here, simplified force-fields[163] and CG biomolecular representations are used.[71,164–170] Integrating all these ideas led to the normal mode-based geometric simulation approach NMSim, which is a three-step protocol for multiscale modeling of protein conformational changes.[171] Initially, static properties of the protein are determined by decomposing the molecule into rigid clusters and flexible regions using FIRST.[82] In a second step, dynamical properties of the molecule are revealed using an ENM representation of the coarse-grained protein (RCNMA approach).[71,172] In the final step, the idea of constrained geometric simulations of diffusive motions in proteins[73] is extended in that new protein conformations are generated by biasing backbone motions toward directions that lie in the subspace spanned by low-frequency normal modes. The generated structures are then iteratively corrected regarding steric clashes and violations of constraints for covalent and noncovalent bonds. In total, when applied repetitively over all three steps, the procedure efficiently generates a series of stereochemically correct conformations that lie preferentially in the subspace spanned by low-frequency normal modes.[171] Recently, NMSim has been used to sample the large-scale domain motions during phosphate group transfer in the pyruvate phosphate dikinase (PPDK). From this, an unknown intermediate state of PPDK has been identified, which was confirmed by X-ray crystallography.[173] In connection with quantitative FRET studies and integrative structure modeling, NMSim has been used for unbiased and FRET-guided generation of structural ensembles.[174] NMSim is accessible via a web server at http://cpclab.uni-duesseldorf.de/nmsim.[175] In a very similar approach subsequently introduced, FRODA simulations were guided by low-frequency modes derived from NMA.[176] The approach was successfully applied in studying protein folding[177] and conformational transitions in biomolecules.[178–180]

Another limitation of the original FRODA approach is the fixed constraint topology, that is, noncovalent constraints cannot break or (re-)form

during simulations, which limits the extent of conformational transitions that can be sampled. FRODAN, a recent re-implementation and extension of FRODA, models noncovalent interactions as maximum-distance constraints that become breakable if they exceed a certain amount of strain, which has been successfully used in targeted simulations between two known conformational states.[74]

Similar in spirit to the FRODA method are approaches that combine constrained geometric simulations with concepts from robotics motion planning[181] or tensegrity principles.[182] The kino-geometric conformation sampler (KGS) is a robotic-inspired, Jacobian-based method for the deformation of interdependent kinematic cycles.[183] Kinematic cycles are connected circular components in biomolecules spanned by (non)covalent interactions. The KGS has been used for sampling the activation pathway of $G_{\alpha s}$ alone and in complex with a GPCR.[184] A variant for RNAs (KGS_{RNA}) correctly reproduced the conformational landscape of noncoding RNA molecules in agreement with NMR experiments.[185,186] In addition, KGS_{RNA} was used to identify transient, exited states of the HIV trans-activation response element.[186] The EASAL (Efficient Atlasing, Analysis, and Search of Molecular Assembly Landscapes) approach is an example where conformations are sampled based on tensegrity principles. Here, structural systems are established where a set of discontinuous compressive components interacts with a set of continuous tensile components to define a stable volume in space.[187] EASAL has been developed for exploring and analyzing high dimensional configuration spaces of biomolecular assemblies and was applied for studying intermonomer interactions of viral capsid assembly[188] and sampling the assembly landscape of two transmembrane helices.[189]

Rigidity Analyses on Perturbed Constraint Networks

The above rigidity analyses were performed on constraint networks in the 'ground state,' that is, as generated from a given biomolecule conformation. Comparing perturbed networks to a 'ground state' network yields additional information in terms of the effect of the perturbation on the rigidity characteristics. Perturbations can affect the constraint network directly, for example, due to removing constraints, inserting a mutation, binding of a ligand, or indirectly, for example, in terms of modeling the influence of temperature on the presence or absence of noncovalent interactions.

Constraint Dilution Simulations to Investigate Protein (Un-)Folding

Information on the heterogeneity of biomolecular stability is obtained by monitoring the decay of network rigidity along a constraint dilution trajectory (see *Analyzing Network States along Constraint Dilution Trajectories*). The gradual removal of noncovalent interactions to generate such a trajectory can be considered a repetitive network perturbation. In 2002, Rader et al. used FIRST and such a perturbation scheme to describe the rigid-to-flexible transition upon the (simulated) unfolding of 26 structurally and functionally different proteins.[83] The authors observed that the phase transitions of all proteins from an overall rigid to a flexible state occur at a similar mean coordination of the atoms and are furthermore analogous to phase transitions found in network glasses.[83] This indicates that, despite their diverse architectures, proteins and network glasses reveal a universal percolation behavior. In two other studies, constraint dilution trajectories generated by FIRST were used to identify folding cores in protein datasets.[69,70] A folding core was defined as the most stable region along the constraint dilution trajectory involving at least two secondary structures.[69] The identified folding cores from both studies were compared with experimentally identified folding cores from H/D exchange experiments, which yielded a very good agreement[69] and an enhancement over random correlation.[70] Subsequently, Rader et al. used FIRST for analyzing folding cores in rhodopsin (Table 1).[190] For this transmembrane protein, the constraint network definition originally introduced for soluble proteins was used. The authors showed that the stable core of the protein contains residues that cause misfolding upon mutation.

Constraint Dilution Simulations to Investigate Protein Thermostability

Monitoring the decay of network rigidity along a constraint dilution trajectory (see *Analyzing Network States along Constraint Dilution Trajectories*) helps to improve the understanding of the relationship between biomolecular structure, activity, and thermostability, which has become important for rational protein engineering.[193,194] Biomolecular thermostability can have a thermodynamic or kinetic origin.[195] In all studies reported below, rigidity analysis was used to investigate only the effect of mutations on the folded state. This was done because rigidity analysis cannot account for the time-dependency of processes,[91] and it is very challenging to generate

TABLE 1 | Selected Applications of Rigidity Analysis to Biomolecules

Author	Data Set/Protein	Application	Experimental Data	Computational Data
Single-point rigidity analysis on proteins				
Jacobs et al.[60]	HIV-1 protease, adenylate kinase, and dihydrofolate reductase	Analyze the flexibility of proteins	Thermal mobility (B-factor) from X-ray crystallography	FIRST, flexibility index f_i
Hespenheide et al.[61]	CCMV protein capsid	Study rigidity of capsid proteins	X-ray crystal structure	FIRST, rigid cluster decomposition
Gohlke et al.[68]	H-Ras and C-Raf1, *apo* states and protein–protein complex	Determine changes in flexibility upon protein–protein complex formation	X-ray crystal structures	FIRST, rigid cluster decomposition using MD-based ensembles
Mamonova et al.[91]	Barnase and GluR2	Compare stability characteristics with NMR data	X-ray crystal structures and NMR ensemble data	FIRST, rigid cluster decomposition from average constraint network based on MD ensembles
Sljoka and Wilson[121]	Acylphosphatase	Compare stability characteristics with H/D exchange data	NMR structures and H/D exchange data	FIRST, rigid cluster decomposition and H/D exchange profile
Li et al.[118]	One VL, three scFv and five Fab antibody fragments	Analyze thermodynamic stability and flexibility of antibody fragment complexes	Heat capacity curves	DCM, cooperativity correlation *CC*
Verma et al.[116]	Wild type human c-type lysozymes, 14 variants with point mutations	Predict the stability of a series of variants	Experimental heat capacity curves C_p	DCM, total conformational entropy S_{conf}, backbone flexibility index *FI*, cooperativity correlation *CC*
Single-point rigidity analysis on RNA and nucleic acid–protein complexes				
Wang et al.[94]	Ribosomal subunits	Investigate flexibility and function of the ribosome, compare FIRST and ANM	X-ray crystal structures	FIRST, rigid cluster decomposition and anisotropic network model (ANM)
Fulle and Gohlke[62]	Ribosomal exit tunnel	Study functional role in cotranslational processes	X-ray crystal structures	FIRST, rigid cluster decomposition using RNA parameterization
Rigidity analyses on perturbed constraint networks				
Rader et al.[83]	26 proteins with different CATH architecture[1]	Loss of structural stability in biomolecules	Unfolding behavior of network glasses upon melting	FIRST, floppy mode density ϕ
Rader et al.[190]	Rhodopsin	Analyze folding cores in biomolecules	Folding cores predicted by H/D exchange NMR experiments	FIRST, rigidity order parameter P_∞, FIRST dilution plots

(*continued overleaf*)

TABLE 1 | Continued

Author	Data Set/Protein	Application	Experimental Data	Computational Data
Investigating protein thermostability				
Radestock and Gohlke[75]	20 pairs of homologous proteins from mesophilic and (hyper-) thermophilic organisms	Analyze the shift in thermostability of pairs of orthologous proteins and identify weak spots in biomolecules	Optimal growth temperatures of the organism or experimentally determined melting temperatures	CNA, rigidity order parameter P_{∞}, cluster configuration entropy H_{type2}
Rader[99]	Rubredoxin structures from the hyperthermophile *P. furiosus* and mesophile *C. pasteurianum*	Analyze thermostability and folding cores, which are responsible for biomolecular stability under extreme environmental conditions.	Folding cores from H/D exchange NMR studies, mutation experiments	FIRST, rigidity order parameter P_{∞}, cluster configuration entropy H_{type1}, FIRST dilution plots, largest rigid cluster propensity P_{lrc}
Radestock and Gohlke[14]	20 pairs of homologous proteins from mesophilic and (hyper-) thermophilic organisms	Analyze flexibility conservation of substrate-binding pockets in enzymes	Same dataset as in Radestock and Gohlke 2008,[75] but using only one pair of structures for each protein family	CNA, stability maps rc_{ij}
Rathi et al.[102]	Five citrate synthase (CS) structures over a temperature range from 37°C to 100°C	Study thermostability within a series of orthologous CS structures and compare predicted weak-spots	Optimal growth temperatures of the organism or experimentally determined melting temperatures	CNA, rigidity order parameter P_{∞}, cluster configuration entropy H_{type2}
Dick et al.[191]	Orthologs from psychrophilic, mesophilic and hyperthermophilic 2-desoxy-D-ribose-5-phosphate aldolase (DERA)	Analyze influence of dimer interface on thermostability and flexibility on substrate access	First crystal structures of psychropilic DERAs, mutation experiments, generating monomeric DERAs, activity assays	CNA, cluster configuration entropy H_{type2}
Prospective application to improve protein thermostability				
Rathi et al.[108]	16 variants of lipase A from *B. subtilis*	Validate thermostability prediction for highly similar variants	Experimentally determined melting temperatures	CNA, percolation index p_i, cluster configuration entropy H_{type2}, median stability of rigid contacts $\widetilde{rc}_{ij,neighbor}$, clustering of unfolding pathways
Rathi et al.[129]	Twelve variants of lipase A from *B. subtilis*	Identify weak spots and predict mutations increasing thermostability	Experimentally determined melting temperatures	CNA, percolation index p_i, cluster configuration entropy H_{type2}
Analysis of allosteric coupling				
Mottonen et al.[65]	Protein CheY	Explore allosteric effect across protein families	X-ray crystal structures and melting temperatures which are used for fitting parameters in DCM due to missing heat capacity curves for CheY	DCM, ΔFI (flexibility index) and ΔCC (cooperativity correlation)
Verma et al.[117]	Wild type human c-type lysozymes, 14 variants with point mutations	Investigate changes in protein flexibility upon single point mutations	Mass spectrometry, H/D exchange NMR experiments, mutation	DCM, backbone flexibility index *FI*, cooperativity correlation *CC*, B-factor

(*continued overleaf*)

TABLE 1 | Continued

Author	Data Set/Protein	Application	Experimental Data	Computational Data
			studies, differential scanning calorimetry	
Hanke and Gohlke (unpublished)	Aptamer domain of the guanine-sensing riboswitch	Investigate aptamer function and the allosteric pathway through the riboswitch	X-ray crystal structure, NMR data on hydrogen bonds	FIRST, rigid cluster decomposition

[1] CATH architecture: alpha, beta, and mixed alpha and beta.

realistic structural models of the unfolded state of a protein.[196] Still, applying rigidity analysis that way provides a wide range of applicability for studying thermostability because increased structural rigidity is in 60% of the cases responsible for increased thermostability.[129]

Radestock et al.[14,75] analyzed protein thermostability of pairs of homologous proteins from mesophilic and thermophilic organisms (Table 1) using CNA. The authors described the macroscopic percolation behavior and predicted phase transition temperatures (T_p) by monitoring the cluster configuration entropy (H) and the rigidity order parameter (P_∞) (see *Global and Local Indices for Characterizing Biomolecular Stability*) during constraint dilution simulations. The comparison between predicted T_p values and optimal growth temperatures of the corresponding organisms (T_{og}) revealed that in two-thirds of the pairs, a higher T_p was predicted for the thermophilic than for the mesophilic homolog.[75] At the microscopic level, the authors identified structural features from which a destabilization originates ('weak spots'), which is very helpful for guiding mutation experiments when prospectively engineering thermostability (see below). From both global and local stability characteristics the authors provided direct evidence for the 'principle of corresponding states,' according to which mesophilic/thermophilic homologs have similar flexibility and rigidity characteristics at the respective T_{og}.[14,75] In addition, by monitoring the local distribution of flexible and rigid regions using stability maps rc_{ij} (see *Global and Local Indices for Characterizing Biomolecular Stability*), adaptive mutations in enzymes were shown to maintain the balance between global (structural) stability, in favor of overall thermostability, and local flexibility, in favor of activity, at appropriate enzyme working temperatures; this important information provides guidelines for what (not) to mutate in prospective studies.[14] Later, Rader[99] applied FIRST in a similar manner to analyze structural mechanisms behind thermostability differences in two homologous structures of rubredoxin (Table 1).[99] The obtained results supported the 'principle of corresponding states' in biomolecular thermostability. On a local level, the study depicted differences in structural stability of the homologs, which agreed with protection factors from H/D exchange experiments.[99]

Extending these studies to series of protein variants, Rathi et al. studied the relationship between structural rigidity and thermostability of citrate synthase from five different species with T_{og} ranging from 37°C to 100°C (Table 1).[102] CNA was applied to conformational ensembles generated by MD simulations. The authors obtained a good correlation (R^2 = 0.88) between predicted T_p and experimental T_{og}. This finding validates that CNA is able to quantitatively discriminate between less and more thermostable proteins even within a series of orthologs. Furthermore, from a local point of view, the study revealed that structural weak spots predominantly occur at sequence positions with a high mutation ratio. Dick et al. applied CNA to study the thermal adaptation of 2-deoxy-D-ribose-5-phosphate aldolase (DERA) originating from psychrophilic to hyperthermophilic organisms (T_{og} = 8 – 100°C).[191] The comparison between predicted T_p and experimental T_{og} revealed a very good correlation (R^2 = 0.97). Interestingly, the authors identified, and validated by experiment, that interface stability contributes to thermostability in the dimeric DERA structures from (hyper)thermophilic organisms. This may be exploited as a design principle when engineering thermostability in multimeric proteins.

Rigidity Analysis on Structurally Perturbed Constraint Networks

So far, perturbations were performed directly on the network by gradually removing constraints associated with noncovalent interactions. Extending the perturbation idea to structural effects, for example, due to a mutation or ligand binding, allows for testing the influence due to adding/removing constraints to/from the network without actually changing the conformation of the 'ground state' structure. This provides an excellent means for investigating alteration in

biomolecular stability upon mutations or binding events in a computationally efficient manner.

Mutation Influences on Unfolding Free Energies

In 2011, KINARI was extended with KINARI-Mutagen (see *KINARI*), allowing for excision mutation studies, essentially mutating (perturbing) a residue of choice to glycine.[13] Here, all noncovalent interactions belonging to the side-chain of the mutated residue are removed from the constraint network. In a first case study, the authors showed that KINARI-Mutagen was able to identify functionally critical residues in crambin based on altered stability characteristics, even though the residues are partially exposed to the solvent accessible surface. In a second case study, predicted changes in stability characteristics upon mutating residues in T4 lysozyme correlate with experimental free energies ($\Delta\Delta G$) of unfolding. Recently, an ensemble-based approach has been implemented in CNA to predict changes in the free energy of biomolecular stability (C. Pfleger, H. Gohlke, unpublished results). The approach combines constraint dilution simulations with structural perturbations due to *in silico* alanine mutations. For a set of 13 single and double mutation variants of eglin c, the predicted free energy changes yield a very good correlation with those from chemical denaturation experiments. Remarkably, almost all mutations involved changes from valine to alanine, demonstrating that it is possible to detect mutation effects in a position-dependent manner even if the type of mutations are similar or identical.

Prospective Application to Improve Protein Thermostability

With the aim to further develop CNA for prospective studies on improving thermostability, Rathi et al. analyzed the thermodynamic stability of a set of 16 variants of lipase A from *Bacillus subtilis*.[108] Eight variants were generated from the wild type structure of lipase A by solely altering the mutated residues while the orientation of neighboring residues was kept unchanged. Three results stood out from this analysis. First, (relative) thermodynamic stability was successfully predicted for variants that differ by only 3–12 mutations from the wild type. Second, a measure for the similarity/dissimilarity of unfolding pathways of variants was introduced for explaining false thermostability predictions (Figure 8). Third, the median stability of rigid contacts $\tilde{rc}_{ij,neighbor}$ was introduced as a new local measure for predicting thermodynamic stability. $\tilde{rc}_{ij,neighbor}$ represents the chemical potential energy due to noncovalent bonding, obtained from the CG, residue-wise network representation of the underlying protein structure. Additionally, the recently developed ENT^{FNC} approach[107] (see *ENT from Fuzzy Noncovalent Constraints*) was used for robust rigidity analysis, which makes it unnecessary to perform computationally demanding MD simulations for each variant. In a subsequent prospective study, Rathi et al. described a strategy to predict amino acid substitutions optimal for thermostability improvement; the predictions were experimentally validated (Table 1).[129] The strategy combines a structural ensemble-based weak spot prediction of the wild type protein by CNA, filtering of weak spots according to sequence conservation, computational site saturation mutagenesis, assessment of variant structures with respect to their structural quality, and screening of the variants for increased structural rigidity by ENT^{FNC}-based CNA. The strategy was applied to predict single-point variants of lipase A from *Bacillus subtilis* and yielded a success rate of 25% (60% when mutations from small-to-large residues and those in the active site were excluded) with respect to experimentally validated mutations that lead to increased thermostability. Notably, an increase in thermostability by 6.6°C compared to wild type due to a single mutation was found.

Analysis of Allosteric Coupling

Allostery is the process by which biomolecules transmit the effect of binding at one site to another, often distal, functional site.[200] Conventionally, models that explain allostery involve a conformational change upon binding of an allosteric effector molecule.[201,202] Over the last decades, the view of allostery has been extended to cover the role of entropy, which can occur in the absence of conformational changes.[203] Owing to the nonlocal character of rigidity percolation, adding constraints to one site of the network, that is, by binding of an allosteric effector, can affect the stability of sites all across the network.[47] Such an effect has first been demonstrated in the context of rigidity analysis on biomolecules for the protein–protein complex Ras/Raf,[68] where the stabilization of the binding partners also affected regions that do not make any direct interactions with the protein–protein interface. Inspired by this observation, a computationally highly efficient approximation of changes in the vibrational entropy (ΔS_{vib}) upon binding to biomolecules has been introduced recently, based on rigidity theory.[192] Here, ΔS_{vib} is estimated from changes in the variation of the number of F with respect to variations in the constraint networks' coordination number. Compared to ΔS_{vib} computed by NMA as a gold standard, this approach yields significant and good to fair correlations for datasets of protein–protein and protein–small-molecule complexes as well as in alanine scanning. This approach may thus serve as a valuable alternative to

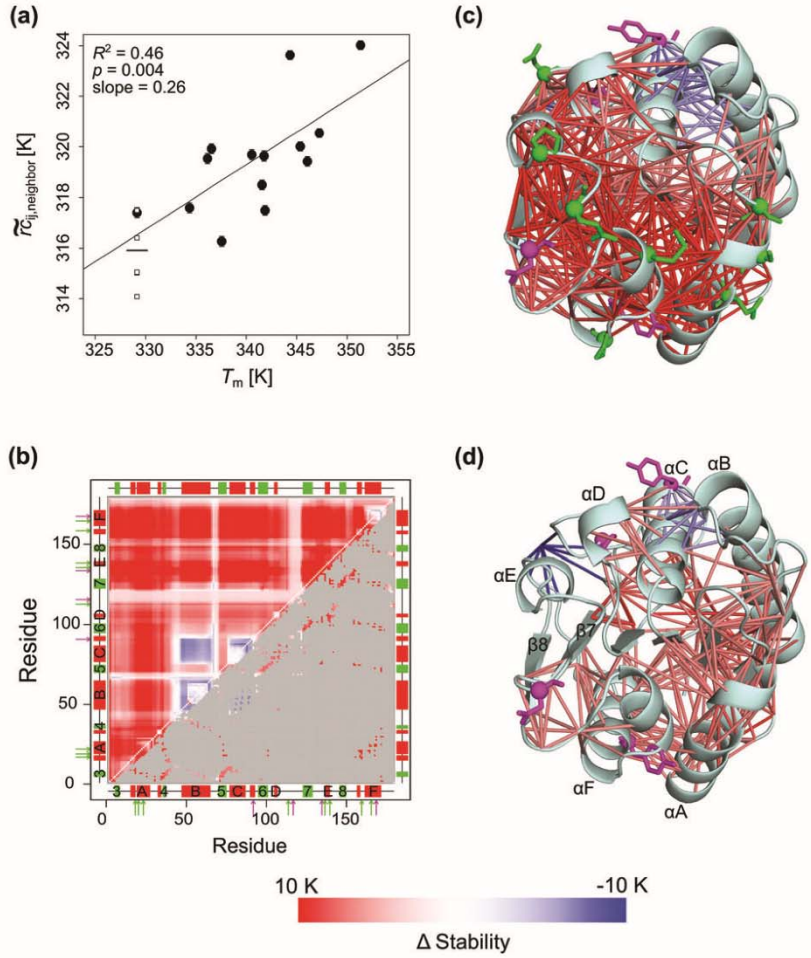


FIGURE 8 | Application of rigidity theory to investigate protein thermostability. (a) Correlation between $\tilde{rc}_{ij,neighbor}$, a local measure for predicting thermodynamic stability, and experimental thermostabilities (T_m) for the six wild type crystal structures (empty squares) and thirteen variants of the *Bacillus subtilis* lipase A. For the six wild type crystal structures, the resulting mean $\tilde{rc}_{ij,neighbor}$ is shown as a horizontal bar. Experimental values were taken from Refs 197–199. Error bars depict the standard error in the mean. (b) Stability map of the variant 6B, T_m of which is 6.6 K higher than that of the wild type. A red/blue color shows that a rigid contact in the variant is more/less stable than in the wild type (see color scale). The upper triangle shows differences in stability values for all residue pairs, and the lower triangle shows differences in stability values only for residue pairs that are within 5 Å of each other. Secondary structure elements are indicated on both abscissa and ordinate and are labeled: α-helix (red rectangle), β-strand (green rectangle), loop (black line). Arrows represent the mutation positions with respect to the wild type sequence. (c, d) Structures of the variants 6B (c) and 1–14F5 (d); T_m of 1–14F5 is 2.1 K higher than that of the wild type. Common mutations in 6B and 1–14F5 are shown in magenta, unique mutations in 6B are shown in green. The differences in the stability of rigid contacts for residue neighbors is displayed by sticks connecting C_α atoms of residue pairs colored according to the scale shown in panel (b); only those contacts that are stabilized by ≥ 4 K or destabilized by ≥ 3 K are shown for clarity. Figure adapted from Ref 108.

NMA-based ΔS_{vib} computation in free energy calculations.

In an extensive study, DCM (see *Distance Constraint Model*) was applied on three bacterial chemotaxis protein Y (CheY) proteins to explore the allosteric response across protein families.[65] A mechanical perturbation method (MPM) was introduced to simulate the binding of ligands by adding extra constraints to a certain site in the constraint network. The authors concluded that perturbed residues with large changes in stability characteristics are likely involved in allosteric signaling. From this, important residues for allosteric signaling were identified, with > 50% of them only occurring in a single ortholog. This finding demonstrates the

complex nature of allostery and might indicate that the conservation of allosteric mechanism exists only across short evolutionary distances. In a second study, the MPM was applied to identify putative allosteric sites in a set of six single chain-Fv fragments of the anti-lymphotoxin-β receptor antibody.[204] The findings from this study on monoclonal antibodies indicate that the allosteric response is sensitive to mutations through changes in the hydrogen bonding network, and results from rigidity analysis support what is found in practice when redesigning monoclonal antibodies either for function and/or thermodynamic stability.

Recently, an ensemble-based perturbation approach has been introduced for gaining a deeper structure-based understanding of the relationship between changes in static properties and allosteric signal transmission in biomolecules (C. Pfleger, H. Gohlke, unpublished results). Applying a free energy perturbation approach to results of rigidity analysis (see *Mutation Influences on Unfolding Free Energies*), free energies of cooperativity and pathways of allosteric signaling are computed. Notably, conformational changes of the biomolecule are excluded in this approach by definition in that *apo* conformations are generated by removing all constraints associated with ligands from the network of the *holo* structures (perturbation). The approach was successfully applied to two systems, lymphocyte function-associated antigen 1 (LFA-1)[205] and protein tyrosine phosphatase 1B (PTP1B),[206] showing ligand-based K- and V-type allostery, respectively. Upon perturbation, altered rigidity characteristics revealed long-range effects in both systems. Remarkably, clusters of residues were identified in both systems that form continuous pathways spreading from the allosteric site to the orthosteric site and to regions known to be important for protein function (Figure 9(a)). Finally, predicted free energies of cooperativity for binding of the allosteric and orthosteric ligands to LFA-1 revealed a nonadditive stabilization in agreement with the experimentally confirmed mechanisms of negative and positive cooperativity.[207,208]

As to nucleic acid systems, Fulle et al. proposed an allosteric signal transmission pathway within the large ribosomal subunit from the exit tunnel region to the peptidyl transferase center based on a hierarchy of regions of varying stabilities (Figure 10).[64] That is, signals are transmitted through structurally stable regions by inducing a conformational change in a domino-like manner. Two independent experimental studies later confirmed the mechanical coupling in the ribosomal tunnel region.[209,210] In another study, Hanke et al. used FIRST with the RNA parameterization[63] (see *Modification of the Constraint Network Representation for RNA Structures*) to investigate the interplay between the ligand

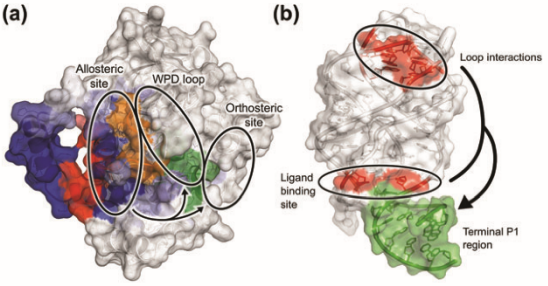


FIGURE 9 | Long-range coupling effects in RNA and protein. (a) Schematic representation of long-range allosteric coupling in the protein tyrosine phosphatase 1B (PTP1B). Upon perturbing the network at the allosteric site by adding constraints mimicking the binding of an allosteric modulator (red), altered stability characteristics are observed for the functionally important WPD loop (orange) and for residues in the orthosteric site (green). (b) Schematic representation of the long-range cooperative stabilization of the P1 region in the aptamer domain of the guanine-sensing riboswitch. Interactions within the tertiary loop-loop region (red) and of the ligand with the binding site (red) together are required to stabilize the terminal P1 region (green) (C.A. Hanke, H. Gohlke, unpublished results).

binding site, tertiary loop-loop interactions, and the switching sequence in the aptamer domain of the guanine-sensing riboswitch (C.A. Hanke, H. Gohlke, unpublished results). Starting from a structural ensemble of the *apo* aptamer domain, the stabilizing effect of the ligand was modeled by adding constraints in the network topologies at the ligand binding site, similar to the study on the CheY proteins.[65] The results suggest that the presence of the ligand has a stabilizing effect on the switching sequence (Figure 9b) and that this stabilizing effect is stronger for the wild type than for a variant in which tertiary interactions ~30 Å away from the ligand binding site had been perturbed. These findings suggest that the distant tertiary interactions and the ligand binding cooperatively stabilize the P1 region, and in this way influence the regulation of genes.

CONCLUSION/OUTLOOK

Studying static properties of biomolecules has come a long way, from Maxwell's mean field approach on constraint counting, the development of constraint network representations for biomolecules, methodological and algorithmic developments for analyzing such networks / characterizing biomolecular stability / linking these results to biomolecular function, and the introduction of software packages for performing rigidity analysis, to applications on biomolecules as complex as the ribosome, viruses, or transmembrane proteins. Key methodological steps along this way were: the realization of the influence of redundant constraints on

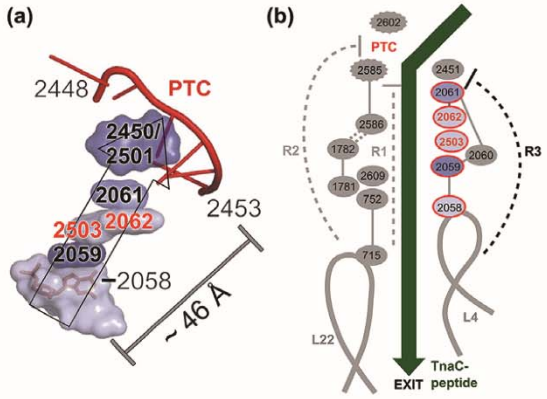


FIGURE 10 | Allosteric pathways in the ribosomal exit tunnel. (a) Rigid cluster decomposition of the allosteric pathway to the peptidyl transferase center (PTC) (red) as predicted by Fulle et al.[64] Different shades of blue correspond to different rigid clusters. Residues in orange were identified to be important for ribosome stalling in experiments.[209] Figure adapted from Ref 64. (b) Allosteric pathways for PTC silencing (R1, R2, R3) when the tryptophanase C (TnaC) peptide (green) is in the exit tunnel[210]; the grey loops marked L4 and L22 indicate ribosomal proteins. Residues that agree with the prediction of the rigidity analyses from (a) are colored accordingly and circled in red. Ribosomal components not identified in the rigidity analysis are colored in grey. Orange residues as in (a). Figure adapted from Ref 210.

constraint counting results and network properties, the development of rules to determine whether noncovalent interactions in biomolecules are strong enough to be included as a constraint, the development of efficient algorithms for determining the DOF in a constraint network locally, concepts to analyze network states along constraint dilution trajectories as well as to compare perturbed to 'ground state' networks, and the introduction of informative indices for linking results from rigidity analysis to biologically relevant characteristics of a structure. As to the applicability, several software packages with, in part, overlapping and, in part, unique features have been made available, and/or web servers have been developed. These software packages allow for generating constraint networks from given biomolecular structures, can consider ligands, ions, or structural water as part of the network, and enable single-point or ensemble-based rigidity analyses. Importantly, ensemble approaches were developed that model the 'flickering' of noncovalent constraints without the need to generate a structural ensemble. The ensemble approaches yield robust results and estimates of uncertainty for rigidity analyses on biomolecules but do not compromise the computational efficiency of such analyses. About 15 years after the first application of rigidity theory to biomolecules, in these authors' view, the field has thus reached a first level of maturity, and we encourage considering rigidity analyses more broadly as a computational biophysical method to scrutinize biomolecular function from a structure-based point of view and to complement approaches focused on biomolecular dynamics. In particular, its computational speed and the inherent long-range aspect to rigidity percolation make this method attractive to investigate signal transduction through biomolecules and distant influences on biomolecular stability.

While the constraint counting itself in terms of the family of (k,l)-pebble games was proven to be correct, modeling constraint networks from given biomolecular structures remains both art and science, similar to force field development in the area of molecular mechanics.[211] Particularly, a biomolecule system-independent parameterization for when to consider a constraint is required for making rigidity analyses broadly applicable. Along these lines, the current parameterizations available in the software packages FIRST/ProFlex, DCM, CNA, and KINARI could be further improved by considering the structural context (e.g., secondary structure, cooperativity between noncovalent interactions, and/or surface accessibility) when evaluating hydrogen bonds and hydrophobic interactions. From an application point of view, parameterizations that reflect different molecular environments will be helpful to evaluate structural stability in different solvents or of membrane-associated and transmembrane systems. Finally, current application studies predominantly focused on investigating a small number of systems, and almost all studies were performed in a retrospective manner. However, both large-scale and prospective studies are required to further evaluate the scope and limitations of rigidity analyses on biomolecules, as pursued in other areas of computational biophysics and structural biology.[172,212,213]

ACKNOWLEDGMENTS

We are grateful to Mike Thorpe, Leslie Kuhn, Donald Jacobs, Ileana Streinu, and Meera Sitharam for many stimulating discussions on the topic of rigidity theory and its application to biomolecules. We thank previous members of our lab (Sebastian Radestock, Elena Schmidt, Simone Fulle, Doris Klein, and Prakash Rathi) for their valuable contributions to this field.

REFERENCES

1. Taverna DM, Goldstein RA. Why are proteins marginally stable? *Proteins* 2002, 46:105–109.
2. Ahmed A, Kazemi S, Gohlke H. Protein flexibility and mobility in structure-based drug design. *Front Drug Des Discov* 2007, 3:455–476.
3. Sterner R, Brunner E. The relationship between catalytic activity, structural flexibility and conformational stability as deduced from the analysis of mesophilic-thermophilic enzyme pairs and protein engineering studies. In: *Thermophiles: Biology and Technology at High Temperatures*. London/New York: CRC Press; 2008, 25–38.
4. Luque I, Freire E. Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins* 2000, 41:63–71.
5. Paloncýová M, Navrátilová V, Berka K, Laio A, Otyepka M. Role of enzyme flexibility in ligand access and egress to active site: bias-exchange metadynamics study of 1,3,7-trimethyluric acid in cytochrome P450 3A4. *J Chem Theory Comput* 2016, 12:2101–2109.
6. Teague SJ. Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov* 2003, 2:527–541.
7. Daniel RM, Dunn RV, Finney JL, Smith JC. The role of dynamics in enzyme activity. *Annu Rev Biophys Biomol Struct* 2003, 32:69–92.
8. Manglik A, Kobilka B. The role of protein dynamics in GPCR function: insights from the β2AR and rhodopsin. *Curr Opin Cell Biol* 2014, 27:136–143.
9. Zavodszky P, Kardos J, Svingor A, Petsko GA. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc Natl Acad Sci USA* 1998, 95:7406–7411.
10. Rathi PC, Pfleger C, Fulle S, Klein DL, Gohlke H. Statics of biomacromolecules. In: *Modeling of Molecular Properties*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA; 2011, 281–299.
11. Vihinen M. Relationship of protein flexibility to thermostability. *Protein Eng* 1987, 1:477–480.
12. Carlson HA. Protein flexibility and drug design: how to hit a moving target. *Curr Opin Chem Biol* 2002, 6:447–452.
13. Jagodzinski F, Hardy J, Streinu I. Using rigidity analysis to probe mutation-induced structural changes in proteins. *J Bioinform Comput Biol* 2011, 10:432–437.
14. Radestock S, Gohlke H. Protein rigidity and thermophilic adaptation. *Proteins* 2011, 79:1089–1108.
15. Tan H, Rader AJ. Identification of putative, stable binding regions through flexibility analysis of HIV-1 gp120. *Proteins* 2009, 74:881–894.
16. Zhang X, Wozniak J, Matthews B. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. *J Mol Biol* 1995, 250:527–552.
17. Ishima R, Torchia DA. Protein dynamics from NMR. *Nat Struct Biol* 2000, 7:740–743.
18. Weiss S. Fluorescence spectroscopy of single biomolecules. *Science* 1999, 283:1676–1683.
19. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. *Protein Sci* 2003, 12:1060–1072.
20. Palmer AG, Kroenke CD, Loria JP. Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods Enzymol* 2001, 339:204–238.
21. Tzeng S-R, Kalodimos CG. Dynamic activation of an allosteric regulatory protein. *Nature* 2009, 462:368–372.
22. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. *Proteins* 1994, 19:141–149.
23. Ikai A. Local rigidity of a protein molecule. *Biophys Chem* 2005, 116:187–191.
24. Karplus M, Kuriyan J. Molecular dynamics and protein function. *Proc Natl Acad Sci USA* 2005, 102:6679–6685.
25. Tozzini V. Coarse-grained models for proteins. *Curr Opin Struct Biol* 2005, 15:144–150.
26. Case DA. Normal mode analysis of protein dynamics. *Curr Opin Struct Biol* 1994, 4:285–290.
27. Hinsen K. Analysis of domain motions by approximate normal mode calculations. *Proteins* 1998, 33:417–429.
28. Dodson G, Verma CS. Protein flexibility: its role in structure and mechanism revealed by molecular simulations. *Cell Mol Life Sci* 2006, 63:207–219.
29. Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G, Emerson A, Fanelli F, Gohlke H, Kuhn LA, Morris GM, et al. Target flexibility: an emerging consideration in drug discovery and design. *J Med Chem* 2008, 51:6237–6255.
30. Aftabuddin M, Kundu S. Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys J* 2007, 93:225–231.
31. Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. *Biophys J* 2004, 86:85–91.
32. Bagler G, Sinha S. Network properties of protein structures. *Physica A* 2005, 346:27–33.
33. Bode C, Kovacs IA, Szalay MS, Palotai R, Korcsmaros T, Csermely P. Network analysis of protein dynamics. *FEBS Lett* 2007, 581:2776–2782.
34. Brinda K, Vishveshwara S. A network representation of protein structures: implications for protein stability. *Biophys J* 2005, 89:4159–4170.

35. Dokholyan NV, Li L, Ding F, Shakhnovich EI. Topological determinants of protein folding. *Proc Natl Acad Sci USA* 2002, 99:8637–8641.
36. Greene LH, Higman VA. Uncovering network systems within protein structures. *J Mol Biol* 2003, 334:781–791.
37. Heringa J, Argos P, Egmond MR, De Vlieg J. Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters. *Protein Eng* 1995, 8:21–30.
38. Kannan N, Vishveshwara S. Identification of side-chain clusters in protein structures by a graph spectral method. *J Mol Biol* 1999, 292:441–464.
39. Krishnan A, Zbilut JP, Tomita M, Giuliani A. Proteins as networks: usefulness of graph theory in protein science. *Curr Protein Pept Sci* 2008, 9:28–38.
40. Kundu S. Amino acid network within protein. *Physica A* 2005, 346:104–109.
41. Vishveshwara S, Ghosh A, Hansia P. Intra and intermolecular communications through protein structure network. *Curr Protein Pept Sci* 2009, 10:146–160.
42. Ghosh A, Sakaguchi R, Liu C, Vishveshwara S, Hou YM. Allosteric communication in cysteinyl tRNA synthetase: a network of direct and indirect readout. *J Biol Chem* 2011, 286:37721–37731.
43. Sethi A, Eargle J, Black AA, Luthey-Schulten Z. Dynamical networks in tRNA:protein complexes. *Proc Natl Acad Sci USA* 2009, 106:6620–6625.
44. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 2005, 435:814–818.
45. Daily MD, Gray JJ. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput Biol* 2009, 5:1–14.
46. Hendrickson B. Conditions for unique graph realizations. *SIAM J Comput* 1992, 21:65–84.
47. Jacobs DJ, Thorpe MF. Generic rigidity percolation: the pebble game. *Phys Rev Lett* 1995, 75:4051–4054.
48. Feng S, Sen PN. Percolation on elastic networks: new exponent and threshold. *Phys Rev Lett* 1984, 52:216–219.
49. Guyon E, Roux S, Hansen A, Bideau D, Troadec JP, Crapo H. Non-local and non-linear problems in the mechanics of disordered systems: application to granular media and rigidity problems. *Rep Prog Phys* 1990, 53:373–419.
50. Jacobs DJ, Hendrickson B. An algorithm for two-dimensional rigidity percolation: the pebble game. *J Comput Phys* 1997, 137:346–365.
51. Jacobs DJ, Thorpe MF. Generic rigidity percolation in two dimensions. *Phys Rev E* 1996, 53:3682–3693.
52. Moukarzel CF, Duxbury P. Comparison of connectivity and rigidity percolation. In: *Rigidity Theory and Applications*. New York: Kluwer Academic/Plenum Publishers; 1999, 69–79.
53. Moukarzel CF, Duxbury PM. Comparison of rigidity and connectivity percolation in two dimensions. *Phys Rev E* 1999, 59:2614–2622.
54. Maxwell JC. On the calculation of the equilibrium and stiffness of frames. *Philos Mag Ser 4* 1864, 27:294–299.
55. Laman G. On graphs and rigidity of plane skeletal structures. *J Eng Math* 1970, 4:331–340.
56. Thorpe MF, Jacobs DJ, Chubynsky NV, Rader AJ. Generic rigidity of network glasses. In: *Rigidity Theory and Applications*. New York: Kluwer Academic/Plenum Publishers; 1999, 239–277.
57. Thorpe M, Jacobs D, Chubynsky M, Phillips J. Self-organization in network glasses. *J Non-Cryst Solids* 2000, 266–269:859–866.
58. Sartbaeva A, Wells SA, Treacy MMJ, Thorpe MF. The flexibility window in zeolites. *Nat Mater* 2006, 5:962–965.
59. Jacobs DJ, Thorpe MF. Computer-implemented system for analyzing rigidity of substructures within a macromolecule, patent US 6 014 449, 1998.
60. Jacobs DJ, Rader A, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. *Proteins* 2001, 44:150–165.
61. Hespenheide BM, Jacobs DJ, Thorpe MF. Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J Phys Condens Matter* 2004, 16: S5055–S5064.
62. Fulle S, Gohlke H. Analyzing the flexibility of RNA structures by constraint counting. *Biophys J* 2008, 94:4202–4219.
63. Fulle S, Gohlke H. Constraint counting on RNA structures: Linking flexibility and function. *Methods* 2009, 49:181–188.
64. Fulle S, Gohlke H. Statics of the ribosomal exit tunnel: implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. *J Mol Biol* 2009, 387:502–517.
65. Mottonen JM, Jacobs DJ, Livesay DR. Allosteric response is both conserved and variable across three CheY orthologs. *Biophys J* 2010, 99:2245–2254.
66. Jacobs DJ, Dallakyan S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys J* 2005, 88:903–915.
67. Del Carpio CA, Florea MI, Suzuki A, Tsuboi H, Hatakeyma N, Endou A, Takaba H, Ichiishi E, Miyamoto A. A graph theoretical approach for assessing bio-macromolecular complex structural stability. *J Mol Model* 2009, 15:1349–1370.
68. Gohlke H, Kuhn LA, Case DA. Change in protein flexibility upon complex formation: analysis

of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* 2004, 56:322–337.

69. Hespenheide BM, Rader AJ, Thorpe MF, Kuhn LA. Identifying protein folding cores from the evolution of flexible regions during unfolding. *J Mol Graph Model* 2002, 21:195–207.
70. Rader AJ, Bahar I. Folding core predictions from network models of proteins. *Polymer* 2004, 45:659–668.
71. Ahmed A, Gohlke H. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins* 2006, 63:1038–1051.
72. Fulle S, Christ NA, Kestner E, Gohlke H. HIV-1 TAR RNA spontaneously undergoes relevant apo-to-holo conformational transitions in molecular dynamics and constrained geometrical simulations. *J Chem Inf Model* 2010, 50:1489–1501.
73. Wells S, Menor S, Hespenheide B, Thorpe MF. Constrained geometric simulation of diffusive motion in proteins. *Phys Biol* 2005, 2:S127–S136.
74. Farrell DW, Speranskiy K, Thorpe MF. Generating stereochemically acceptable protein pathways. *Proteins* 2010, 78:2908–2921.
75. Radestock S, Gohlke H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng Life Sci* 2008, 8:507–522.
76. Livesay DR, Jacobs DJ. Conserved quantitative stability / flexibility relationships (QSFR) in an orthologous RNAse H pair. *Proteins* 2006, 62:130–143.
77. Jacobs DJ. Generic rigidity in three-dimensional bond-bending networks. *J Phys A Math Gen* 1998, 31:6653–6668.
78. Whiteley W. Counting out to the flexibility of molecules. *Phys Biol* 2005, 2:S116–S126.
79. Fox N, Jagodzinski F, Li Y, Streinu I. KINARI-Web: a server for protein rigidity analysis. *Nucleic Acids Res* 2011, 39:177–183.
80. Whiteley W. Rigidity of molecular structures: generic and geometric analysis. In: *Rigidity Theory and Applications*. New York: Kluwer Academic/Plenum Publishers; 1999, 21–46.
81. Tay T, Whiteley W. Recent advances in the generic ridigity of structures. *Struct Topol* 1984, 9:31–38.
82. Thorpe MF, Lei M, Rader AJ, Jacobs DJ, Kuhn LA. Protein flexibility and dynamics using constraint theory. *J Mol Graph Model* 2001, 19:60–69.
83. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF. Protein unfolding: rigidity lost. *Proc Natl Acad Sci USA* 2002, 99:3540–3545.
84. Pfleger C, Rathi PC, Klein DL, Radestock S, Gohlke H. Constraint network analysis (CNA): A python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. *J Chem Inf Model* 2013, 53:1007–1015.
85. Fox N, Streinu I. Towards accurate modeling of noncovalent interactions for protein rigidity analysis. *BMC Bioinformatics* 2013, 14:1–22.
86. Streinu I. Large scale rigidity-based flexibility analysis of biomolecules. *Struct Dyn* 2016, 3:1–16.
87. Wells SA, Jimenez-Roldan JE, Römer RA. Comparative analysis of rigidity across protein families. *Phys Biol* 2009, 6:1–11.
88. Dahiyat BI, Gordon DB, Mayo SL. Automated design of the surface positions of protein helices. *Protein Sci* 1997, 6:1333–1337.
89. Cheatham TE, Cieplak P, Kollman PA. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J Biomol Struct Dyn* 1999, 16:845–862.
90. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 1995, 117:5179–5197.
91. Mamonova T, Hespenheide B, Straub R, Thorpe MF, Kurnikova M. Protein flexibility using constraints from molecular dynamics simulations. *Phys Biol* 2005, 2:S137–S147.
92. Pfleger C, Radestock S, Schmidt E, Gohlke H. Global and local indices for characterizing biomolecular flexibility and rigidity. *J Comput Chem* 2013, 34:220–233.
93. Van Wynsberghe AW, Cui Q. Comparison of mode analyses at different resolutions applied to nucleic acid systems. *Biophys J* 2005, 89:2939–2949.
94. Wang Y, Rader AJ, Bahar I, Jernigan RL. Global ribosome motions revealed with elastic network model. *J Struct Biol* 2004, 147:302–314.
95. Sljoka A. Counting for rigidity, flexibility and extensions via the pebble game algorithm. *York Univ*. Thesis, 2006:1–173.
96. Lee A, Streinu I. Pebble game algorithms and sparse graphs. *Discret Math* 2008, 308:1425–1437.
97. Lee A, Streinu I, Theran L. Graded sparse graphs and matroids. *J Univ Comput Sci* 2007, 13:1671–1679.
98. Katoh N, Tanigawa SI. A proof of the molecular conjecture. *Discrete Comput Geom* 2011, 45:647–700.
99. Rader AJ. Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys Biol* 2010, 7:1–11.
100. Privalov PL, Gill SJ. Stability of protein structrure and hydrophobic interaction. *Adv Protein Chem* 1988, 39:191–234.
101. Schellman JA. Temperature, stability, and the hydrophobic interaction. *Biophys J* 1997, 73:2960–2964.
102. Rathi PC, Radestock S, Gohlke H. Thermostabilizing mutations preferentially occur at structural weak

spots with a high mutation ratio. *J Biotechnol* 2012, 159:135–144.

103. Stauffer D. Scaling theory of percolation clusters. *Phys Rep* 1979, 54:1–74.
104. Stauffer D, Aharony A. *Introduction to Percolation Theory*. 2nd ed. London: Taylor and Francis; 1994, 1–194.
105. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* 1948, 27:623–656.
106. Andraud C, Beghdadi A, Lafait J. Entropic analysis of random morphologies. *Physica A* 1994, 207:208–212.
107. Pfleger C, Gohlke H. Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. *Structure* 2013, 21:1725–1734.
108. Rathi PC, Jaeger KE, Gohlke H. Structural rigidity and protein thermostability in variants of lipase A from *Bacillus subtilis*. *PLoS One* 2015, 10:1–24.
109. Zwanzig RW. High-temperature equation of state by a perturbation method. I. nonpolar gases. *J Chem Phys* 1954, 22:1420–1426.
110. Jorgensen W, Ravimohan C. Monte Carlo simulation of differences in free energies of hydration. *J Chem Phys* 1985, 83:3050–3054.
111. Livesay DR, Dallakyan S, Wood GG, Jacobs DJ. A flexible approach for understanding protein stability. *FEBS Lett* 2004, 576:468–476.
112. Jacobs DJ, Dallakyan S, Wood GG, Heckathorne A. Network rigidity at finite temperature: Relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, 68:1–51.
113. Livesay DR, Huynh DH, Dallakyan S, Jacobs DJ. Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family. *Chem Cent J* 2008, 2:17.
114. Jacobs DJ, Livesay DR, Hules J, Tasayco ML. Elucidating quantitative stability/flexibility relationships within thioredoxin and its fragments using a distance constraint model. *J Mol Biol* 2006, 358:882–904.
115. Mottonen JM, Xu M, Jacobs DJ, Livesay DR. Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family. *Proteins* 2009, 75:610–627.
116. Verma D, Jacobs DJ, Livesay DR. Predicting the melting point of human C-type lysozyme mutants. *Curr Protein Pept Sci* 2010, 11:562–572.
117. Verma D, Jacobs DJ, Livesay DR. Changes in lysozyme flexibility upon mutation are frequent, large and long-ranged. *PLoS Comput Biol* 2012, 8: e1002409.
118. Li T, Verma D, Tracka MB, Casas-Finet J, Livesay DR, Jacobs DJ. Thermodynamic stability and flexibility characteristics of antibody fragment complexes. *Protein Pept Lett* 2014, 21:752–765.
119. Makhatadze GI, Privalov PL. On the entropy of protein folding. *Protein Sci* 1996, 5:507–510.
120. Gohlke H, Case DA. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J Comput Chem* 2004, 25:238–250.
121. Sljoka A, Wilson D. Probing protein ensemble rigidity and hydrogen-deuterium exchange. *Phys Biol* 2013, 10:56013.
122. Rathi PC, Mulnaes D, Gohlke H. VisualCNA: a GUI for interactive constraint network analysis and protein engineering for improving thermostability. *Bioinformatics* 2015, 31:2394–2396.
123. Krüger DM, Rathi PC, Pfleger C, Gohlke H. CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. *Nucleic Acids Res* 2013, 41: W340–W348.
124. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. *J Mol Biol* 1994, 238:777–793.
125. Zaccai G. How soft is a protein? A protein dynamics force constant measured by neutron scattering. *Science* 2000, 288:1604–1607.
126. Crivelli S, Eskow E, Bader B, Lamberti V, Byrd R, Schnabel R, Head-Gordon T. A physical approach to protein structure prediction. *Biophys J* 2002, 82:36–49.
127. Forli S, Olson AJ. A force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *J Med Chem* 2012, 55:623–638.
128. Huey R, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* 2007, 28:1145–1152.
129. Rathi PC, Fulton A, Jaeger KE, Gohlke H. Application of rigidity theory to the thermostabilization of lipase A from *Bacillus subtilis*. *PLoS Comput Biol* 2016, 12:e1004754.
130. González LC, Wang H, Livesay DR, Jacobs DJ. Calculating ensemble averaged descriptions of protein rigidity without sampling. *PLoS One* 2012, 7:1–13.
131. González LC, Livesay DR, Jacobs DJ. Improving protein flexibility predictions by combining statistical sampling with a mean-field virtual Pebble Game. *ACM-BCB* 2011:294–298. doi: 10.1145/2147805.2147838.
132. Vriend G. What if: a molecular modeling and drug design program. *J Mol Graph* 1990, 8:52–56.
133. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol* 1999, 285:1711–1733.

134. Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* 2005, 33:368–371.

135. Metz A, Pfleger C, Kopitz H, Pfeiffer-Marek S, Baringhaus KH, Gohlke H. Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein-protein interface. *J Chem Inf Model* 2012, 52:120–133.

136. Raschka S, Bemister-Buffington J, Kuhn LA. Detecting the native ligand orientation by interfacial rigidity: SiteInterlock. *Proteins* 2016, 84:1888–1901.

137. Verma D, Jacobs DJ, Livesay DR. Variations within class-A β-lactamase physiochemical properties reflect evolutionary and environmental patterns, but not antibiotic specificity. *PLoS Comput Biol* 2013, 9:1–16.

138. Brown MC, Verma D, Russell C, Jacobs DJ, Livesay DR. A case study comparing quantitative stability/flexibility relationships across five metallo-β-lactamases highlighting differences within NDM-1. *Methods Mol Biol* 2014, 1084:227–238.

139. Brown JR, Livesay DR. Flexibility correlation between active site regions is conserved across four AmpC β-lactamase enzymes. *PLoS One* 2015, 10:1–19.

140. Li T, Tracka MB, Uddin S, Casas-Finet J, Jacobs DJ, Livesay DR. Redistribution of flexibility in stabilizing antibody fragment mutants follows Le Châtelier's principle. *PLoS One* 2014, 9:1–14.

141. Marcia M, Humphris-Narayanan E, Keating KS, Somarowthu S, Rajashankar K, Pyle AM. Solving nucleic acid structures by molecular replacement: examples from group II intron studies. *Acta Crystallogr D Biol Crystallogr* 2013, 69:2174–2185.

142. Stoddard CD, Montange RK, Hennelly SP, Rambo RP, Sanbonmatsu KY, Batey RT. Free state conformational sampling of the SAM-I riboswitch aptamer domain. *Structure* 2010, 18:787–797.

143. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. *Nature* 2007, 450:964–972.

144. Arkhipov A, Yin Y, Schulten K. Four-scale description of membrane sculpting by BAR domains. *Biophys J* 2008, 95:2806–2821.

145. Arkhipov A, Freddolino PL, Schulten K. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure* 2006, 14:1767–1777.

146. Gohlke H, Thorpe MF. A natural coarse graining for simulating large biomolecular motion. *Biophys J* 2006, 91:2115–2120.

147. Lei M, Zavodszky MI, Kuhn LA, Thorpe MF. Sampling protein conformations and pathways. *J Comput Chem* 2004, 25:1133–1148.

148. Belfield WJ, Cole DJ, Martin IL, Payne MC, Chau PL. Constrained geometric simulation of the nicotinic acetylcholine receptor. *J Mol Graph Model* 2014, 52:1–10.

149. Kozuska JL, Paulsen IM, Belfield WJ, Martin IL, Cole DJ, Holt A, Dunn SMJ. Impact of intracellular domain flexibility upon properties of activated human 5-HT3 receptors. *Br J Pharmacol* 2014, 171:1617–1628.

150. Fokas AS, Cole DJ, Chin AW. Constrained geometric dynamics of the Fenna-Matthews-Olson complex: the role of correlated motion in reducing uncertainty in excitation energy transfer. *Photosynth Res* 2014, 122:275–292.

151. Sun M, Rose MB, Ananthanarayanan SK, Jacobs DJ, Yengo CM. Characterization of the pre-force-generation state in the actomyosin cross-bridge cycle. *Proc Natl Acad Sci USA* 2008, 105:8631–8636.

152. Jolley CC, Wells SA, Hespenheide BM, Thorpe MF, Fromme P. Docking of photosystem I subunit C using a constrained geometric simulation. *J Am Chem Soc* 2006, 128:8803–8812.

153. Jolley CC, Wells SA, Fromme P, Thorpe MF. Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. *Biophys J* 2008, 94:1613–1621.

154. Glembo TJ, Ozkan SB. Union of geometric constraint-based simulations with molecular dynamics for protein structure prediction. *Biophys J* 2010, 98:1046–1054.

155. de Groot BL, van Aalten DMF, Scheek RM, Amadei A, Vriend G, Berendsen HJC. Prediction of protein conformational freedom from distance constrains. *Proteins* 1997, 29:240–251.

156. Seeliger D, Haas J, de Groot BL. Geometry-based sampling of conformational transitions in proteins. *Structure* 2007, 15:1482–1492.

157. Hayward S, Kitao A, Berendsen HJC. Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins* 1997, 27:425–437.

158. Go N, Noguti T, Nishikawa T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc Natl Acad Sci USA* 1983, 80:3696–3700.

159. Brooks B, Karplus M. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci USA* 1983, 80:6571–6575.

160. He J, Zhang Z, Shi Y, Liu H. Efficiently explore the energy landscape of proteins in molecular dynamics simulations by amplifying collective motions. *J Chem Phys* 2003, 119:4005–4017.

161. Tatsumi R, Fukunishi Y, Nakamura H. A hybrid method of molecular dynamics and harmonic

dynamics for docking of flexible ligand to flexible receptor. *J Comput Chem* 2004, 25:1995–2005.

162. Zhang Z, Shi Y, Liu H. Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophys J* 2003, 84:3583–3593.
163. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* 1996, 77:1905–1908.
164. Durand P, Trinquier G, Sanejouand YH. A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers* 1994, 34:759–771.
165. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des* 1997, 2:173–181.
166. Tama F, Gadea FX, Marques O, Sanejouand Y-H. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* 2000, 41:1–7.
167. Kurkcuoglu O, Jernigan RL, Doruker P. Mixed levels of coarse-graining of large proteins using elastic network model succeeds in extracting the slowest motions. *Polymer* 2004, 45:649–657.
168. Doruker P, Jernigan RL, Bahar I. Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J Comput Chem* 2002, 23:119–127.
169. Li G, Cui Q. A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca^{2+}-ATPase. *Biophys J* 2002, 83:2457–2474.
170. Bahar I, Erman B, Haliloglu T, Jernigan RL. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. *Biochemistry* 1997, 36:13512–13523.
171. Ahmed A, Rippmann F, Barnickel G, Gohlke H. A normal mode-based geometric simulation approach for exploring biologically relevant conformational transitions in proteins. *J Chem Inf Model* 2011, 51:1604–1622.
172. Ahmed A, Villinger S, Gohlke H. Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. *Proteins* 2010, 78:3341–3352.
173. Minges ARM, Ciupka D, Winkler C, Höppner A, Gohlke H, Groth G. Structural intermediates and directionality of the swiveling motion of Pyruvate Phosphate Dikinase. *Sci Rep* 2017, 7:45389.
174. Dimura M, Peulen TO, Hanke CA, Prakash A, Gohlke H, Seidel CAM. Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. *Curr Opin Struct Biol* 2016, 40:163–185.
175. Krüger DM, Ahmed A, Gohlke H. NMSim web server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins. *Nucleic Acids Res* 2012, 40:310–316.
176. Jimenez-Roldan JE, Freedman RB, Römer RA, Wells SA. Rapid simulation of protein motion: merging flexibility, rigidity and normal mode analyses. *Phys Biol* 2012, 9:16008.
177. Burkoff NS, Várnai C, Wells SA, Wild DL. Exploring the energy landscapes of protein folding simulations with Bayesian computation. *Biophys J* 2012, 102:878–886.
178. Amin NT, Wallis AK, Wells SA, Rowe ML, Williamson RA, Howard MJ, Freedman RB. High-resolution NMR studies of structure and dynamics of human ERp27 indicate extensive interdomain flexibility. *Biochem J* 2013, 450:321–332.
179. Wells SA, Crennell SJ, Danson MJ. Structures of mesophilic and extremophilic citrate synthases reveal rigidity and flexibility for function. *Proteins* 2014, 82:2657–2670.
180. Erskine PT, Fokas A, Muriithi C, Rehman H, Yates LA, Bowyer A, Findlow IS, Hagan R, Werner JM, Miles AJ, et al. X-ray, spectroscopic and normal-mode dynamics of calexcitin: structure-function studies of a neuronal calcium-signalling protein. *Acta Crystallogr D Biol Crystallogr* 2015, 71:615–631.
181. Kavraki LE, Svestka P, Latombe J-C, Overmars MH. Probabilistic roadmaps for path planning in high-dimensionalconfiguration spaces. *IEEE Trans Robot Autom* 1996, 12:566–580.
182. Sitharam M, Ozkan A, Pence J, Peters J. EASAL: efficient atlasing, analysis and search of molecular assembly landscapes. *arXiv:1203.3811*, 2012, 1–26.
183. Yao P, Zhang L, Latombe J-C. Sampling-based exploration of folded state of a protein under kinematic and geometric constraints. *Proteins* 2012, 80:25–43.
184. Pachov DV, Fonseca R, Arnol D, Bernauer J, van den Bedem H. Coupled motions in β2AR:Gαs conformational ensembles. *J Chem Theory Comput* 2016, 12:946–956.
185. Fonseca R, van den Bedem H, Bernauer J. Probing RNA native conformational ensembles with structural constraints. *J Comput Biol* 2016, 23:362–371.
186. Fonseca R, Pachov DV, Bernauer J, van den Bedem H. Characterizing RNA ensembles from NMR data with kinematic models. *Nucleic Acids Res* 2014, 42:9562–9572.
187. Bansod YD, Nandanwar D. Overview of tensegrity—I: basic structures. *Eng Mech* 2014, 21:355–367.
188. Wu R, Ozkan A, Bennett A, Agbandje-Mckenna M, Sitharam M. Robustness measure for an adeno-associated viral shell self-assembly is accurately predicted by configuration space atlasing using EASAL. *ACM-BCB* 2012:690–695. doi: :10.1145/2147805.2147838.

189. Ozkan A, Flores-Canales JC, Sitharam M, Kurnikova M. Fast and flexible geometric method for enhancing MC sampling of compact configurations for protein docking problem. *arXiv:1408.2481*, 2014, 1–29.

190. Rader AJ, Anderson G, Isin B, Khorana HG, Bahar I, Klein-Seetharaman J. Identification of core amino acids stabilizing rhodopsin. *Proc Natl Acad Sci USA* 2004, 101:7246–7251.

191. Dick M, Weiergräber OH, Classen T, Bisterfeld C, Bramski J, Gohlke H, Pietruszka J. Trading off stability against activity in extremophilic aldolases. *Sci Rep* 2016, 6:17908.

192. Gohlke H, Ben-Shalom IY, Kopitz H, Pfeiffer-Marek S, Baringhaus K-H. Rigidity theory-based approximation of vibrational entropy changes upon binding to biomolecules. *J Chem Theory Comput* 2017, 13:1495–1502.

193. van den Burg B. Extremophiles as a source for novel enzymes. *Curr Opin Microbiol* 2003, 6:213–218.

194. Vieille C, Zeikus GJ, Vieille C. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* 2001, 65:1–43.

195. Polizzi KM, Bommarius AS, Broering JM, Chaparro-Riggers JF. Stability of biocatalysts. *Curr Opin Chem Biol* 2007, 11:220–225.

196. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. *Structure* 2003, 11:1453–1459.

197. Ahmad S, Kamal MZ, Sankaranarayanan R, Rao NM. Thermostable Bacillus subtilis lipases: in vitro evolution and structural insight. *J Mol Biol* 2008, 381:324–340.

198. Ahmad S, Rao NM. Thermally denatured state determines refolding in lipase: mutational analysis. *Protein Sci* 2009, 18:1183–1196.

199. Kamal MZ, Ahmad S, Molugu TR, Vijayalakshmi A, Deshmukh MV, Sankaranarayanan R, Rao NM. In vitro evolved nonaggregating and thermostable lipase: structural and thermodynamic investigation. *J Mol Biol* 2011, 413:726–741.

200. Nussinov R, Tsai C-J, Ma B. The underappreciated role of allostery in the cellular network. *Annu Rev Biophys* 2013, 42:169–189.

201. Monod J, Wyman J, Changeux J-P. On the nature of allosteric transitions: a plausible model. *J Mol Biol* 1965, 12:88–118.

202. Koshland DE, Némethy G, Filmer D. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* 1966, 5:365–385.

203. Cooper A, Dryden D. Allostery without conformational change—a plausible model. *Eur Biophys J* 1984, 11:103–109.

204. Srivastava A, Tracka MB, Uddin S, Casas-finet J, Livesay DR, Jacobs DJ. Mutations in antibody fragments modulate allosteric response via hydrogen-bond network fluctuations. *Biophys J* 2016, 110:1933–1942.

205. Kallen J, Welzenbach K, Ramage P, Geyl D, Kriwacki R, Legge G, Cottens S, Weitz-Schmidt G, Hommel U. Structural basis for LFA-1 inhibition upon lovastatin binding to the CD11a I-domain. *J Mol Biol* 1999, 292:1–9.

206. Wiesmann C, Barr KJ, Kung J, Zhu J, Erlanson DA, Shen W, Fahr BJ, Zhong M, Taylor L, Randal M, et al. Allosteric inhibition of protein tyrosine phosphatase 1B. *Nat Struct Mol Biol* 2004, 11:730–737.

207. Guckian KM, Lin EYS, Silvian L, Friedman JE, Chin D, Scott DM. Design and synthesis of a series of meta aniline-based LFA-1 ICAM inhibitors. *Bioorg Med Chem Lett* 2008, 18:5249–5251.

208. Hintersteiner M, Kallen J, Schmied M, Graf C, Jung T, Mudd G, Shave S, Gstach H, Auer M. Identification and X-ray co-crystal structure of a small-molecule activator of LFA-1-ICAM-1 binding. *Angew Chem Int Ed Engl* 2014, 53:4322–4326.

209. Vázquez-Laslop N, Ramu H, Klepacki D, Kannan K, Mankin AS. The key function of a conserved and modified rRNA residue in the ribosomal response to the nascent peptide. *EMBO J* 2010, 29:3108–3117.

210. Seidelt B, Innis CA, Wilson DN, Gartmann M, Armache J-P, Villa E, Trabuco LG, Becker T, Mielke T, Schulten K, et al. Structural insight into nascent polypeptide chain-mediated translational stalling. *Science* 2009, 326:1412–1415.

211. Bowen JP, Allinger NL. Molecular mechanics: the art and science of parameterization. In: *Reviews in Computational Chemistry*, vol. 2. Hoboken, NJ: John Wiley & Sons; 1991, 81–97.

212. Mikulskis P, Genheden S, Ryde U. A large-scale test of free-energy simulation estimates of protein–ligand binding affinities. *J Chem Inf Model* 2014, 54:2794–2806.

213. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 2005, 15:285–289.

ORIGINAL PUBLICATION II

Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for *Bacillus subtilis* lipase A

Nutschel, C., Fulton, A., Zimmermann, O., Schwaneberg, U., Jaeger, K.-E., Gohlke, H.

J Chem Inf Model. 2020, 60, 3, 1568-1584.

https://pubs.acs.org/doi/10.1021/acs.jcim.9b00954

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

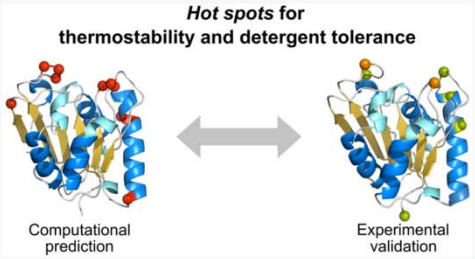
Systematically Scrutinizing the Impact of Substitution Sites on Thermostability and Detergent Tolerance for *Bacillus subtilis* Lipase A

Christina Nutschel, Alexander Fulton, Olav Zimmermann, Ulrich Schwaneberg, Karl-Erich Jaeger, and Holger Gohlke*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 1568−1584 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Improving an enzyme's (thermo-)stability or tolerance against solvents and detergents is highly relevant in protein engineering and biotechnology. Recent developments have tended toward data-driven approaches, where available knowledge about the protein is used to identify substitution sites with high potential to yield protein variants with improved stability, and subsequently, substitutions are engineered by site-directed or site-saturation (SSM) mutagenesis. However, the development and validation of algorithms for data-driven approaches have been hampered by the lack of availability of large-scale data measured in a uniform way and being unbiased with respect to substitution types and locations. Here, we extend our knowledge on guidelines for protein engineering following a data-driven approach by scrutinizing the impact of substitution sites on thermostability or/and detergent tolerance for *Bacillus subtilis* lipase A (*Bs*LipA) at very large scale. We systematically analyze a complete experimental SSM library of *Bs*LipA containing all 3439 possible single variants, which was evaluated as to thermostability and tolerances against four detergents under respectively uniform conditions. Our results provide systematic and unbiased reference data at unprecedented scale for a biotechnologically important protein, identify consistently defined hot spot types for evaluating the performance of data-driven protein-engineering approaches, and show that the rigidity theory and ensemble-based approach Constraint Network Analysis yields hot spot predictions with an up to ninefold gain in precision over random classification.



1. INTRODUCTION

Improving a protein's (thermo-)stability[1−8] or tolerance against solvents[9−16] and detergents[17−19] has become of utmost importance in protein engineering: Considering that enzymes are predominantly used as detergent additives[20] and that the global industrial enzyme market has been forecast to reach $7.0 billion by 2023 from $5.5 billion in 2018 makes it clear that an increasing demand exists for enzymes that are adapted to harsh temperature, solvent, and detergent conditions.[20−22]

Modifying protein stability based on rational approaches has a long history,[23,24] and a number of, usually, structure-based algorithms have been developed that estimate the effect of a substitution on the stability of a protein.[25−28] However, despite successful applications in single cases (e.g., see Table 2 in ref 20), the general reliability of these approaches is still unsatisfactory.[25,29−32] One reason is that multiple attempts to identify key features in protein sequences and/or structures associated with protein stability have failed to paint a clear picture, which makes it difficult to define rules of universal validity and general applicability.[20,33] Another reason lies in the data used in the design and evaluation of rational design algorithms. The ProTherm database,[34,35] which has been most frequently used for such endeavors, contains on average ∼12 single, ∼12 double, and ∼1 multiple substitutions for each of the ∼1000 proteins stored.[33] Thus, while overall exhaustive, the data may not include a sufficient number of variants per protein to compensate for outliers and, therefore, may not allow a stratification of the data to derive a generally applicable set of rules. As such data, furthermore, originate from different experimental methods, it is not surprising that different changes in protein stability have been found associated with the same variant.[36] In addition, the data are strongly biased toward substitutions to alanine, whereas it is very limited for some other substitutions.[37] Recently, comprehensive mutagenesis data on a domain level associated with protein stabilities against a denaturating agent have been reported as a means to overcome these limitations.[38]

Following the principles of natural evolution, albeit on a reduced time scale, protein engineering by directed evolution has emerged as an attractive strategy to improve stability

Received: October 12, 2019
Published: January 6, 2020

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

through iterative cycles of mutagenesis and screening or selection.[20,39] However, the highly labor-intensive method can become technically challenging if beneficial mutations need to be accumulated over generations of mutagenesis and screening or selection to reach a desired effect.[40] After all, evolution is not good for problems that require multiple, simultaneous, low-probability events.[41] To successfully investigate the then necessary large protein libraries, powerful automated techniques for rapid high-throughput screenings were established.[20,39]

As an intermediate, third route recent developments have tended toward data-driven approaches,[42] where available knowledge about the protein is used to first identify a substitution site with high potential to yield protein variants with improved stability, and second, substitutions are engineered by site-directed (SDM) or site-saturation (SSM) mutagenesis.[33] The "knowledge" can arise from sequence information,[42,43] structure information,[44−46] or computational techniques.[2,4,7,8,47,48] By such data-driven approaches, the challenge of accurately predicting the effect of a substitution on protein stability is circumvented, and substitution efforts are guided to a few, distinguished sequence positions, making subsequent combinations feasible. However, even with high-throughput screening techniques, it is difficult to handle all variants based on combinations of the 20 proteinogenic AAs at more than six substitution sites (i.e., more than $20^6 = 6.4 \times 10^7$ variants).[20,39,49,50]

Here, to extend our knowledge on guidelines for time- and cost-efficient protein engineering following a data-driven approach, we scrutinize the impact of substitution sites on thermostability or/and detergent tolerance for one protein at very large scale. To do so, we systematically analyze a complete experimental SSM library of *Bs*LipA produced by us,[15,16,19] which contains all 3439 theoretically possible single variants (181 substitution sites of *Bs*LipA × 19 naturally occurring AAs) and was evaluated as to different protein stabilities under respectively uniform conditions. Previously, the SSM library has been characterized regarding solvent and detergent tolerance (D) data.[15,16,19] Here, we characterize the SSM library for the first time regarding thermostability (T_{50}) as well as combined T_{50} and D data. *Bs*LipA is a particularly interesting protein for such analysis because a high-resolution X-ray crystal structure (PDB ID: 1ISP, 1.3 Å) is known,[51] which provides valuable insights in atomic details. Furthermore, the protein has considerable biotechnological importance,[52,53] possesses an α/β-hydrolase fold[54] such that the impact of substitution sites at α-helices, β-strands, and other secondary structure elements can be tested, and has been used frequently as a model system in related experimental and computational small-scale studies.[7,8]

Our systematic large-scale analysis focuses on the following five aspects: (I) We determined the likelihoods to find substitution sites showing significantly increased T_{50} or D and investigated the frequencies and magnitudes of effects caused by single AA substitutions. (II) We analyzed at which substitution sites variants result with increased T_{50} or/and D across the protein and compared the findings to random mutagenesis. (III) From these results, we defined *hot spot* classes, i.e., classes of substitution sites particularly promising to increase T_{50} or/and D. (IV) We probed to what extent hot spots can be predicted based on structure or sequence characteristics. (V) We tested the predictive power of the rigidity theory-based approach Constraint Network Analysis (CNA) previously applied in related scenarios,[2,4−8] i.e., how accurately hot spots can be predicted as structural *weak spots* identified in a thermal unfolding simulation of the protein.

The main outcomes from our analyses are that we provide systematic and unbiased reference data at large scale for thermostability measured as T_{50} values and detergent tolerance measured as D for a biotechnologically important protein, we identify and consistently define hot spot types for evaluating the performance of data-driven protein-engineering approaches, and we show that CNA-based hot spot prediction can yield a gain in precision over *random classification* up to ninefold.

2. MATERIALS AND METHODS

2.1. Generation and Screening of the *Bs*LipA SSM Library toward Changes in T_{50} or D. The *Bs*LipA library was constructed by site-saturation mutagenesis (SSM) and site-directed mutagenesis (SDM) as described by Frauenkron-Machedjou et al.[15,16] and Fulton et al.[19] In the present study, we defined all 3439 single variants (181 substitution sites of *Bs*LipA × 19 naturally occurring AAs) generated with SSM and SDM as the "SSM library".

Previously, the SSM library has been screened toward its tolerance against four different classes of detergents: anionic (sodium dodecyl sulfate, SDS), cationic (cetyltrimethylammonium bromide, CTAB), zwitterionic (3-[hexadecyl(dimethyl)-azaniumyl]propane-1-sulfonate, SB3-16), and nonionic (polyoxyethylenesorbitan monooleate, Tween 80) by Fulton et al.[19] Residual activities of the variants after incubation in the presence of the respective detergent (D) were obtained as described in ref 19.

As to the screening procedure regarding thermostability, the screening cultures were incubated as described in ref 19. The culture supernatant was collected by centrifugation (1500 g, 40 min) and diluted 2.5-fold with Sørensen buffer (42.5 mL of Na_2HPO_4 (8.9 g l^{-1}), 2.5 mL of KH_2PO_4 (6.8 g l^{-1})) before screening. The protein-containing supernatant was incubated in a 0.2 mL PCR microtiter plate (MTP) in a programmable thermal cycler (Eppendorf Mastercycler Thermal Cycler PCR). The supernatant samples were incubated at temperatures between 40 and 60 °C for 20 min. A dry block incubator (MRK 23 Cooling-ThermoMixer, DITABIS) was equipped with a "15 and 50 mL falcon tube adaptor" (BT 03, DITABIS). Three falcon tubes with 19.8 mL of *para*-nitrophenyl palmitate (*p*NPP) solution A (19.8 mL of Sørensen buffer, 45.54 mg of sodium deoxycholate, 22 mg of gum arabic) were inserted into the falcon tube incubator. All dry block incubators were set to 40 °C, 30 min prior to the beginning of the experiment. Twenty seconds before the end of the incubation, 2.2 mL of *p*NPP solution B (48 mg of *p*NPP in 8 mL of 2-propanol) was added into prewarmed *p*NPP solution A and briefly mixed. The substrate mixture was applied to the wells of the MTPs in 50 μL aliquots to start the measurement of thermostability and measured in a MTP reader (Molecular Devices Spectramax). The enzymatic activity in each sample was measured by the rate of increase in absorption at O.D. 410 nm. The residual activity in each sample was calculated from the slope of the change in absorption at O.D. 410 nm relative to the slope of the sample heated to 40 °C during a measurement time of 3 min. From that, T_{50} was obtained from the inflection point of a sigmoid curve fit. Control experiments with just *p*NPP, or *p*NPP in the presence of *Bs*LipA at temperatures up to 60.6 °C, that way leading to denaturation of *Bs*LipA, show no change in the *para*-nitrophenolate (*p*NP) absorption over time, demonstrating that *p*NP is only produced in the presence of a functional enzyme (Figure

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

S1). The T_{50} values are provided as an Excel sheet in the Supporting Information.

2.2. Global Characterization of *Bs*LipA Variants' Changes in T_{50} or D. For analyzing the changes in T_{50} (eq 1) or D (eq 2) of *Bs*LipA variants, the values of wt*Bs*LipA were used as references; i.e., the differences between the values of the variants and those of wt*Bs*LipA were calculated. Positive (negative) Δ-values indicate variants with increased (decreased) T_{50} or D.

$$\Delta T_{50} = T_{50}(\text{variant}) - T_{50}(\text{wt}Bs\text{LipA}) \quad (1)$$

$$\Delta D = D(\text{variant}) - D(\text{wt}Bs\text{LipA}) \quad (2)$$

For the large-scale analysis, only ΔT_{50} of variants higher (lower) than the experimental uncertainty, taken as the standard deviation σ_T for the respective variant determined from three screenings of T_{50}, were considered significantly increased (decreased) in T_{50} compared to wt*Bs*LipA. Furthermore, only ΔD of variants higher (lower) than two times the experimental standard deviation ($2\sigma_D$) of wt*Bs*LipA determined from screenings of 2997 wt*Bs*LipA replicates[19] toward the respective detergent were considered significantly increased (decreased) in D compared to wt*Bs*LipA. Here, σ_D of wt*Bs*LipA was used as significance criterion, as the experimental standard deviation for each variant was not available. $2\sigma_D$ was chosen because it corresponds to a p-value below 0.05.

2.3. Definitions of Classes of *Bs*LipA Substitution Sites. The different classes of substitution sites regarding significantly increased T_{50} or/and D were defined based on the set theory. Therefore, the following binary operations on sets were applied:

The *union* of the sets A and B is the set of elements which are in A, in B, or in both A and B (eq 3).[55]

$$(A \cup B) = \{x: x \in A \vee x \in B\} \quad (3)$$

The *intersection* of the sets A and B is the set of elements which are in A and B (eq 4).[55]

$$(A \cap B) = \{x: x \in A \wedge x \in B\} \quad (4)$$

Finally, the Jaccard index (J) was used to compare the similarity of two sets A and B, i.e., the cardinal number of the respective intersection divided by the cardinal number of the respective union (eq 5).[56,57] The range of J is [0, 1], with 1 indicating identical sets A and B.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

Based on the different classes of substitution sites, we defined *hot spots*, which are substitution sites particularly promising to yield significantly increased T_{50} or/and D.

2.4. Structural Determinants of *Bs*LipA Hot Spots. Hot spots were assigned to groups according to their location in secondary structure elements (yielding 20 subgroups), solvent-accessible surface areas (SASAs) (yielding five subgroups), and physicochemical properties (yielding five subgroups). The secondary structure elements of the wt*Bs*LipA crystal structure (PDB ID: 1ISP with highest resolution of 1.3 Å[51]) were identified with the DSSP program.[58] Additionally, the SASAs of the wt*Bs*LipA were analyzed with the DSSP program.[58] The fractional solvent-accessible surface areas (fSASAs) were calculated with respect to the maximum solvent-accessible surface area of each hot spot (maxSASA) (eq 6).[59]

$$\text{fSASA} = 100 \cdot \frac{\text{SASA}}{\text{maxSASA}} \quad (6)$$

As the screening studies were performed at pH 8,[19] hot spots were subgrouped by their physicochemical properties as follows: aliphatic (Ile, Ala, Val, Leu, Gly), aromatic (Phe, Tyr, Trp), neutral (Cys, Pro, Met, Ser, Thr, Asn, Gln), positively charged (His, Lys, Arg), and negatively charged (Asp, Glu).

2.5. Conservation of wt*Bs*LipA Residues within Bacterial Lipases. Apart from the catalytic triad (S77, D133, and H156), also variants at conserved sequence positions were considered because the SSM library revealed significantly increased T_{50} or/and D at such positions. The conservation of wt*Bs*LipA residues within the bacterial lipases was calculated using the available sequences from the Pfam database[60] for the lipase class 2 (PF01674). The sequences were limited to the bacterial sources, which contain 1138 sequences from 603 bacterial species. All sequences were aligned using Clustal Omega.[61,62] For the alignment, the full-length sequence of wt*Bs*LipA (UniProt ID: P37957) was used.[63] The conservation was calculated using AACon Calculations[64] through Jalview.[65] The conservation range is [0, 10] with 0 (10) showing no (high) conservation.

2.6. Constraint Network Analysis. The Constraint Network Analysis (CNA) aims at linking structural rigidity and flexibility to the biomolecule's structure, (thermo)stability, and function.[66−68] The CNA software acts as front- and back-end to the graph theory-based rigidity analysis software Floppy Inclusions and Rigid Substructure Topography (FIRST).[69] In CNA, proteins are modeled as constraint networks in a *body-and-bar* representation, which has been described in detail by Hesphenheide et al.[70] Based on the modeled constraint network of the protein structure, a *pebble game algorithm* decomposes the network into flexible and rigid subparts.[71,72] In order to monitor the decay of network rigidity and to identify the *rigidity percolation threshold*, CNA performs thermal unfolding simulations by consecutively removing noncovalent constraints (hydrogen bonds, including salt bridges) from a network in increasing order of their strength.[73] For this, a hydrogen bond energy E_{HB} is computed by a modified version of the potential by Mayo et al.[73] During the thermal unfolding simulations, phase transitions can be identified where the network switches from overall rigid to flexible states. For a given network state $\sigma = f(T)$, hydrogen bonds with an energy $E_{HB} > E_{cut}(\sigma)$ are removed from the network at temperature T. In this study, the thermal unfolding simulation was carried out by decreasing E_{cut} from −0.1 to −6.0 kcal mol^{-1} with a step size of 0.1 kcal mol^{-1}. E_{cut} can be converted to a temperature T using the linear equation introduced by Radestock et al. (eq 7).[2,4] The range of E_{cut} is equivalent to increasing the temperature from 302 to 420 K with a step size of 2 K. Because hydrophobic interactions remain constant or become even stronger as the temperature increases,[74,75] the number of hydrophobic tethers was kept unchanged during the thermal unfolding simulation, as done previously.[7,8]

$$T = \frac{-20\ \text{K}}{\text{kcal}\cdot\text{mol}^{-1}} E_{cut} + 300\ \text{K} \quad (7)$$

The CNA software is available under academic licenses from http://cpclab.uni-duesseldorf.de/index.php/Software, and the CNA web server is accessible at http://cpclab.uni-duesseldorf.de/cna/.

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

2.7. Generation of a Structural Ensemble of wt*Bs*LipA. MD simulations of wt*Bs*LipA were carried out with the GPU-accelerated version of PMEMD[76] of the AMBER14 suite of programs[77] together with the ff14SB force field.[78] As a starting structure, the X-ray crystal structure of wt*Bs*LipA (PDB ID: 1ISP) was used.[51] Hydrogens were added, and side-chain orientations ("flips") of Asn, Gln, and His were optimized by the REDUCE program[79] based on suitable hydrogen-bonding geometries and avoiding potential steric clashes. This was done to take into account that O versus N or N versus C is difficult to distinguish in X-ray crystallography experiments.[79] For neutralization of the system, sodium counterions were added. Subsequently, the system was solvated by a truncated octahedral box of TIP3P water[80] such that a layer of water molecules of at least 11 Å widths covers the protein surface. The particle mesh Ewald method[81] was used with a direct-space nonbonded cutoff of 8 Å. Bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm,[82] and the time step for the simulation was 2 fs. As done before,[8] a trajectory of 100 ns length was generated after thermalization and adjustment of the pressure, simulating in the canonical (NVT) ensemble at T = 300 K, with conformations extracted every 40 ps from the last 80 ns, resulting in a structural ensemble of 2000 conformations. We assessed the statistical independence of the extracted conformations by calculating the autocorrelation function of the cluster configuration entropy H_{type2}, the measure used to identify phase transitions in the constraint networks (see section 2.9 below) (Figure S2). Because fluctuations of H_{type2} decorrelate already within the first two snapshots, the snapshots used for CNA, which were extracted at time intervals of 40 ps, are considered independent.

2.8. Thermal Unfolding Simulation of wt*Bs*LipA. For analyzing the rigid cluster decomposition of wt*Bs*LipA, a thermal unfolding simulation was performed by CNA on an ensemble of network topologies (ENT^{MD}) generated from a molecular dynamics (MD) trajectory. The ensemble-based CNA was pursued to increase the robustness of the rigidity analyses.[5,83]Subsequently, the unfolding trajectory was visually inspected by VisualCNA[84] for identifying secondary structure elements that segregate from the largest rigid cluster at each major phase transition. VisualCNA is an easy-to-use PyMOL plug-in that allows setting up CNA runs and analyzing CNA results linking data plots with molecular graphics representations.[84] VisualCNA is available under an academic license from https://cpclab.uni-duesseldorf.de/index.php/Software.

2.9. Local and Global Indices for Analyzing Structural Rigidity of wt*Bs*LipA. From the thermal unfolding simulation, CNA computes a comprehensive set of indices to quantify biologically relevant characteristics of the biomolecule's stability.[85] *Global* indices are used for determining the flexibility and rigidity at a macroscopic level. *Local* indices determine the flexibility and rigidity at a microscopic level of bonds.

The cluster configuration entropy H_{type2} is a global index, which has been introduced by Radestock and Gohlke.[2] H_{type2} is used to identify the phase transition temperature T_p at which a biomolecule switches from a rigid to a floppy state and the largest rigid cluster stops to dominate the whole protein network. As long as the largest rigid cluster dominates the whole protein network, H_{type2} is low because of the limited number of possible ways to configure a system with a very large cluster. When the largest rigid cluster starts to decay or stops to dominate the network, H_{type2} jumps. There, the network is in a partially flexible state with many ways to configure a system consisting of many small clusters. The percolation behavior of protein networks is usually complex, and multiple phase transitions can be observed.[2,4,5,7,8] In order to identify T_p, a double sigmoid fit was applied to an H_{type2} versus $T(E_{cut})$ curve as done previously,[2,4,5,7,8] and T_p taken as that T value associated with the largest slope of the fit.

The stability map rc_{ij} is a local index, which has been introduced by Radestock and Gohlke.[4] rc_{ij} represents the local stability within a protein structure for all residue pairs at which a rigid contact rc between two residues i and j (represented by their C_α atoms) is lost during the thermal unfolding. rc exists if i and j belong to the same rigid cluster c of the set of rigid clusters $C^{E_{cut}}$.[85] Thus, rc_{ij} contains information cumulated over all network states along the unfolding trajectory as to which parts of the network are (locally) mechanically stable at a given σ and which are not.[7] This stability information is not only available in a qualitative manner but also quantitatively in that each rc_{ij} has been associated with E_{cut} at which the rigid contact is lost. The sum over all entries in rc_{ij} represents the chemical potential energy due to noncovalent bonding, obtained from the coarse-grained, residue-wise network representation of the underlying protein structure. To focus only on the stability of rc between structurally close residues, rc_{ij} was filtered such that only rigid contacts between two residues that are at most 5 Å apart from each other were considered (neighbor stability map $rc_{ij,neighbor}$).

Finally, CNA predicts *unfolding nuclei* as structural features from which macroscopic (in)stability originates.[2] Unfolding nuclei are represented by residues that percolate from the largest rigid cluster at the latest phase transition. If such residues become flexible, it will have a detrimental effect on protein stability. Fringe residues of the unfolding nuclei percolate from the largest rigid cluster during earlier steps of the thermal unfolding. We follow the hypothesis that the more structurally stable the fringes of unfolding nuclei are, the more structurally stable will be those unfolding nuclei.[2] Therefore, if such fringe residues (termed *weak spots*) are targeted by substitutions, the likelihood to stabilize the rigid core of a protein should be high. If two unfolding nuclei were only separated by one residue, this residue was also considered a weak spot. This procedure of identifying weak spots is in agreement with a previous study by us.[2]

2.10. Statistical Evaluation of CNA as a Binary Classifier. The performance of CNA was investigated as a binary classifier with the following possible outcomes: true positives (TP) are predicted weak spots that are hot spots, whereas false positives (FP) are predicted weak spots that are non-hot spots. In turn, true negatives (TN) are predicted non-weak spots that are non-hot spots, whereas false negatives (FN) are predicted non-weak spots that are hot spots. Different metrics were then applied to evaluate CNA.

The *recall* (r) answers the question how many hot spots were predicted as weak spots (eq 8).[86]

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\text{No. of predicted weak spots that are hot spots}}{\text{No. of hot spots}} \quad (8)$$

The *precision* (p) evaluates how many predicted weak spots are actually hot spots (eq 9).[86]

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

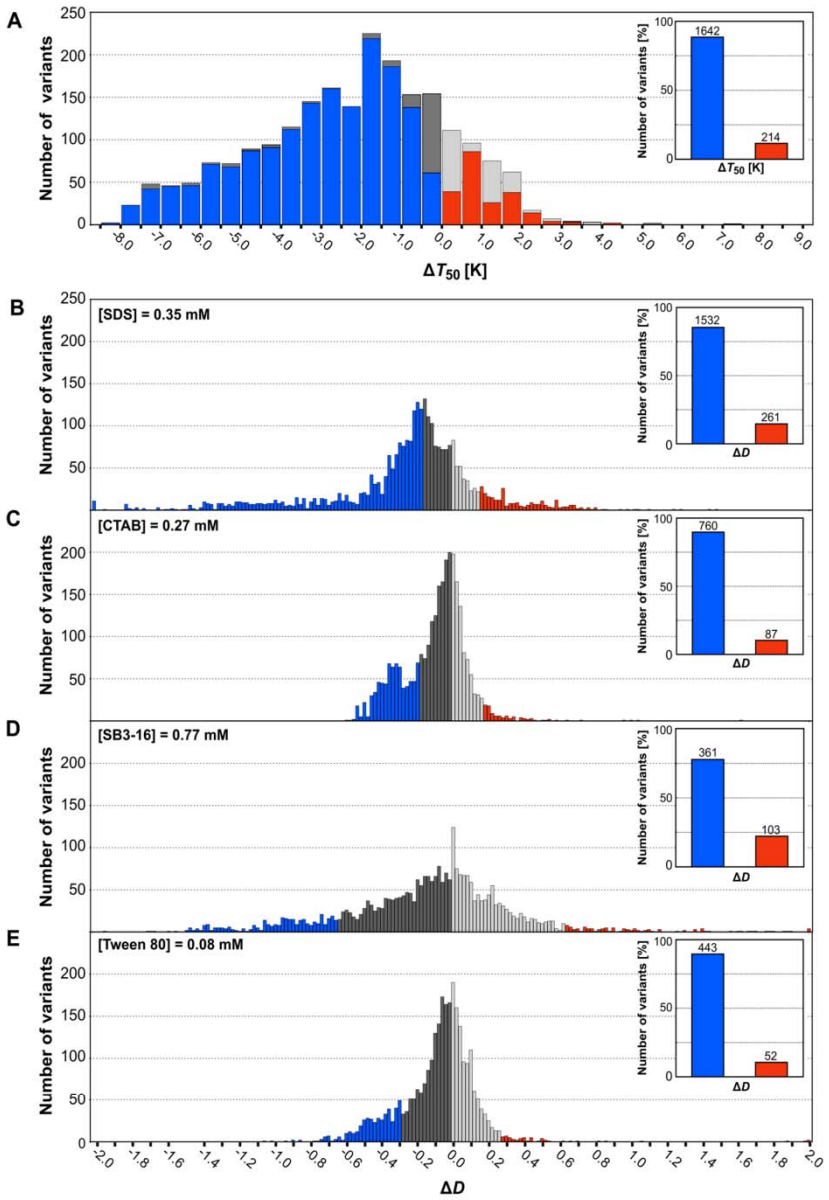


Figure 1. Distribution of *Bs*LipA variants' changes in T_{50} or D toward one detergent. Distribution of *Bs*LipA variants' changes in (A) T_{50} (ΔT_{50}) or D (ΔD) with respect to (B) SDS, (C) CTAB, (D) SB3-16, and (E) Tween 80 at the indicated concentrations compared to wt*Bs*LipA ($\Delta T_{50}/\Delta D = 0$). (A) Variants with ΔT_{50} lower than the experimental uncertainty (standard deviation σ_T for the respective variant) were excluded from further analyses (gray). (B–E) Variants within $2\sigma_D$ of ΔD of wt*Bs*LipA determined from screenings of 2997 wt*Bs*LipA replicates toward the respective detergent were excluded from further analyses (gray). The insets show the numbers of variants which cause a significant increase or decrease in T_{50} or D toward one detergent. A red (blue) color indicates a significantly increased (decreased) T_{50} or D toward one detergent.

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = \frac{\text{No. of predicted weak spots that are hot spots}}{\text{No. of weak spots}} \quad (9)$$

$$p_{\mathrm{random}} = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} = \frac{\text{No. of hot spots}}{\text{181 residues of } Bs\text{LipA}} \quad (10)$$

The precision in *random classification* (p_{random}) indicates how many of the 181 *Bs*LipA residues are actually hot spots (eq 10).[86]

The *gain in precision* over random classification (gip) represents how many predicted weak spots are actually hot spots in comparison to random classification (eq 11).[86] The gip

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

Table 1. Identified Classes of Substitution Sites

class[a]	definition	no. of substitution sites	no. of weak spots[b]	gip[c]
I	{substitution site$_x$ | $1 \le x \le 181$, $T_{50}(x)$ is significantly increased}	69	nd[d]	nd[d]
II	{substitution site$_x$ | $1 \le x \le 181$, $D_{SDS}(x)$ is significantly increased}	74	nd[d]	nd[d]
III	{substitution site$_x$ | $1 \le x \le 181$, $D_{CTAB}(x)$ is significantly increased}	42	nd[d]	nd[d]
IV	{substitution site$_x$ | $1 \le x \le 181$, $D_{SB3\text{-}16}(x)$ is significantly increased}	46	nd[d]	nd[d]
V	{substitution site$_x$ | $1 \le x \le 181$, $D_{Tween80}(x)$ is significantly increased}	34	nd[d]	nd[d]
VI	II ∪ III ∪ IV ∪ V	109	nd[d]	nd[d]
VII	I ∪ VI	124	nd[d]	nd[d]
VIII	II ∩ III ∩ IV ∩ V	11	2	3.30
IX	I ∩ VIII	7	2	5.17
X	{substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $T_{50}(x)$}	6	1	3.02
XI	{substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $D_{SDS}(x)$}	6	1	3.02
XII	{substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $D_{CTAB}(x)$}	6	3	9.05
XIII	{substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $D_{SB3\text{-}16}(x)$}	6	2	6.03
XIV	{substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $D_{Tween80}(x)$}	6	0	–
XV	XI ∪ XII ∪ XIII ∪ XIV	20	nd[d]	nd[d]
XVI	X ∪ XV	24	nd[d]	nd[d]
XVII	XI ∩ XII ∩ XIII ∩ XIV	0	nd[d]	nd[d]
XVIII	X ∩ XVII	0	nd[d]	nd[d]

[a]Class of substitution sites; underlined classes represent hot spots. [b]Numbers of hot spots that are predicted as weak spots. [c]Gain in precision over random classification (eq 11). [d]Not determined.

range is $[0, \infty]$, with values <1 indicating a lower precision than obtained by random classification.

$$\text{gip} = \frac{p}{p_{\text{random}}} \tag{11}$$

The F_1-score (F_1) is a measure of the test's accuracy. It represents the harmonic mean of p and r; i.e., if there is an uneven class distribution, it is used to seek a balance between p and r (eq 12).[87] The F_1 range is $[0, 1]$, with 1 indicating perfect r and p.

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r} \tag{12}$$

2.11. Markov Chain Monte Carlo-Based Unfolding Simulations of wt*Bs*LipA. As an independent method to assess the order of unfolding of wt*Bs*LipA, we used a Markov Chain Monte Carlo (MCMC) simulation with an all-atom model restricted to dihedral degrees of freedom.[88] This method has been successfully used for protein-folding simulations[89] and has been shown to reproduce the order of melting temperatures for a set of protein variants.[90] In this MCMC model, implemented in the open source tool ProFASi (Protein Folding and Aggregation Simulator), the protein conformation is modified by changing one or few dihedral angles in each step. A step is accepted according to the Metropolis criterion, i.e., with a probability that depends on the absolute temperature and the resulting change of energy of the system. In ProFASi, the energy is calculated by an all-atom implicit solvent force field.[90,91] While MCMC simulations allow arbitrarily large changes to the molecule, the unfolding simulations for this study have been restricted to side chain dihedral updates and small, locally correlated updates of main chain dihedral angles.[92] To ensure adequate sampling, 96 MCMC simulations at 330 K were performed with a total of 3.05×10^{10} elementary updates.

3. RESULTS

3.1. About One-Tenth of All Variants in the Complete SSM Library Show Significantly Increased T_{50} or D toward at Least One Detergent, and Such Variants Were Found at Two-Thirds of All Substitution Sites. The *Bs*LipA SSM library contained T_{50} as well as D data toward the four detergents SDS, CTAB, SB3-16, and Tween 80 for all 3439 single variants (181 substitution sites of *Bs*LipA × 19 naturally occurring AAs), including also inactive variants (see section 2.1). Initially, the results of both experimental screening studies of the SSM library with respect to changes in T_{50} (ΔT_{50}) or D toward one detergent (ΔD) were assessed in terms of the variance of the data and its significance (see section 2.2).

As to the T_{50} data, only variants with ΔT_{50} higher (lower) than the experimental uncertainty, taken as the standard deviation σ_T for the respective variant determined from three screenings of T_{50}, were considered significantly increased (decreased) in T_{50} compared to wt*Bs*LipA ($\Delta T_{50} = 0$ K) (eq 1). The average σ_T is 0.44 K. In total, 1856 variants with significantly increased T_{50} were obtained, of which 214 (~12%) show an increase and 1642 (~88%) a decrease (Figure 1A, Table S1). This proportion represents what one would obtain in the case of *random mutagenesis*. The distribution of ΔT_{50} is left-skewed, with extreme ΔT_{50} values of −8.3 and +7.7 K, with the most frequent ΔT_{50} range being −2 to −1.5 K (~12% out of 1856 variants), followed by ΔT_{50} between −1.5 and −1 K (~10% out of 1856 variants) (Figure 1A). In turn, for each of 69 substitution sites (~38% out of 181 substitution sites) at least one variant with significantly increased T_{50} was found. These substitution sites are summarized in class **I** (**I** = {Substitution site$_x$ \| $1 \le x \le 181$, $T_{50}(x)$ is significantly increased}) (Tables 1 and S2).

Likewise, only variants with ΔD higher (lower) than two times the experimental standard deviation ($2\sigma_D$) of wt*Bs*LipA determined from screenings of 2997 wt*Bs*LipA replicates[19] toward the respective detergent were considered significantly increased (decreased) in D compared to wt*Bs*LipA ($\Delta D = 0$) (eq 2). The screening revealed the highest σ_D in the presence of SB3-16, followed by Tween 80, CTAB, and SDS (Table S1).[19] This may be related to the fact that SB3-16 and Tween 80 were tested above the critical micelle concentration (cmc), while CTAB and SDS were tested below it.[19,93] The respective detergent concentration had been chosen based on the inactivation of purified wt*Bs*LipA (Table S1).[19] On average, 900 variants with

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

Journal of Chemical Information and Modeling pubs.acs.org/jcim Article

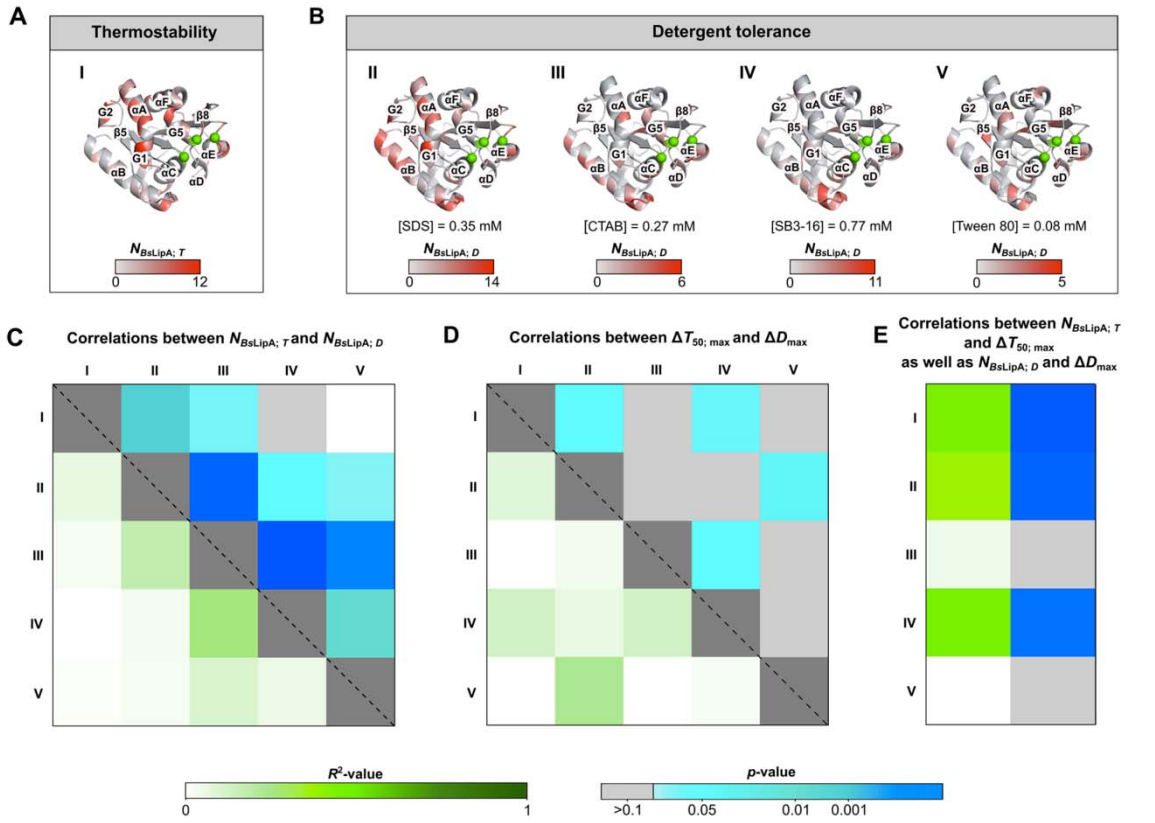


Figure 2. Localization of *Bs*LipA variants as to the frequency of substitution occurrences and highest effects regarding significantly increased T_{50} or D toward one detergent. (A) The maximum number of substitutions that cause significantly increased (A) T_{50} ($N_{BsLipA;T}$) of **I** (**I** = {Substitution site$_x$ | $1 \leq x \leq 181$, $T_{50}(x)$ is significantly increased}) or (B) D ($N_{BsLipA;D}$) of **II**–**V** (**II**–**V** = {Substitution site$_x$ | $1 \leq x \leq 181$, $D_{SDS/CTAB/SB3\text{-}16/Tween\ 80}(x)$ is significantly increased}) are mapped onto wt*Bs*LipA (PDB ID: 1ISP). C_α atoms of the catalytic triad S77/D133/H156 are shown as green spheres. A red (gray) color indicates a high (low) $N_{BsLipA;T}$ of **I** or $N_{BsLipA;D}$ of **II**–**V**. (C) R^2- and p-values for correlations between $N_{BsLipA;T}$ of **I** or $N_{BsLipA;D}$ of **II**–**V**. (D) Additionally, an analysis of the respective highest effects in significantly increased T_{50} ($\Delta T_{50;\,max}$) of **I** or D (ΔD_{max}) of **II**–**V** was performed. Here, R^2- and p-values for correlations between $\Delta T_{50;\,max}$ of **I** or ΔD_{max} of **II**–**V** are shown. (E) R^2- and p-values for correlations between $N_{BsLipA;T}$ and $\Delta T_{50;max}$ of **I** or $N_{BsLipA;D}$ and ΔD_{max} of **II**–**V**.

significantly increased D were obtained, of which 126 (~14%) show an increase and 774 (~86%) a decrease, on average across each detergents (Figures 1B–E, Table S1). This proportion represents what one would obtain in the case of random mutagenesis. The distribution of ΔD is left-skewed. The magnitude of the increase (decrease) in ΔD is between 1.6-fold and 2.4-fold (0.6-fold and 2.9-fold) of the residual activity of wt*Bs*LipA. Furthermore, variants tested against SDS and SB3-16 showed an up to two times higher ΔD than against CTAB and Tween 80 (Figures 1B–E). This may be related to the different classes of the detergents.[19,93] In turn, for each of 74, 42, 46, or 34 substitution sites at least one variant with significantly increased D toward SDS, CTAB, SB3-16, or Tween 80 (~41, 23, 25, or 19% out of 181 substitution sites) was found. These substitution sites are summarized in classes **II**–**V** (**II**–**V** = {Substitution site$_x$ | $1 \leq x \leq 181$, $D_{SDS/CTAB/SB3\text{-}16/Tween\ 80}(x)$ is significantly increased}) (Tables 1 and S2). The union of **II**–**V** contains 109 substitution sites (~60% out of 181 substitution sites) and is represented by class **VI** (**VI** = **II** ∪ **III** ∪ **IV** ∪ **V**) (Tables 1 and S2, eq 3). For each of these substitution sites at least one variant shows significantly increased D toward at least one detergent.

Finally, 124 substitution sites are summarized in the union of **I** and **VI** (~69% out of 181 substitution sites) (**VII** = **I** ∪ **VI**) (Tables 1 and S2, eq 3). Thus, only for two-thirds of all substitution sites at least one variant with significantly increased T_{50} or D toward at least one detergent was obtained.

To conclude, for the first time, we performed a systematic large-scale analysis of a complete experimental SSM library toward two types of stabilities of one protein containing all single variants. The likelihoods to generate variants with significantly increased T_{50} (~12%) or D toward one detergent (~14% on average across all detergents) by random mutagenesis (**I**–**V**) are similar. Variants with significantly increased T_{50} or D toward at least one detergent were obtained at only two-thirds of all substitution sites (**VII**), and this value falls to about one-third or below if T_{50} and D toward one detergent are considered separately (**I**–**V**). Hence, such substitution sites are not uniformly distributed across the protein. For the following analyses, only substitution sites with at least one variant yielding significantly increased T_{50} or D toward at least one detergent were considered.

3.2. The Higher the Frequency of Substitution Occurrences That Lead to Significantly Increased T_{50} or

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

***D* toward One Detergent, the More Pronounced the Highest Effect, and Vice Versa.** Next, we investigated the *Bs*LipA SSM library regarding the respective frequency of substitution occurrences at substitution sites that lead to significantly increased T_{50} ($N_{BsLipA;T}$) or D ($N_{BsLipA;D}$) toward one detergent. Additionally, we analyzed the respective highest effects in significantly increased T_{50} ($\Delta T_{50;max}$) or D (ΔD_{max}) toward one detergent at substitution sites. Finally, we address the question if the frequency of substitution occurrences and the highest effects per substitution site are related to each other.

The highest $N_{BsLipA;T}$ of **I** was 12 (F17) (Figure 2A), whereas the highest $N_{BsLipA;D}$ of **II**−**V** were 14 (E65), 6 (I135 and D144), 11 (G46), and 5 (V99) (Figure 2B, Table S14), respectively, indicating that up to ∼60% and more of the variants for some substitution sites yield significantly increased T_{50} or D toward one detergent. Correlations between $N_{BsLipA;T}$ of **I** and $N_{BsLipA;D}$ of **II**−**V** yielded, on average, $R^2 = 0.03$; $p > 0.1$ (Figure 2C, Table S3). The highest correlation was found between $N_{BsLipA;T}$ of **I** and $N_{BsLipA;D}$ of **II** ($R^2 = 0.07$, $p < 0.001$). With respect to $N_{BsLipA;D}$ of **II**−**V**, overall very weak to weak but mostly significant correlations were obtained (on average: $R^2 = 0.11$, $p < 0.01$) (Figure 2C, Table S3). The highest correlation was observed between $N_{BsLipA;D}$ of **III** and **IV** ($R^2 = 0.26$, $p < 0.001$).

The highest $\Delta T_{50;max}$ of **I** was 7.7 K (M137), whereas the highest ΔD_{max} of **II**−**V** were 1.49 (M137), 1.63 (T110), 2.41 (G46), and 2.29 (S127), respectively (Table S9), indicating that specific single AA substitutions have a great impact on the magnitudes of the effects. Correlations between $\Delta T_{50;max}$ of **I** and ΔD_{max} of **II**−**V** shown, on average, $R^2 = 0.06$; $p > 0.1$ (Figure 2D, Table S4). The highest correlation was observed between $\Delta T_{50;max}$ of **I** and ΔD_{max} of **IV** ($R^2 = 0.13$, $p < 0.1$). With respect to ΔD_{max} of **II**−**V**, overall very weak to weak and mostly insignificant correlations were obtained (on average: $R^2 = 0.08$, $p > 0.1$) (Figure 2D, Table S4). The highest correlations were observed between ΔD_{max} of **II** and **V** ($R^2 = 0.24$, $p < 0.05$) as well as ΔD_{max} of **III** and **IV** ($R^2 = 0.13$, $p < 0.1$).

Finally, mostly good to fair and significant correlations between $N_{BsLipA;T}$ and $\Delta T_{50;max}$ of **I** as well as $N_{BsLipA;D}$ and ΔD_{max} of **II**−**V** were found (on average for increase: $R^2 = 0.27$, $p < 0.01$) (Figure 2E, Table S5).

To conclude, these findings indicate that the relation "the higher the frequency of substitution occurrences that lead to significantly increased T_{50} or D towards one detergent, the more pronounced the highest effect, and vice versa" holds for substitution sites at which at least one variant shows significantly increased T_{50} or D toward one detergent (**I**−**V**). Together with the results from the previous chapter, this result suggests that identifying a priori substitution sites with a high likelihood for significantly increased T_{50} or D toward one detergent will also be beneficial with respect to the magnitude of effects that can be achieved there by substitutions.

3.3. Eleven Substitution Sites Yield a ∼4.6-fold Higher Likelihood To Find for Each Detergent Variants with Significantly Increased *D* than Random Mutagenesis. Next, we focused on pairwise intersections of **II**−**V** to investigate if there are substitution sites at which for two detergents at least one variant shows significantly increased D, regardless of the magnitude of the single effect (see section 2.3). We compared the pairwise similarities between **II**−**V** by calculating the Jaccard index (J), i.e., the cardinal number of the respective intersection divided by the cardinal number of the respective union (Table S6, eq 5).[56,57] The highest similarity was found between **III** and **IV** with J(**III**, **IV**) = 0.47, whereas the lowest similarity was observed between **II** and **V** with J(**II**, **V**) = 0.23. This may be related to the different classes of the detergents[19,93]

Encouraged by the findings of overlapping **II**−**V**, we also looked at the overall intersection of **II**−**V** (**VIII** = **II** ∩ **III** ∩ **IV** ∩ **V**), i.e., substitution sites at which for each detergent at least one variant shows significantly increased D, regardless of the magnitude of the single effect (Tables 1 and S2, eq 4). **VIII** contains the 11 substitution sites E2, G13, D43, T45, Y49, N51, V54, E65, N98, M134, and M137 (∼6% out of 181 substitution sites) (Tables 1, S2, and S14). These substitution sites are associated with 50 variants causing a significant change in D, of which 32 (∼64%) show a significant increase, on average across all detergents (Table S7). Thus, this likelihood is ∼4.6-fold higher in comparison to random mutagenesis. The most promising substitution sites of **VIII** are M134, N51, and T45 with variants showing increased ΔD_{max} of 2.25, 2.10, and 1.90, respectively.

To conclude, a dramatically reduced number of 11 substitution sites (**VIII**) yield a ∼4.6-fold higher likelihood to find for each detergent variants with significantly increased D compared to random mutagenesis. These findings indicate that if a protein-engineering study aims at identifying variants showing significantly increased D toward each detergent, such substitution sites (**VIII**) should be identified prior to SDM.

3.4. Seven Substitution Sites Yield a ∼3.4-fold Higher Likelihood To Find Variants with Significantly Increased T_{50} and a ∼4.7-fold Higher Likelihood To Find for Each Detergent Variants with Significantly Increased *D* than Random Mutagenesis. The same analyses were repeated for intersections of **I** and **II**−**V**, respectively, regarding substitution sites at which at least one variant shows significantly increased T_{50} and for one detergent significantly increased D, regardless of the magnitude of the single effect (see section 2.3). We compared the pairwise similarities between **I** and **II**−**V**, respectively, by calculating J (Table S6, eq 5). The highest similarity was found between **I** and **II** with J(**I**, **II**) = 0.42, whereas the lowest similarity was observed between **I** and **V** with J(**I**, **V**) = 0.16.

Encouraged by the findings of overlapping **I** and **II**−**V**, respectively, we also looked at the overall intersection of **I** and **II**−**V** (**IX** = **I** ∩ **VIII**), i.e., substitution sites at which at least one variant shows significantly increased T_{50} and for each detergent significantly increased D, regardless of the magnitude of the single effect (Tables 1 and S2, eq 4). **IX** contains the seven substitution sites, E2, G13, T45, Y49, V54, M134, and M137 (∼4% out of 181 substitution sites) (Tables 1, S2, and S14). Associated with these are 86 variants causing a significant change in T_{50}, of which 35 (∼41%) show a significant increase (Table S8). Thus, this likelihood is ∼3.4-fold higher in comparison to random mutagenesis. The most promising substitution sites of **IX** are M137, M134, and Y49 with variants showing increased $\Delta T_{50;max}$ of 7.7, 5.6, and 1.6 K, respectively. Furthermore, associated with substitution sites of **IX** are 29 variants causing a significant change in D, of which 19 (∼66%) show a significant increase, on average across all detergents (Table S8). Thus, this likelihood is ∼4.7-fold higher in comparison to random mutagenesis. The most promising substitution sites of **IX** are M134, T45, and M137 with variants showing increased ΔD_{max} of 2.25, 1.90, and 1.67, respectively.

To conclude, a dramatically reduced number of seven substitution sites (**IX**) yield a ∼3.4-fold higher likelihood to find variants with significantly increased T_{50} and a ∼4.7-fold higher likelihood to find for each detergent variants with

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

significantly increased D compared to random mutagenesis. These findings indicate that if a protein-engineering study aims at identifying variants showing significantly increased T_{50} and D toward each detergent, such substitution sites (**IX**) should be identified prior to SDM.

3.5. Six Substitution Sites with Highest $\Delta T_{50;max}$ (ΔD_{max}) Yield a ~5.3-fold (~4.5-fold) Higher Likelihood To Find Variants with Significantly Increased T_{50} (D) than Random Mutagenesis. The above analyses focused on substitution sites at which significantly increased T_{50} or D toward one detergent (**I**–**V**), significantly increased D toward each detergent (**VIII**), as well as significantly increased T_{50} and D toward each detergent (**IX**) were observed, regardless of the magnitude of the effect. Now, we identified those six substitution sites for which the respective highest effects ($\Delta T_{50;max}$ or ΔD_{max}) were found. The number of 6 is motivated by the current technical limitation to screen more than 20^6 variants.[20,39,49,50]

The six substitution sites M137, M134, G155, F17, I157, and Y139 yield variants with the highest $\Delta T_{50;\,max}$ of 7.7, 5.6, 4.5, 3.8, 3.6, and 3.2 K, respectively, and constitute class **X** (**X** = {Substitution sites$_x$ | $1 \leq x \leq 181$, six highest effects in significantly increased $T_{50}(x)$}) (Tables 1, S2, and S9). The substitution sites of **X** are associated with 68 variants causing a significant change in T_{50}, of which 43 (~63%) yield a significantly increased T_{50} (Table S10). Thus, this likelihood is ~5.3-fold higher in comparison to random mutagenesis.

The most promising substitution sites exhibiting variants with the highest ΔD_{max} toward one detergent (**XI**–**XIV** = {Substitution sites$_x$ | $1 \leq x \leq 181$, six highest effects in significantly increased $D_{SDS/CTAB/SB3\text{-}16/Tween\ 80}(x)$}) are M137 (**XI**), T110 (**XII**), G46 (**XIII**), and S127 (**XIV**) with variants showing highest ΔD_{max} of 1.49, 1.63, 2.41, and 2.29, respectively (Tables 1, S2, and S9). With these substitution sites, 43 variants are associated causing a significant change in D, of which 27 (~63%) cause significantly increased D, on average across all detergents (Table S10). Thus, this likelihood is ~4.5-fold higher in comparison to random mutagenesis.

Furthermore, we determined the union of **XI**–**XIV**, the set of 20 substitution sites (~11% out of 181 substitution sites) that yield variants showing the respective highest ΔD_{max} toward at least one detergent (**XV** = **XI** ∪ **XII** ∪ **XIII** ∪ **XIV**) (Tables 1 and S2, eq 3). Additionally, the union of **X** and **XV** was defined as the set of 24 substitution sites (~13% out of 181 substitution sites), which exhibit variants showing the respective highest $\Delta T_{50;max}$ or ΔD_{max} toward at least one detergent (**XVI** = **X** ∪ **XV**) (Tables 1 and S2, eq 3).

The intersection between **XI**–**XIV** (**XVII** = **XI** ∩ **XII** ∩ **XIII** ∩ **XIV**) is empty; i.e., there are no common substitution sites among those six at which for each detergent variants with highest ΔD_{max} were found (Tables 1 and S2, eq 4). The intersection between **X** and **XVII** (**XVIII** = **X** ∩ **XVII**) is necessarily empty, too; i.e., there are no common substitution sites among those six at which variants with highest $\Delta T_{50;max}$ and ΔD_{max} for each detergent were found (Tables 1 and S2, eq 4). Thus, **XVII** and **XVIII** were not considered for the following analyses.

Additionally, we compared the pairwise similarities between **X**–**XIV** by calculating J (eq 5). Regarding the highest ΔD_{max}, only **XII** and **XIII** overlap to some extent (J(**XII**, **XIII**) = 0.2) (Table S6). Regarding the highest $\Delta T_{50;max}$ and ΔD_{max}, only **X** and **XI**, **XII**, or **XIII**, respectively, slightly overlap (J(**X**, **XI**) ≈ J(**X**, **XII**) ≈ J(**X**, **XIII**) = 0.1) (Table S6).

To conclude, a highest $\Delta T_{50;\,max}$ of 7.7 K and a highest ΔD_{max} of 2.41 were found. The six substitution sites with highest $\Delta T_{50;max}$ yield a ~5.3-fold higher likelihood to find variants with significantly increased T_{50} (**X**); the six substitution sites with highest ΔD_{max} yield a ~4.5-fold higher likelihood to find variants with significantly increased D (**XI**–**XIV**). There are no common substitution sites among those six at which for each detergent variants with highest ΔD_{max} were found (**XVII**). Neither are there common substitution sites among those six at which variants with highest $\Delta T_{50;max}$ and ΔD_{max} for each detergent were found (**XVIII**).

3.6. Definition of Hot Spots. Based on these results, we defined seven types of hot spots, i.e., substitution sites particularly promising to cause a significant increase in T_{50} or/and D. First, the respective six substitution sites of **X**–**XIV** are considered hot spots because variants yield the respective highest $\Delta T_{50;max}$ or ΔD_{max} toward one detergent for these substitution sites (Tables 1, S2, and S9). Furthermore, we showed that there is a correlation between the magnitude of an effect found at a substitution site and the frequency of substitution occurrences that lead to significantly increased T_{50} or D toward one detergent (see section 3.2). Finally, generating and evaluating variants based on combinations of all 20 AAs at six substitution sites is still manageable with current protein-engineering techniques.[20,39,49,50]

As shown above, **XVII** and **XVIII**, which would constitute the substitution sites with the broadest impact on ΔD_{max}, or $\Delta T_{50;\,max}$ and ΔD_{max}, are empty (see section 3.5). Hence, we resorted to defining, second, the 11 substitution sites of **VIII** showing significantly increased D toward each detergent, regardless of the magnitude of the single effect (see section 3.3) and, third, the seven substitution sites of **IX** showing significantly increased T_{50} and D toward each detergent, regardless of the magnitude of the single effect (see section 3.4) as hot spots (Tables 1 and S2). With 11 and 7 substitution sites, these classes are also the smallest besides **X**–**XIV**.

3.7. Hot Spots Are Diverse in Terms of Localization in Secondary Structure Elements, Degree of Burial, and Sequence-Based Characteristics of the Substituted AAs. Ideally, one would identify such hot spots based on structural or sequence characteristics of the protein (see sections 2.4 and 2.5) prior to performing experiments. Suitable structure-based characteristics are localization in secondary structure elements (Table S11)[19,94−96] and the degree of burial as measured by fSASAs (Table S12, eq 6).[19,97,98]

As to localization in secondary structure elements (Table S11), hot spots are rarely found in 3_{10}-helices and β-strands. Exceptions are hot spots of class **XIV**, which are enriched in strand β7. With respect to α-helices, at least one and at most four hot spot(s) of each class is (are) found in that secondary structure class, mainly in helices αB and αE. However, without further information, one would not know which particular secondary structure element to choose for hot spot prediction. Hence, if all sites of a certain secondary structure class were chosen as hot spots, in the best case, a gain in precision (gip, eq 11) over random classification of 4.71 is found for β-strands, albeit at the expense of predicting 32 substitution sites (~18% of 181 AAs), far more than the 6 sought. As to bridges, turns, loops, and bends defined by DSSP,[58] no hot spot is found in the first secondary structure type. At most three hot spots are found in any of the other three types, but only for hot spots of class **XI** and **VIII**. These cases are related to a maximal gip of 1.93, albeit at the expense of predicting 47 substitution sites (~26% of 181 AAs). Thus, in our study, identifying hot spots based on this secondary structure type results in a low precision.

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

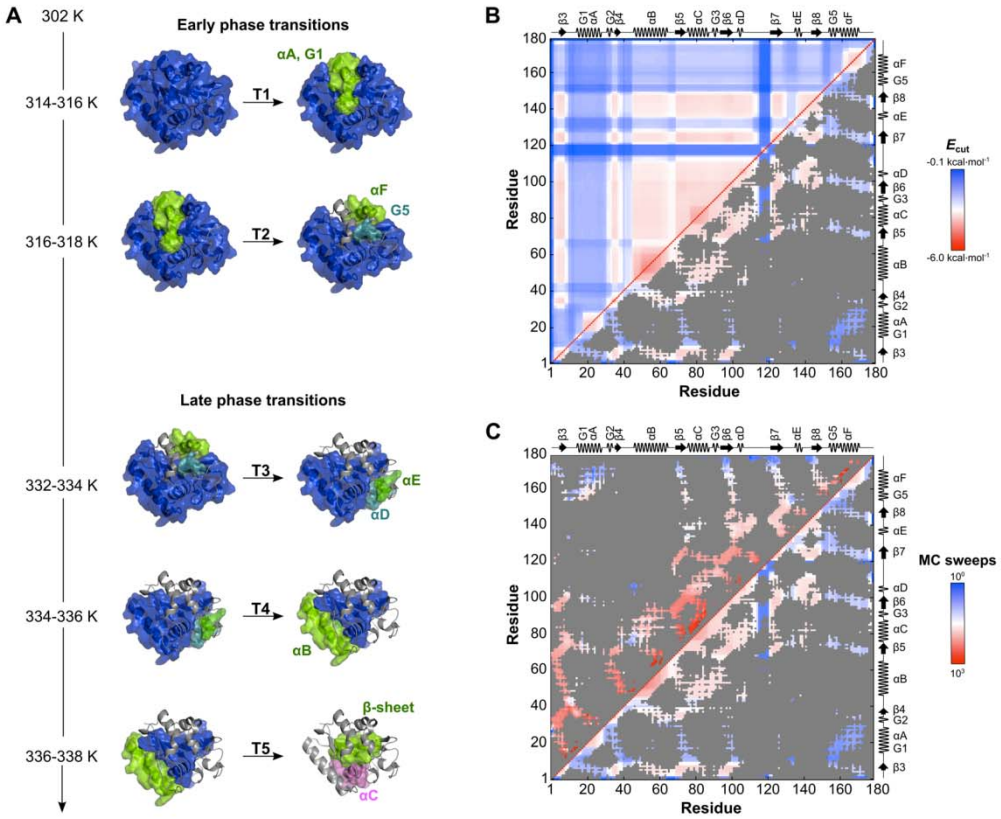


Figure 3. Prediction of the thermal unfolding pathway of wt*Bs*LipA. (A) Thermal unfolding pathway of wt*Bs*LipA (PDB ID: 1ISP) showing the early (T1−T2) and late (T3−T5) phase transitions. Rigid clusters are represented as uniformly colored blue, green, magenta, and cyan bodies in the descending order of their sizes. (B) For wt*Bs*LipA the stability map rc_{ij} including E_{cut} values at which a rigid contact between two residues is lost for all residue pairs during the thermal unfolding simulation (upper triangle); the neighbor stability map $rc_{ij,neighbor}$ considering only the rigid contacts between two residues that are at most 5 Å apart from each other, with values for all other residue pairs colored gray (lower triangle). The E_{cut} values are calculated with CNA based on a structural ensemble (ENT^{MD}). A red (blue) color indicates that contacts between residue pairs are more (less) rigid. (C) The aforementioned $rc_{ij,neighbor}$ (lower triangle) was compared with a contact map simulated by ProFASi (upper triangle). A red (blue) color indicates contacts between residue pairs that have a longer (shorter) lifetime (in MC sweeps) than the contacts of the residue pairs of the initial protein structure. 3_{10}-helices are represented as G-helices.

As to the degree of burial (Table S12), the least hot spots are associated with substitution sites that are mostly solvent-exposed (0.8 < fSASA ≤ 1.0). By contrast, the most hot spots are associated with substitution sites that are partially solvent-exposed (0.6 < fSASA ≤ 0.8), although this statement does not hold for hot spots of class **XIV**. This case is related to a maximal gip of 6.70, albeit at the expense of predicting 18 substitution sites (∼10% of 181 AAs).

Suitable sequence-based characteristics are physicochemical properties of the substituted AAs (Table S13)[19,99−101] and the degree of AA conservation (Table S14).[19,102,103] As to the physicochemical properties of the substituted AAs (Table S13), the distribution of hot spots over the classes is generally broad. Exceptions are hot spots of classes **XIII** and **XIV** (in both cases preferentially found at aliphatic and neutral AAs (Table S15)) and class **X** (preferentially found at aliphatic, aromatic, and neutral AAs (Table S15)). Therefore, gip values are generally low, with the largest one being 4.02 for the case of hot spots of class **X** at aromatic AAs, albeit at the expense of predicting 15 substitution sites (∼8% of 181 AAs). As to the degree of AA conservation, hot spots are located at nonconserved and semiconserved positions (conservation in the range of 0−6) (Table S14). The highest conservations were found for I128 (conservation = 6) and V99, T126, and I128 (conservation = 5).

To conclude, while predicting hot spots based on structural characteristics can lead to marked gip values, usually many predicted hot spots result, which would require considerable experimental efforts. Still, if a higher number of predicted hot spots is acceptable, partially solvent-exposed residues are good hot spot candidates. Applying sequence-based characteristics, substituting aliphatic and neutral residues should more likely improve T_{50} or/and D. Additionally, nonconserved and semiconserved regions preferentially contain hot spots.

3.8. Rigidity Theory-Based (CNA) and Markov Chain Monte Carlo Simulation-Based (ProFASi) Approaches Predict Similar Thermal Unfolding Pathways of wt*Bs*LipA. We intend to test if hot spots can be predicted as structural weak spots by our rigidity theory-based approach CNA[66−68] (see section 2.6). As a prerequisite, information on the hierarchy of rigid and flexible regions in a protein structure is required. Therefore, a thermal unfolding simulation of wt*Bs*LipA was carried out with CNA as done previously[7,8] to predict major

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

phase transitions at which the network switches from overall rigid to flexible states (see sections 2.7, 2.8, and 2.9).

From the thermal unfolding pathway of wt*Bs*LipA, five major phase transitions, T1−T5, were predicted based on the *global* index H_{type2} (Figure 3A). Depending on the energy cutoff E_{cut}, the phase transitions were characterized as either *early* (T1−T2; with -0.8 kcal mol$^{-1} \geq E_{cut} \geq -0.9$ kcal mol^{-1}) or *late* (T3−T5; with -1.7 kcal mol$^{-1} \geq E_{cut} \geq -1.9$ kcal mol^{-1}). E_{cut} can be converted to a temperature T using a linear equation (eq 7),[2] according to which the ranges of E_{cut} in this study are equivalent to 316 K $\leq T \leq$ 318 K for T1−T2, and 334 K $\leq T \leq$ 338 K for T3−T5. During the early phase transitions αA, 3_{10}-1, αF, and 3_{10}-5 segregate from the largest rigid cluster. αD, αE, αB, αC, and β-strands segregate from the largest rigid cluster during the late phase transitions. Afterward, the β-sheet becomes sequentially flexible, beginning with β4 and β8, followed by β3, β7, β5, and β6. For the analysis, ∼3 h of computational time on a single GPU is required to generate a 100 ns long MD trajectory as well as ∼4 h of computational time on a single core for the thermal unfolding simulation.

Since the percolation behavior of a protein network is complex due to the protein's structural hierarchy and composition of different modules, it is often challenging to assign a phase transition with H_{type2}.[85] Thus, in addition to using H_{type2}, we also characterized the hierarchy of rigid and flexible regions of wt*Bs*LipA at a local level by computing $rc_{ij,neighbor}$ (lower triangle in Figure 3B) based on rc_{ij} (upper triangle in Figure 3B). $rc_{ij,neighbor}$ showed that residue pairs at the *N*-terminus revealed higher E_{cut} values than residue pairs at the *C*-terminus. Thus, $rc_{ij,neighbor}$ demonstrates that the rigid contacts between neighboring residues are stronger at the *N*-terminus than at the *C*-terminus along the thermal unfolding simulation, i.e., the *C*-terminus of wt*Bs*LipA starts to unfold first.

As an independent approach to assess the order of unfolding of wt*Bs*LipA, we used the Markov Chain Monte Carlo (MCMC) simulation software ProFASi (Protein Folding and Aggregation Simulator) (see section 2.11).[88] The results of the simulation were represented in a contact map (upper triangle in Figure 3C). They reveal that the contacts between the residue pairs of the *N*-terminus have a longer lifetime (in terms of MC sweeps) than the contacts of the residue pairs of the *C*-terminus compared to the initial structure. Thus, although methodologically different, ProFASi predicts a very similar unfolding pathway of wt*Bs*LipA with respect to CNA.

To conclude, five major phase transitions, T1−T5, were predicted by thermal unfolding simulations using CNA at which first the different helices and, finally, the β-strands segregate from the largest rigid cluster during thermal unfolding simulations of wt*Bs*LipA by CNA. Structural rigidity is initially lost at the *C*-terminus, which is uniformly revealed by the global index H_{type2} and the local index $rc_{ij,neighbor}$. Finally, the two independent approaches CNA and ProFASi predict very similar unfolding pathways of wt*Bs*LipA. The results suggest that the loss of rigidity predicted by CNA along the thermal unfolding simulation closely mimics the temperature-induced unfolding of wt*Bs*LipA.

3.9. Unfolding Nuclei and Major Phase Transitions Are Predictive Markers of Structural Weak Spots. We next probed to what extent structural weak spots predicted by CNA agree with the above-defined hot spots. Following previous work,[2] weak spots are fringe residues of unfolding nuclei that percolate from the largest rigid cluster during earlier steps of the thermal unfolding (see section 2.9). In total, we predicted 10 weak spots (∼6% out of 181 substitution sites), i.e., I12, G13, G46, G52, P53, T66, M134, I135, V136, and H152 (Figure 4A, Tables 1, 2, and S2). Three weak spots each segregate from the largest rigid cluster at T1 or T2, and four from the largest rigid cluster at T4 (Table 2).

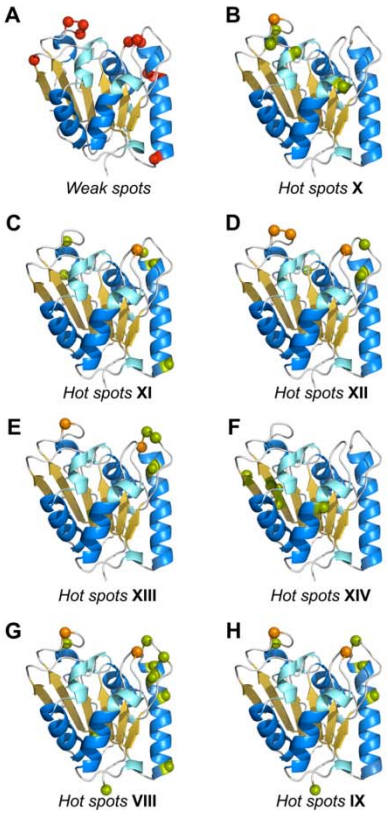


Figure 4. Localization of CNA-predicted weak spots and experimental hot spots of*Bs*LipA. (A) Weak spots and (B) hot spots of **X**, (C) **XI**, (D) **XII**, (E) **XIII**, (F) **XIV**, (G) **VIII**, and (H) **IV** are mapped onto wt*Bs*LipA (PDB ID: 1ISP). (A) Ten weak spots, i.e., I12, G13, G46, G52, P53, T66, M134, I135, V136, and H152, were predicted by CNA (red spheres). (B−F) The respective six substitution sites of **X**−**XIV** are considered hot spots as variants yield the respective six highest $\Delta T_{50;max}$ or ΔD_{max} toward one detergent for these substitution sites. (G) The 11 substitution sites of **VIII** showing significantly increased D toward each detergent, regardless of the magnitude of the single effect, and (H) the seven substitution sites of **IX** showing significantly increased T_{50} and D toward each detergent, regardless of the magnitude of the single effect, are considered hot spots. A green sphere represents a hot spot, and an orange sphere indicates a hot spot that was correctly predicted as a weak spot.

The performance of predicting hot spots as weak spots by CNA was evaluated in terms of a binary classification, considering predicted *weak spots* at hot spots true positives (TP) and predicted weak spots at not-hot spots false positives (FP) (see section 2.10). In particular, the gain in precision over random classification (gip) (eq 11) and the F_1-score (F_1) (eq 12), a measure of a classifier's accuracy, were used as performance measures. Over all seven classes of hot spots, between one and three of the predicted weak spots are hot spots (except for **XIV**, where no weak spot was met), resulting in gip

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

Table 2. CNA-Predicted Weak Spots of *Bs*LipA

weak spot	location at secondary structure elements	phase transition
I12	turn	T1
G13	turn	T1
G46	toop	T4
G52	αB	T4
P53	αB	T4
T66	αB	T4
M134	bend	T2
I135	bend	T2
V136	bend	T2
H152	bend	T1

values between 3.02 and 9.05 (Tables 1 and S2). Note that these results are associated with only 10 predicted weak spots, about half as many predictions than in the case of identifying hot spots as partially solvent-exposed residues (Table S12). As the numbers of hot spots in **VIII**–**XIV** are of a very similar magnitude, the CNA predictions are also associated with similar recall (r) (eq 8) and precision (p) values (eq 9) in each case (Table S2), indicating a well-balanced classifier. In the case of **XII**, the CNA predictions yield an F_1-score of 0.38, higher than any F_1-score associated with hot spot predictions based on structure or sequence characteristics (Tables S2, S11, S12, S13, and S14), and the F_1-score for **IX** is 0.24, generally higher than F_1-scores associated with structure- or sequence-based predictions for this class and on par with the result obtained for identifying these hot spots as partially solvent-exposed residues (Tables S2, S11, S12, S13, and S14).

To conclude, predicting hot spots as weak spots by CNA results in several cases in very good to good gip values and good to fair accuracies and is associated with a very low number of predicted weak spots, such that also only few experimental efforts are required later. Considering the low computing time required to perform a CNA analysis, these results indicate that applying CNA-based weak spot prediction before experimental engineering is beneficial, in particular if the number of substitution sites that can be dealt with in experiment is low.

4. DISCUSSION

In this study, for the first time, we performed a systematic large-scale analysis of a complete experimental SSM library of a biotechnologically highly relevant protein, *Bs*LipA,[52,53] with respect to two types of protein stability. The library covers all 181 residues of *Bs*LipA and results in 3439 variants, each with a single AA substitution as confirmed by DNA sequencing. Considering the screening results of the library toward thermostability and detergent tolerance together is unique compared to related studies[2,4−8,17−19] and important in view of the challenges of multidimensional property optimization of modern biocatalysts.[104−106] The measured T_{50} and D values provide valuable reference data for future analyses because, in contrast to other data sources,[34−37] the different protein stabilities were measured under respectively uniform conditions, and there is no bias toward any particular substitution type or site. Note, though, that other factors than protein stability may influence T_{50} or D values measured here,[52] including that substitutions can directly impact *Bs*LipA function, e.g., when occurring in the vicinity of the active site.[8] Moreover, the measured T_{50} and D values may be influenced by thermodynamic or kinetic factors.[7,8] Therefore, in our analysis, we focused on scrutinizing the impact of substitution sites on thermostability or/and detergent tolerance to gain generally applicable rules for data-driven protein engineering. The following results stand out:

First, across the library, the likelihoods to find variants with significantly increased T_{50} (∼12%) or D toward one detergent (∼14%) are almost identical and small. The finding that the overwhelming number of single AA substitutions introduced by random mutagenesis causes a destabilizing effect is in agreement with previous studies.[33,107−110] This finding becomes even more statistically relevant if multiple mutations need to be accumulated over generations to reach a desired effect because frequently a single, yet rather likely, destabilizing mutation is sufficient to annihilate the effect of several stabilizing ones.[20] The proportions of variants with increased T_{50} or D found here are in line with the composition of databases such as ProTherm[30] but markedly larger than the success rate of ∼2% used as a reference to evaluate the performance of FoldX.[111] Hence, beyond the single T_{50} and D data, due to the completeness of our library and the model character of our protein, our results also constitute unbiased reference data as to what efficiency can be expected for a protein system when optimizing thermostability or detergent tolerance by random mutagenesis. In turn, largest increases in T_{50} of 7.7 K and D of 2.4 found demonstrate that considerable improvements of protein stability can already be achieved by single AA substitutions. In that respect, previous studies on *Bs*LipA applying either directed evolution[44] or rational design[7,8] already yielded close-to-optimal results in terms of increased thermostability.

Second, in the context of data-driven protein engineering, we identified substitution sites for which variants yield significantly increased T_{50} or/and D. Not considering the magnitude of the increase, only about one-third or below of all *Bs*LipA residues constitute such favorable substitution sites if T_{50} and D are considered separately, demonstrating that the location of a residue within a protein structure matters with respect to a substitution effect. This result corroborates previous studies.[5,7,8] In addition, our complete SSM library allowed us to reveal for such substitution sites a significant and fair correlation between the frequency of T_{50} or/and D-increasing substitutions and the magnitude of the maximum effect. Together, these results show that addressing all substitution sites in an unbiased manner by random mutagenesis results in a considerable experimental effort coupled to low efficiency. In turn, approaches that can identify substitution sites with a high likelihood for significantly increased T_{50} or D prior to doing experiments will be of great value in protein engineering studies.

Third, notably, the conclusions from the last paragraph also hold if more than one protein property is considered at a time. As such, we showed that at 11 substitution sites a ∼4.6-fold higher likelihood to find for each detergent variants with significantly increased D compared to random mutagenesis is found. Additionally, seven substitution sites yield a ∼3.4-fold higher likelihood to find significantly increased T_{50} and a ∼4.7-fold higher likelihood to find for each detergent variants with significantly increased D compared to random mutagenesis. The latter finding suggests that approaches that can identify substitution sites with a high likelihood for significantly increased T_{50} should also be beneficial for identifying substitution sites with a high likelihood for significantly increased D, or vice versa. This is an important finding for practical applications as many more algorithms have been devised that address thermostability than detergent tolerance.

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

Fourth, as another set of reference data, we defined hot spot types together with the associated substitution sites to provide benchmark data for evaluating the performance of data-driven approaches. The first five classes follow the strict criterion that only the six substitution sites with the respective highest $\Delta T_{50;max}$ or ΔD_{max} are considered, according to that all combinations of the 20 proteinogenic AAs at such sites could still be experimentally investigated.[20,39,49,50] The intersections comprising the substitution sites with the broadest impact on ΔD_{max}, or $\Delta T_{50;\ max}$ and ΔD_{max}, are empty. Thus, we resorted to defining two further classes with the somewhat relaxed criterion that the comprised substitution sites show significantly increased D toward each detergent, or significantly increased T_{50} and D toward each detergent, regardless of the magnitude of the single effect.

Fifth, we used the complete, unbiased, and uniformly generated T_{50} and D data to probe if universal rules for modulating thermostability or detergent tolerance can be identified. We thereby focused on "one-dimensional" descriptors in terms of location in secondary structure elements, degree of burial, and physicochemical properties and conservation degree of substituted AA. Such descriptors have been widely analyzed before[112,113] and play a role in data-driven consensus approaches.[114,115] Analyzing "two- or higher dimensional" descriptors in terms of residue−residue interactions, entropic contributions or other collective phenomena, or cross-correlations of "one-dimensional" descriptors[33] remains for future work. Notably, considering our descriptors, many (up to 98 substitution sites) predicted hot spots result, which would require considerable experimental efforts particularly if beneficial substitutions need to be accumulated to reach a desired effect. This finding demonstrates on a single protein level that, with these descriptors, no universal and sufficiently discriminating rule(s) can be identified, a finding that is mirrored in studies across protein families[116,117] and with respect to low successes in assessing thermostabilities.[112] Still, if a higher number of predicted hot spots is acceptable, partially solvent-exposed residues are good hot spot candidates. This result differs from previous experimental studies showing that especially surface remodeling emerged as an effective strategy to improve protein stability.[118,119] Furthermore, loop positions, which have elsewhere been identified to preferentially carry favorable substitution sites,[120,121] show mostly destabilizing effects. Finally, and likely surprisingly, hot spots were preferentially found at nonconserved and semiconserved position, a finding that may help refine future consensus concepts where multiple sequence alignments are used to substitute nonconsensus residues by consensus ones.[42,122]

Sixth, we made use of the reference data to unequivocally benchmark our ensemble- and rigidity theory-based CNA approach with respect to predicting hot spots as structural weak spots of the protein. In contrast to previous studies on much smaller data sets,[2,4,5,8] the present work allows to systematically assess the quality of our predictions. To do so, and in contrast to other assessments of protein stability predictors,[29,30] we apply recall and precision as basic statistical measures, rather than sensitivity and specificity, because we are interested in the accuracy of predicting hot spots and not not-hot spots, the latter of which furthermore clearly dominate the data set in terms of occurrence frequency. Methodologically, CNA differs from other state-of-the art methods that do not consider ensemble representations of the protein.[113,123−127] Furthermore, CNA does not require system-specific weighting or fitting parameters.[113,125,128,129] This should make the results obtained here with CNA transferable to other protein systems. Weak spot prediction by CNA relies on a realistic modeling of the thermal unfolding of a protein.[66−68] The predicted major phase transitions and the order of the segregating secondary structure elements are in agreement with previous computational studies and experimental observations on other proteins with an α/β hydrolase fold.[130,131] Furthermore, we confirmed the unfolding pathway of wt*Bs*LipA predicted by CNA with the independent MCMC-based ProFASi approach. From a practical point of view, it is relevant that CNA predicted only 10 weak spots, allowing to focus subsequent substitution efforts on only ∼6% of the protein residues. Furthermore, the gain in precision over random classification is between ∼3 and ∼9, depending on the hot spot class. Considering the properties of the majority of predicted weak spots, i.e., a location in a loop, turn, or bend and a neutral or aliphatic amino acid type (Table 2), the notion may arise that these two properties, when correlated, characterize hot spots. The gain in precision over random classification is only between ∼0.7 and ∼2.1, however, depending on the hot spot class (Table S16), and, hence, more than fourfold lower than when hot spots are predicted as weak spots by CNA (Table 1). Together with the low computational demand on the order of hours only, these results lead to the strong recommendation to apply CNA-based weak spot prediction for data-driven protein engineering toward increased T_{50} or/and D.

In summary, we provide systematic and unbiased reference data at large scale for thermostability measured as T_{50} values and detergent tolerance measured as D for a biotechnologically important protein, identified consistently defined hot spot types for evaluating the performance of data-driven protein-engineering approaches, and showed that CNA-based hot spot prediction can yield a gain in precision over random classification up to ninefold.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b00954.

Tables S1−S16 and Figures S1−S2 as described in text; supplemental references (PDF)

T_{50} values (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Holger Gohlke − *Forschungszentrum Jülich GmbH, Jülich, Germany, and Heinrich Heine University Düsseldorf, Düsseldorf, Germany*; orcid.org/0000-0001-8613-1447; Email: gohlke@uni-duesseldorf.de, h.gohlke@fz-juelich.de

Other Authors

Christina Nutschel − *Forschungszentrum Jülich GmbH, Jülich, Germany*

Alexander Fulton − *Heinrich Heine University Düsseldorf, Jülich, Germany*

Olav Zimmermann − *Forschungszentrum Jülich GmbH, Jülich, Germany*

Ulrich Schwaneberg − *RWTH Aachen University, Aachen, Germany, and DWI-Leibniz-Institute for Interactive*

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

Materials, Aachen, Germany; orcid.org/0000-0003-4026-701X

Karl-Erich Jaeger − *Heinrich Heine University Düsseldorf, Jülich, Germany, and Forschungszentrum Jülich GmbH, Jülich, Germany;* orcid.org/0000-0002-6036-0708

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.9b00954

Author Contributions

H.G., K.-E.J., and U.S. conceived the study. C.N. analyzed the data, performed MD simulations and CNA computations, analyzed the results, and wrote the manuscript together with H.G. A.F. performed experimental work. O.Z. performed ProFASi simulations and analyzed the results. H.G. supervised and managed the project. All authors reviewed and approved the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

C.N. is funded through a grant ("Vernetzungsdoktorand") provided by the Forschungszentrum Jülich. Parts of the study were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through funding no. INST 208/704-1 FUGG to H.G. and INST 208/654-1 FUGG to K.-E.J. U.S. and K.-E.J. additionally received funding within the DFG research training group 1166 "Biocatalysis using Non-Conventional Media− BioNoCo". H.G. is grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich Heine University Düsseldorf. H.G. gratefully acknowledges the computing time granted by the John von Neumann Institute for Computing (NIC) and provided on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user IDs: HKF7; protil (project ID: 15956)).[132,133]

■ REFERENCES

(1) Salazar, O.; Cirino, P. C.; Arnold, F. H. Thermostabilization of a cytochrome P450 peroxygenase. *ChemBioChem* **2003**, *4*, 891−893.
(2) Radestock, S.; Gohlke, H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* **2008**, *8*, 507−522.
(3) Rader, A. Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys. Biol.* **2010**, *7*, 016002.
(4) Radestock, S.; Gohlke, H. Protein rigidity and thermophilic adaptation. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 1089−1108.
(5) Rathi, P. C.; Radestock, S.; Gohlke, H. Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* **2012**, *159*, 135−144.
(6) Dick, M.; Weiergräber, O. H.; Classen, T.; Bisterfeld, C.; Bramski, J.; Gohlke, H.; Pietruszka, J. Trading off stability against activity in extremophilic aldolases. *Sci. Rep.* **2016**, *6*, 17908.
(7) Rathi, P. C.; Jaeger, K.-E.; Gohlke, H. Structural rigidity and protein thermostability in variants of lipase A from Bacillus subtilis. *PLoS One* **2015**, *10*, No. e0130289.
(8) Rathi, P. C.; Fulton, A.; Jaeger, K.-E.; Gohlke, H. Application of rigidity theory to the thermostabilization of lipase A from Bacillus subtilis. *PLoS Comput. Biol.* **2016**, *12*, No. e1004754.
(9) Pottkämper, J.; Barthen, P.; Ilmberger, N.; Schwaneberg, U.; Schenk, A.; Schulte, M.; Ignatiev, N.; Streit, W. R. Applying metagenomics for the identification of bacterial cellulases that are stable in ionic liquids. *Green Chem.* **2009**, *11*, 957−965.
(10) Liu, H.; Zhu, L.; Bocola, M.; Chen, N.; Spiess, A. C.; Schwaneberg, U. Directed laccase evolution for improved ionic liquid resistance. *Green Chem.* **2013**, *15*, 1348−1355.
(11) Carter, J. L.; Bekhouche, M.; Noiriel, A.; Blum, L. J.; Doumèche, B. Directed evolution of a formate dehydrogenase for increased tolerance to ionic liquids reveals a new site for increasing the stability. *ChemBioChem* **2014**, *15*, 2710−2718.
(12) Chen, Z.; Pereira, J. H.; Liu, H.; Tran, H. M.; Hsu, N. S.; Dibble, D.; Singh, S.; Adams, P. D.; Sapra, R.; Hadi, M. Z.; Simmons, B. A.; Sale, K. L. Improved activity of a thermophilic cellulase, Cel5A, from Thermotoga maritima on ionic liquid pretreated switchgrass. *PLoS One* **2013**, *8*, No. e79725.
(13) Lehmann, C.; Bocola, M.; Streit, W. R.; Martinez, R.; Schwaneberg, U. Ionic liquid and deep eutectic solvent-activated CelA2 variants generated by directed evolution. *Appl. Microbiol. Biotechnol.* **2014**, *98*, 5775−5785.
(14) Nordwald, E. M.; Armstrong, G. S.; Kaar, J. L. NMR-guided rational engineering of an ionic-liquid-tolerant lipase. *ACS Catal.* **2014**, *4*, 4057−4064.
(15) Frauenkron-Machedjou, V. J.; Fulton, A.; Zhu, L.; Anker, C.; Bocola, M.; Jaeger, K. E.; Schwaneberg, U. Towards understanding directed evolution: more than half of all amino acid positions contribute to ionic liquid resistance of Bacillus subtilis lipase A. *ChemBioChem* **2015**, *16*, 937−945.
(16) Zhao, J.; Frauenkron-Machedjou, V. J.; Fulton, A.; Zhu, L.; Davari, M. D.; Jaeger, K.-E.; Schwaneberg, U.; Bocola, M. Unraveling the effects of amino acid substitutions enhancing lipase resistance to an ionic liquid: a molecular dynamics study. *Phys. Chem. Chem. Phys.* **2018**, *20*, 9600−9609.
(17) Brissos, V.; Eggert, T.; Cabral, J.; Jaeger, K.-E. Improving activity and stability of cutinase towards the anionic detergent AOT by complete saturation mutagenesis. *Protein Eng., Des. Sel.* **2008**, *21*, 387−393.
(18) Akbulut, N.; Öztürk, M. T.; Pijning, T.; Öztürk, S. İ.; Gümüşel, F. Improved activity and thermostability of Bacillus pumilus lipase by directed evolution. *J. Biotechnol.* **2013**, *164*, 123−129.
(19) Fulton, A.; Frauenkron-Machedjou, V. J.; Skoczinski, P.; Wilhelm, S.; Zhu, L.; Schwaneberg, U.; Jaeger, K. E. Exploring the protein stability landscape: Bacillus subtilis lipase A as a model for detergent tolerance. *ChemBioChem* **2015**, *16*, 930−936.
(20) Rigoldi, F.; Donini, S.; Redaelli, A.; Parisini, E.; Gautieri, A. Engineering of thermostable enzymes for industrial applications. *APL Bioeng* **2018**, *2*, 011501.
(21) Littlechild, J. A. Enzymes from extreme environments and their industrial applications. *Front. Bioeng. Biotechnol.* **2015**, *3*, 161.
(22) https://www.bccresearch.com/market-research/biotechnology/global-markets-for-enzymes-in-industrial-applications-bio030k.html.
(23) Lehmann, M.; Wyss, M. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr. Opin. Biotechnol.* **2001**, *12*, 371−375.
(24) Eijsink, V. G.; Bjørk, A.; Gåseidnes, S.; Sirevåg, R.; Synstad, B.; van den Burg, B.; Vriend, G. Rational engineering of enzyme stability. *J. Biotechnol.* **2004**, *113*, 105−120.
(25) Thiltgen, G.; Goldstein, R. A. Assessing predictors of changes in protein stability upon mutation using self-consistency. *PLoS One* **2012**, *7*, No. e46084.
(26) Johnston, M. A.; Søndergaard, C. R.; Nielsen, J. E. Integrated prediction of the effect of mutations on multiple protein characteristics. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 165−178.
(27) Li, Y.; Zhang, J.; Tai, D.; Russell Middaugh, C.; Zhang, Y.; Fang, J. Prots: A fragment based protein thermo-stability potential. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 81−92.
(28) Capriotti, E.; Fariselli, P.; Rossi, I.; Casadio, R. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinf.* **2008**, *9*, S6.
(29) Potapov, V.; Cohen, M.; Schreiber, G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng., Des. Sel.* **2009**, *22*, 553−560.

(30) Khan, S.; Vihinen, M. Performance of protein stability predictors. *Hum. Mutat.* **2010**, *31*, 675−684.
(31) Usmanova, D. R.; Bogatyreva, N. S.; Ariño Bernad, J.; Eremina, A. A.; Gorshkova, A. A.; Kanevskiy, G. M.; Lonishin, L. R.; Meister, A. V.; Yakupova, A. G.; Kondrashov, F. A.; Ivankov, D. N. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics* **2018**, *34*, 3653−3658.
(32) Pucci, F.; Bernaerts, K. V.; Kwasigroch, J. M.; Rooman, M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* **2018**, *34*, 3659−3665.
(33) Modarres, H. P.; Mofrad, M.; Sanati-Nezhad, A. Protein thermostability engineering. *RSC Adv.* **2016**, *6*, 115252−115270.
(34) Bava, K. A.; Gromiha, M. M.; Uedaira, H.; Kitajima, K.; Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **2004**, *32*, D120−D121.
(35) Kumar, M. S.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. ProTherm and ProNIT: thermodynamic databases for proteins and protein−nucleic acid interactions. *Nucleic Acids Res.* **2006**, *34*, D204−D206.
(36) Zhang, Z.; Wang, L.; Gao, Y.; Zhang, J.; Zhenirovskyy, M.; Alexov, E. Predicting folding free energy changes upon single point mutations. *Bioinformatics* **2012**, *28*, 664−671.
(37) Kang, S.; Chen, G.; Xiao, G. Robust prediction of mutation-induced protein stability change by property encoding of amino acids. *Protein Eng., Des. Sel.* **2008**, *22*, 75−83.
(38) Nisthal, A.; Wang, C. Y.; Ary, M. L.; Mayo, S. L. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 16367−16377.
(39) Zeymer, C.; Hilvert, D. Directed evolution of protein catalysts. *Annu. Rev. Biochem.* **2018**, *87*, 131−157.
(40) Arnold, F. H. Directed evolution: bringing new chemistry to life. *Angew. Chem., Int. Ed.* **2018**, *57*, 4143−4148.
(41) Romero, P. A.; Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866.
(42) Chaparro-Riggers, J. F.; Polizzi, K. M.; Bommarius, A. S. Better library design: data-driven protein engineering. *Biotechnol. J.* **2007**, *2*, 180−191.
(43) Wijma, H. J.; Floor, R. J.; Janssen, D. B. Structure-and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.* **2013**, *23*, 588−594.
(44) Reetz, M. T.; Carballeira, J. D.; Vogel, A. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem., Int. Ed.* **2006**, *45*, 7745−7751.
(45) Huang, X.; Gao, D.; Zhan, C.-G. Computational design of a thermostable mutant of cocaine esterase via molecular dynamics simulations. *Org. Biomol. Chem.* **2011**, *9*, 4138−4143.
(46) Badieyan, S.; Bevan, D. R.; Zhang, C. Study and design of stability in GH5 cellulases. *Biotechnol. Bioeng.* **2012**, *109*, 31−44.
(47) Brooks, B. R.; Brooks III, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(48) Craig, D. B.; Dombkowski, A. A. Disulfide by Design 2.0: a web-based tool for disulfide engineering in proteins. *BMC Bioinf.* **2013**, *14*, 346.
(49) Singh, J.; Ator, M. A.; Jaeger, E. P.; Allen, M. P.; Whipple, D. A.; Soloweij, J. E.; Chowdhary, S.; Treasurywala, A. M. Application of genetic algorithms to combinatorial synthesis: A computational approach to lead identification and lead optimization. *J. Am. Chem. Soc.* **1996**, *118*, 1669−1676.
(50) Moore, K. W.; Pechen, A.; Feng, X.-J.; Dominy, J.; Beltrani, V. J.; Rabitz, H. Why is chemical synthesis and property optimization easier than expected? *Phys. Chem. Chem. Phys.* **2011**, *13*, 10048−10070.
(51) Kawasaki, K.; Kondo, H.; Suzuki, M.; Ohgiya, S.; Tsuda, S. Alternate conformations observed in catalytic serine of Bacillus subtilis lipase determined at 1.3 Å resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 1168−1174.
(52) Skoczinski, P.; Volkenborn, K.; Fulton, A.; Bhadauriya, A.; Nutschel, C.; Gohlke, H.; Knapp, A.; Jaeger, K.-E. Contribution of single amino acid and codon substitutions to the production and secretion of a lipase by Bacillus subtilis. *Microb. Cell Fact.* **2017**, *16*, 160.
(53) Schallmey, M.; Singh, A.; Ward, O. P. Developments in the use of Bacillus species for industrial production. *Can. J. Microbiol.* **2004**, *50*, 1−17.
(54) Van Pouderoyen, G.; Eggert, T.; Jaeger, K.-E.; Dijkstra, B. W. The crystal structure of Bacillus subtilis lipase: a minimal α/β hydrolase fold enzyme. *J. Mol. Biol.* **2001**, *309*, 215−226.
(55) Jech, T. *Set Theory*; Springer Science & Business Media: Berlin/Heidelberg, 2013.
(56) Hamers, L.; Hemeryck, Y.; Herweyers, G.; Janssen, M.; Keters, H.; Rousseau, R.; Vanhoutte, A. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Inf. Process. Manage.* **1989**, *25*, 315−318.
(57) Leydesdorff, L. On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 77−85.
(58) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(59) Creighton, T. E. *Proteins: Structures and Molecular Properties*; Freeman: New York, 1992.
(60) Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; et al. The Pfam protein families database. *Nucleic Acids Res.* **2012**, *40*, D290−D301.
(61) Sievers, F.; Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. In *Multiple sequence alignment methods*; Springer: New York, 2014; pp 105−116.
(62) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.
(63) Dartois, V.; Baulard, A.; Schanck, K.; Colson, C. Cloning, nucleotide sequence and expression in Escherichia coli of a lipase gene from Bacillus subtilis 168. *Biochim. Biophys. Acta, Gene Struct. Expression* **1992**, *1131*, 253−260.
(64) Golicz, A.; Troshin, P. V.; Madeira, F.; Martin, D. M. A.; Procter, J. B.; Barton, G. J. AACon: A Fast Amino Acid Conservation Calculation Service. Submitted, **2018**.
(65) Waterhouse, A. M.; Procter, J. B.; Martin, D. M.; Clamp, M.; Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009**, *25*, 1189−1191.
(66) Krüger, D. M.; Rathi, P. C.; Pfleger, C.; Gohlke, H. CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure,(thermo-) stability, and function. *Nucleic Acids Res.* **2013**, *41*, W340−W348.
(67) Pfleger, C.; Rathi, P. C.; Klein, D. L.; Radestock, S.; Gohlke, H. Constraint Network Analysis (CNA): a Python software package for efficiently linking biomacromolecular structure, flexibility,(thermo-) stability, and function. *J. Chem. Inf. Model.* **2013**, *53*, 1007−15.
(68) Hermans, S. M.; Pfleger, C.; Nutschel, C.; Hanke, C. A.; Gohlke, H. Rigidity theory for biomolecules: concepts, software, and applications. *Wiley Interdiscip Rev. Comput. Mol. Sci.* **2017**, *7*, No. e1311.
(69) Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 150−165.
(70) Hespenheide, B.; Jacobs, D.; Thorpe, M. Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J. Phys.: Condens. Matter* **2004**, *16*, S5055.
(71) Jacobs, D. J.; Thorpe, M. F. Generic rigidity percolation: the pebble game. *Phys. Rev. Lett.* **1995**, *75*, 4051.
(72) Jacobs, D. J. Generic rigidity in three-dimensional bond-bending networks. *J. Phys. A: Math. Gen.* **1998**, *31*, 6653.
(73) Dahiyat, B. I.; Benjamin Gordon, D.; Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6*, 1333−1337.

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

(74) Folch, B.; Rooman, M.; Dehouck, Y. Thermostability of salt bridges versus hydrophobic interactions in proteins probed by statistical potentials. *J. Chem. Inf. Model.* **2008**, *48*, 119−127.
(75) Privalov, P. L.; Gill, S. J. Stability of protein structure and hydrophobic interaction. In *Advances in Protein Chemistry*; Elsevier: Amsterdam, Netherlands, 1988; Vol. *39*, pp 191−234.
(76) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878−3888.
(77) Case, D. A.; Babin, V.; Berryman, J.; Betz, R.; Cai, Q.; Cerutti, D.; Cheatham, T. E., III; Darden, T.; Duke, R.; Gohlke, H. *Amber 14*; 2014.
(78) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(79) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735−1747.
(80) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(81) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(82) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(83) Pfleger, C.; Gohlke, H. Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. *Structure* **2013**, *21*, 1725−1734.
(84) Rathi, P. C.; Mulnaes, D.; Gohlke, H. VisualCNA: a GUI for interactive constraint network analysis and protein engineering for improving thermostability. *Bioinformatics* **2015**, *31*, 2394−2396.
(85) Pfleger, C.; Radestock, S.; Schmidt, E.; Gohlke, H. Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* **2013**, *34*, 220−33.
(86) Buckland, M.; Gey, F. The relationship between recall and precision. *J. Am. Soc. Inf. Sci.* **1994**, *45*, 12−19.
(87) Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval*, 2005; Springer: Amsterdam, Netherlands, 2005; pp 345−359.
(88) Irbäck, A.; Mohanty, S. PROFASI: a Monte Carlo simulation package for protein folding and aggregation. *J. Comput. Chem.* **2006**, *27*, 1548−1555.
(89) Mohanty, S.; Meinke, J. H.; Zimmermann, O. Folding of Top7 in unbiased all-atom Monte Carlo simulations. *Proteins: Struct., Funct., Genet.* **2013**, *81*, 1446−1456.
(90) Irbäck, A.; Mitternacht, S.; Mohanty, S. An effective all-atom potential for proteins. *PMC Biophys.* **2009**, *2*, 2.
(91) Irbäck, A.; Samuelsson, B.; Sjunnesson, F.; Wallin, S. Thermodynamics of α-and β-structure formation in proteins. *Biophys. J.* **2003**, *85*, 1466−1473.
(92) Favrin, G.; Irbäck, A.; Sjunnesson, F. Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. *J. Chem. Phys.* **2001**, *114*, 8154−8158.
(93) Neugebauer, J. *Guide to the Properties and Uses of Detergents in Biology and Biochemistry*; Calbiochem-Novabiochem International: San Diego, CA, 1987.
(94) Villegas, V.; Viguera, A. R.; Avilés, F. X.; Serrano, L. Stabilization of proteins by rational design of α-helix stability using helix/coil transition theory. *Folding Des.* **1996**, *1*, 29−34.
(95) Taddei, N.; Chiti, F.; Fiaschi, T.; Bucciantini, M.; Capanni, C.; Stefani, M.; Serrano, L.; Dobson, C. M.; Ramponi, G. Stabilisation of α-helices by site-directed mutagenesis reveals the importance of secondary structure in the transition state for acylphosphatase folding. *J. Mol. Biol.* **2000**, *300*, 633−647.
(96) Yu, H.; Yan, Y.; Zhang, C.; Dalby, P. A. Two strategies to engineer flexible loops for improved enzyme thermostability. *Sci. Rep.* **2017**, *7*, 41212.
(97) Chen, H.; Zhou, H.-X. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res.* **2005**, *33*, 3193−3199.
(98) Yokot, K.; Satou, K.; Ohki, S.-y. Comparative analysis of protein thermostability: Differences in amino acid content and substitution at the surfaces and in the core regions of thermophilic and mesophilic proteins. *Sci. Technol. Adv. Mater.* **2006**, *7*, 255.
(99) Watanabe, K.; Masuda, T.; Ohashi, H.; Mihara, H.; Suzuki, Y. Multiple Proline Substitutions Cumulatively Thermostabilize Bacillus Cereus ATCC7064 Oligo-1, 6-Glucosidase: Irrefragable Proof Supporting the Proline Rule. *Eur. J. Biochem.* **1994**, *226*, 277−283.
(100) Kumar, S.; Tsai, C.-J.; Nussinov, R. Factors enhancing protein thermostability. *Protein Eng., Des. Sel.* **2000**, *13*, 179−191.
(101) Ikai, A. Thermostability and aliphatic index of globular proteins. *J. Biochem.* **1980**, *88*, 1895−1898.
(102) Ashenberg, O.; Gong, L. I.; Bloom, J. D. Mutational effects on stability are largely conserved during protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 21071−21076.
(103) Greene, L. H.; Chrysina, E. D.; Irons, L. I.; Papageorgiou, A. C.; Acharya, K. R.; Brew, K. Role of conserved residues in structure and stability: Tryptophans of human serum retinol-binding protein, a model for the lipocalin superfamily. *Protein Sci.* **2001**, *10*, 2301−2316.
(104) Hafizah, N. F.; Teh, A.-H.; Furusawa, G. Biochemical Characterization of Thermostable and Detergent-Tolerant β-Agarase, PdAgaC, from Persicobacter sp. CCB-QB2. *Appl. Biochem. Biotechnol.* **2019**, *187*, 770−781.
(105) Lu, M.; Dukunde, A.; Daniel, R. Biochemical profiles of two thermostable and organic solvent−tolerant esterases derived from a compost metagenome. *Appl. Microbiol. Biotechnol.* **2019**, *103*, 3421−3437.
(106) Annamalai, N.; Rajeswari, M. V.; Balasubramanian, T. Extraction, purification and application of thermostable and halostable alkaline protease from Bacillus alveayuensis CAS 5 using marine wastes. *Food Bioprod. Process.* **2014**, *92*, 335−342.
(107) Araya, C. L.; Fowler, D. M.; Chen, W.; Muniez, I.; Kelly, J. W.; Fields, S. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 16858−16863.
(108) Foit, L.; Morgan, G. J.; Kern, M. J.; Steimer, L. R.; von Hacht, A. A.; Titchmarsh, J.; Warriner, S. L.; Radford, S. E.; Bardwell, J. C. Optimizing protein stability in vivo. *Mol. Cell* **2009**, *36*, 861−871.
(109) Deng, Z.; Huang, W.; Bakkalbasi, E.; Brown, N. G.; Adamski, C. J.; Rice, K.; Muzny, D.; Gibbs, R. A.; Palzkill, T. Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *J. Mol. Biol.* **2012**, *424*, 150−167.
(110) Klesmith, J. R.; Bacik, J.-P.; Wrenbeck, E. E.; Michalczyk, R.; Whitehead, T. A. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 2265−2270.
(111) Buß, O.; Rudat, J.; Ochsenreither, K. FoldX as protein engineering tool: better than random based approaches? *Comput. Struct. Biotechnol. J.* **2018**, *16*, 25−33.
(112) Pack, S. P.; Kang, T. J.; Yoo, Y. J. Protein thermostabilizing factors: high relative occurrence of amino acids, residual properties, and secondary structure type in different residual state. *Appl. Biochem. Biotechnol.* **2013**, *171*, 1212−1226.
(113) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **2002**, *320*, 369−387.
(114) Polizzi, K. M.; Chaparro-Riggers, J. F.; Vazquez-Figueroa, E.; Bommarius, A. S. Structure-guided consensus approach to create a more thermostable penicillin G acylase. *Biotechnol. J.* **2006**, *1*, 531−536.
(115) Porebski, B. T.; Buckle, A. M. Consensus protein design. *Protein Eng., Des. Sel.* **2016**, *29*, 245−251.

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

(116) Vogt, G.; Woell, S.; Argos, P. Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **1997**, *269*, 631−643.
(117) Chakravarty, S.; Varadarajan, R. Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Lett.* **2000**, *470*, 65−69.
(118) Perl, D.; Mueller, U.; Heinemann, U.; Schmid, F. X. Two exposed amino acid residues confer thermostability on a cold shock protein. *Nat. Struct. Biol.* **2000**, *7*, 380.
(119) Pace, C. N. Single surface stabilizer. *Nat. Struct. Biol.* **2000**, *7*, 345.
(120) Pokkuluri, P.; Raffen, R.; Dieckman, L.; Boogaard, C.; Stevens, F.; Schiffer, M. Increasing protein stability by polar surface residues: domain-wide consequences of interactions within a loop. *Biophys. J.* **2002**, *82*, 391−398.
(121) Jung, S.; Plückthun, A. Improving in vivo folding and stability of a single-chain Fv antibody fragment by loop grafting. *Protein Eng., Des. Sel.* **1997**, *10*, 959−966.
(122) Pantoliano, M. W.; Whitlow, M.; Wood, J. F.; Dodd, S. W.; Hardman, K. D.; Rollence, M. L.; Bryan, P. N. Large increases in general stability for subtilisin BPN'through incremental changes in the free energy of unfolding. *Biochemistry* **1989**, *28*, 7205−7213.
(123) Worth, C. L.; Preissner, R.; Blundell, T. L. SDM−a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.* **2011**, *39*, W215−22.
(124) Pokala, N.; Handel, T. M. Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol.* **2005**, *347*, 203−27.
(125) Capriotti, E.; Fariselli, P.; Casadio, R. I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **2005**, *33*, W306−W310.
(126) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. Protein structure prediction using Rosetta. In *Methods in Enzymology*; Elsevier: Amsterdam, Netherlands, 2004; Vol. *383*, pp 66−93.
(127) Potapov, V.; Cohen, M.; Inbar, Y.; Schreiber, G. Protein structure modelling and evaluation based on a 4-distance description of side-chain interactions. *BMC Bioinf.* **2010**, *11*, 374.
(128) Benedix, A.; Becker, C. M.; de Groot, B. L.; Caflisch, A.; Böckmann, R. A. Predicting free energy changes using structural ensembles. *Nat. Methods* **2009**, *6*, 3.
(129) Bordner, A.; Abagyan, R. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins: Struct., Funct., Genet.* **2004**, *57*, 400−413.
(130) Beermann, B.; Guddorf, J.; Boehm, K.; Albers, A.; Kolkenbrock, S.; Fetzner, S.; Hinz, H.-J. Stability, Unfolding, and Structural Changes of Cofactor-Free 1 H-3-Hydroxy-4-oxoquinaldine 2, 4-Dioxygenase. *Biochemistry* **2007**, *46*, 4241−4249.
(131) Hung, H.-C.; Chang, G.-G. Multiple unfolding intermediates of human placental alkaline phosphatase in equilibrium urea denaturation. *Biophys. J.* **2001**, *81*, 3456−3471.
(132) Krause, D. JUWELS: Modular Tier-0/1 Supercomputer at the Jülich Supercomputing Centre. *JLSRF* **2019**, *5*, A135.
(133) Krause, D.; Thörnig, P. JURECA: Modular Supercomputer at Jülich Supercomputing Centre. *Journal of Large-scale Research Facilities JLSRF* **2018**, *4*, 132.

https://dx.doi.org/10.1021/acs.jcim.9b00954
J. Chem. Inf. Model. 2020, 60, 1568−1584

ORIGINAL PUBLICATION II-SUPPORTING INFORMATION

Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for *Bacillus subtilis* lipase A

Nutschel, C., Fulton, A., Zimmermann, O., Schwaneberg, U., Jaeger, K.-E., Gohlke, H.

J Chem Inf Model. 2020, 60, 3, 1568-1584.

https://pubs.acs.org/doi/10.1021/acs.jcim.9b00954

Supporting Information

Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for *Bacillus subtilis* lipase A

Christina Nutschel[1,2], Alexander Fulton[3], Olav Zimmermann[2], Ulrich Schwaneberg[5,6], Karl-Erich Jaeger[3,4], Holger Gohlke[1,2,7,*]

[1] John von Neumann Institute for Computing (NIC) and Institute for Complex Systems - Structural Biochemistry (ICS-6), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[2] Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[3] Institute of Molecular Enzyme Technology, Heinrich Heine University Düsseldorf, 52425 Jülich, Germany

[4] Institute of Bio- and Geosciences IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[5] Institute of Biotechnology, RWTH Aachen University, 52074 Aachen, Germany

[6] DWI-Leibniz-Institute for Interactive Materials, 52056 Aachen, Germany

[7] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

*Corresponding author:
John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC), and Institute for Complex Systems - Structural Biochemistry (ICS-6)
Forschungszentrum Jülich GmbH
Wilhelm-Johnen-Straße
52425 Jülich
Germany
Email: gohlke@uni-duesseldorf.de or h.gohlke@fz-juelich.de

Table of Contents

Supplemental Tables

Table S1: Variants with significantly changed T_{50} or D towards one detergent by *random mutagenesis*.

Type of protein stability	No. of variants with $\Delta T_{50} > 0$ K or $\Delta D > 0$[a]	No. of variants with $\Delta T_{50} < 0$ K or $\Delta D < 0$[a]	Total no. of variants	Concentration [mM][b]	σ_D [mM][c]
T_{50}	214 (11.5)	1642 (88.5)	1856	/	/
D_{SDS}	261 (14.6)	1532 (85.4)	1793	0.35	0.08
D_{CTAB}	87 (10.3)	760 (89.7)	847	0.27	0.09
$D_{SB3\text{-}16}$	103 (22.2)	361 (77.8)	464	0.77	0.32
$D_{Tween\ 80}$	52 (10.5)	443 (89.5)	495	0.08	0.14
Mean (D)	126 (14.4)	774 (85.6)	900	/	/

[a] Number of variants; values in brackets represent the likelihood [%] to find variants with significantly changed T_{50} or D in relation to the total number of variants, respectively.

[b] Used detergent concentration; CMC values according to published data: SDS (7 mM); CTAB (1 mM); SB3-16 (0.01 mM); Tween 80 (0.012 mM)[1].

[c] Standard deviations of 2997 wt*Bs*LipA replicates for each concentration [2].

Table S2: Identified classes of substitution sites.

Class[a]	Definition	No. of substitution sites	*Random classification*[b]	CNA				
				No. of *weak spots*[c]	r[d]	p[e]	*gip*[f]	*F1*[g]
I	{Substitution site$_x$ | $1 \le x \le 181$, $T_{50}(x)$ is significantly increased}	69	0.38	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
II	{Substitution site$_x$ | $1 \le x \le 181$, $D_{SDS}(x)$ is significantly increased}	74	0.41	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
III	{Substitution site$_x$ | $1 \le x \le 181$, $D_{CTAB}(x)$ is significantly increased}	42	0.23	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
IV	{Substitution site$_x$ | $1 \le x \le 181$, $D_{SB3\text{-}16}(x)$ is significantly increased}	46	0.25	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
V	{Substitution site$_x$ | $1 \le x \le 181$, $D_{Tween\ 80}(x)$ is significantly increased}	34	0.19	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
VI	**II** ∪ **III** ∪ **IV** ∪ **V**	109	0.60	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
VII	**I** ∪ **VI**	124	0.69	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
VIII	**II** ∩ **III** ∩ **IV** ∩ **V**	11	0.06	2	0.18	0.20	3.30	0.19
IX	**I** ∩ **VIII**	7	0.04	2	0.29	0.20	5.17	0.24
X	{Substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $T_{50}(x)$}	6	0.03	1	0.17	0.10	3.02	0.13
XI	{Substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $D_{SDS}(x)$}	6	0.03	1	0.17	0.10	3.02	0.13
XII	{Substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $D_{CTAB}(x)$}	6	0.03	3	0.50	0.30	9.05	0.38
XIII	{Substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $D_{SB3\text{-}16}(x)$}	6	0.03	2	0.33	0.20	6.03	0.25
XIV	{Substitution site$_x$ | $1 \le x \le 181$, six highest effects in significantly increased $D_{Tween\ 80}(x)$}	6	0.03	0	/	/	/	/

Table S2 continued.

Class[a]	Definition	No. of substitution sites	*Random classification*[b]	CNA				
				No. of *weak spots*[c]	*r*[d]	*p*[e]	*gip*[f]	*F1*[g]
XV	**XI** ∪ **XII** ∪ **XIII** ∪ **XIV**	20	0.11	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
XVI	**X** ∪ **XV**	24	0.13	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
XVII	**XI** ∩ **XII** ∩ **XIII** ∩ **XIV**	0	0	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]
XVIII	**X** ∩ **XVII**	0	0	nd[h]	nd[h]	nd[h]	nd[h]	nd[h]

[a] Class of substitution sites; underlined classes represent *hot spots*.

[b] Likelihood for randomly choosing substitution sites / *hot spots* of the respective class.

[c] Numbers of *hot spots* predicted as *weak spots*.

[d] Recall **(Eq. 8 in the main text)**.

[e] Precision **(Eq. 9 in the main text)**.

[f] Gain in precision over *random classification* **(Eq. 11 in the main text)**.

[g] F1-score **(Eq. 12 in the main text)**.

[h] Not determined.

Table S3: Correlations between the frequency of substitution occurrences per substitution site of classes **I** – **V**, where variants yield significantly increased T_{50} ($N_{BsLipA;\ T}$) or D ($N_{BsLipA;\ D}$) towards one detergent.

		Class of substitution site[a]				
		I	**II**	**III**	**IV**	**V**
Class of substitution site[a]	**I**	/	< 0.001	< 0.05	> 0.1	< 0.1
	II	0.066	/	< 0.001	< 0.05	< 0.05
	III	0.033	0.176	/	< 0.001	< 0.001
	IV	0.004	0.036	0.263	/	< 0.001
	V	0.015	0.031	0.105	0.059	/

[a] Upper values are *p*-values; lower values are R^2-values.

Table S4: Correlations between the highest effects per substitution site of classes **I** – **V**, considering variants with significantly increased T_{50} ($\Delta T_{50;\ max}$) or D (ΔD_{max}) towards one detergent.

		Class of substitution site[a]				
		I	**II**	**III**	**IV**	**V**
Class of substitution site[a]	**I**	/	< 0.1	> 0.1	< 0.1	> 0.1
	II	0.089	/	> 0.1	> 0.1	< 0.05
	III	0.002	0.041	/	< 0.1	> 0.1
	IV	0.132	0.064	0.128	/	> 0.1
	V	$4 \square 10^{-5}$	0.235	0.007	0.029	/

[a] Upper values are *p*-values; lower values are R^2-values.

Table S5: Correlations between $N_{BsLipA;\ T}$ and $\Delta T_{50;\ max}$ as well as $N_{BsLipA;\ D}$ and ΔD_{max} of substitution sites of classes **I** – **V**.

Class of substitution sites	R^2**-value**	***p*-value**
I	0.449	< 0.001
II	0.382	< 0.001
III	0.054	> 0.1
IV	0.464	< 0.001
V	0.008	> 0.1

Table S6: Jaccard indices (*J*) for substitution sites of classes **I** – **V** or **X** – **XIV**.

		Class of substitution site[a]				
		I / X	**II / XI**	**III / XII**	**IV / XIII**	**V / XIV**
Class of substitution site[a]	**I / X**	/	0.091	0.091	0.091	0
	II / XI	0.416	/	0	0	0
	III / XII	0.291	0.333	/	0.200	0
	IV / XIII	0.264	0.250	0.467	/	0
	V / XIV	0.157	0.227	0.310	0.250	/

[a] Lower values are *J* of substitution sites for which variants yield significantly increased T_{50} and *D* towards one detergent as well as significantly increased *D* towards two detergents (**I** – **V**); upper values are the *J* of the six substitution sites for which variants yield the respective highest effects regarding significantly increased T_{50} and *D* towards one detergent as well as significantly increased *D* towards two detergents (**X** – **XIV**).

Table S7: Variants with significantly changed *D* at substitution sites of class **VIII**.

Type of protein stability	No. of variants with $\Delta D > 0$[a]	No. of variants with $\Delta D < 0$[a]	Total no. of variants
D_{SDS}	63 (54.8)	52 (45.2)	115
D_{CTAB}	19 (76.0)	6 (24.0)	25
$D_{SB3\text{-}16}$	31 (79.5)	8 (20.5)	39
$D_{Tween\ 80}$	16 (84.2)	3 (15.8)	19
Mean (*D*)	32 (64.0)	17 (36.0)	50

[a] Number of variants; values in brackets represent the likelihood [%] to find variants with significantly changed *D* in relation to the total number of variants, respectively.

Table S8: Variants with significantly changed T_{50} or D at substitution sites of class **IX**.

Type of protein stability	No. of variants with $\Delta T_{50} > 0$ or $\Delta D > 0$[a]	No. of variants with $\Delta T_{50} < 0$ or $\Delta D < 0$[a]	Total no. of variants
T_{50}	35 (40.7)	51 (59.3)	86
D_{SDS}	39 (56.5)	30 (43.5)	69
D_{CTAB}	10 (79.9)	3 (23.1)	13
$D_{SB3\text{-}16}$	18 (78.3)	5 (21.7)	23
$D_{Tween\ 80}$	10 (83.3)	2 (16.7)	12
Mean (D)	19 (65.5)	10 (34.5)	29

[a] Number of variants; values in brackets represent the likelihood [%] to find variants with significantly changed T_{50} or D in relation to the total number of variants, respectively.

Table S9: Substitution sites of class **X** – **XIV** with the respective six highest effects in significantly increased T_{50} or D towards each detergent.

Class of substitution site	Substitution site[a]	$\Delta T_{50;\,max}$ [K] or ΔD_{max}	$N_{BsLipA;\,T}$ or $N_{BsLipA;\,D}$	Location[b]
X	M137	7.7	10	Loop
	M134	5.6	9	Bend
	G155	4.5	4	Loop
	F17	3.8	12	G1
	I157	3.6	1	G5
	Y139	3.2	7	αE
XI	M137	1.49	6	Loop
	R142	1.45	7	Loop
	T47	1.29	8	Loop
	E65	1.10	14	αB
	G13	1.03	6	Turn
	Y49	0.94	9	αB
XII	T110	1.63	1	Loop
	K44	1.04	4	Turn
	I135	1.01	6	Bend
	G13	0.72	1	Turn
	M134	0.58	4	Bend
	N51	0.55	1	αB
XIII	**G46**	2.41	11	Loop
	K44	2.26	3	Turn
	M134	2.25	1	Bend
	N51	2.10	7	αB
	T45	1.90	4	Turn
	V99	1.87	5	β6
XIV	S127	2.29	3	β7
	I128	2.00	1	β7
	T126	1.98	1	β7
	L123	1.89	1	β7
	Q150	1.07	2	β8
	A20	0.86	1	αA

[a] Substitution sites highlighted in bold are predicted as *weak spots* by CNA.

[b] Location of the substitution site in terms of secondary structure elements.

Table S10: Variants with significantly changed T_{50} or D at substitution sites of classes **X** – **XIV**.

Type of protein stability	No. of variants with $\Delta T_{50} > 0$ or $\Delta D > 0$[a]	No. of variants with $\Delta T_{50} < 0$ or $\Delta D < 0$[a]	Total no. of variants
T_{50}	43 (63.2)	25 (36.8)	68
D_{SDS}	50 (70.4)	21 (29.6)	71
D_{CTAB}	17 (53.1)	15 (46.9)	32
$D_{SB3\text{-}16}$	31 (79.5)	8 (20.5)	39
$D_{Tween\ 80}$	9 (32.1)	19 (67.9)	28
Mean (D)	27 (62.8)	16 (37.2)	43

[a] Number of variants; values in brackets represent the likelihood [%] to find variants with significantly increased T_{50} or D in relation to the total number of variants, respectively.

Table S11: Distribution of *hot spots* regarding secondary structure elements.

Secondary structure element	No of substitution sites	Class of *hot spots*																																					
		X					XI					XII					XIII					XIV					VIII					IX									
		#	*r*[a]	*p*[b]	*F1*[c]	*gip*[d]	*#*	*r*	*p*	*F1*	*gip*	*#*	*r*	*p*	*F1*	*gip*	*#*	*r*	*p*	*F1*	*gip*	*#*	*r*	*p*	*F1*	*gip*	*#*	*r*	*p*	*F1*	*gip*	*#*	*r*	*p*	*F1*	*gip*				
αA	**9**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	**1**	0.17	0.11	0.13	3.35	/	/	/	/	/	/	/	/	/				
αB	**19**	/	/	/	/	/	2	0.33	0.11	0.16	3.18	**1**	0.17	0.05	0.08	1.59	1	0.17	0.05	0.08	1.59	/	/	/	/	/	4	0.36	0.21	0.27	3.46	2	0.29	0.11	0.15	2.72				
αC	**12**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
αE	**4**	**1**	0.17	0.25	0.29	7.54	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
αF	**11**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
α(all)	**55**	**1**	**0.17**	**0.02**	**0.03**	**0.55**	**2**	**0.33**	**0.04**	**0.07**	**1.10**	**1**	**0.10**	**0.02**	**0.03**	**0.55**	**1**	**0.10**	**0.02**	**0.03**	**0.55**	**1**	**0.10**	**0.02**	**0.03**	**0.55**	4	**0.36**	**0.07**	**0.12**	**1.20**	**2**	**0.29**	**0.04**	**0.06**	**0.94**				
G1	**4**	**1**	0.17	0.25	0.29	7.54	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
G2	**3**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
G3	**3**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
G4	**3**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
G5	**5**	**1**	0.17	0.20	0.18	6.03	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
G(all)	**18**	**2**	**0.33**	**0.11**	**0.17**	**3.35**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
β3	**4**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
β4	**3**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
β5	**6**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
β6	**7**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	1	0.17	0.14	0.15	4.31	/	/	/	/	/	1	0.09	0.14	0.11	2.35	/	/	/	/	/				
β7	**7**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	4	0.67	0.57	0.62	17.24	/	/	/	/	/	/	/	/	/				
β8	**5**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	**1**	0.17	0.20	0.18	6.03	/	/	/	/	/	/	/	/	/				
β(all)	**32**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	**1**	**0.10**	**0.03**	**0.05**	**0.94**	**5**	**0.83**	0.16	**0.26**	4.71	**1**	**0.09**	**0.03**	**0.05**	**0.51**	/	/	/	/	/				
Bridge	**3**	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/				
Turn	**13**	/	/	/	/	/	1	0.17	0.03	0.11	2.32	2	0.33	0.15	0.21	4.64	2	0.33	0.15	0.21	4.64	/	/	/	/	/	2	0.18	0.15	0.17	2.53	2	0.29	0.15	0.20	3.98				
Loop	**47**	2	0.33	0.04	0.08	1.28	3	0.50	0.06	0.11	1.93	**1**	0.17	0.02	0.04	0.64	1	0.17	0.02	0.04	0.64	/	/	/	/	/	3	0.27	0.05	0.10	1.05	2	0.29	0.04	0.07	1.10				
Bend	**13**	**1**	0.17	0.08	0.11	2.32	/	/	/	/	/	2	0.33	0.15	0.21	4.64	1	0.17	0.08	0.11	2.32	/	/	/	/	/	1	0.09	0.08	0.08	1.27	1	0.14	0.08	0.10	1.59				
Rest(all)	**76**	**3**	**0.50**	**0.04**	**0.07**	**1.19**	**4**	**0.67**	**0.05**	**0.10**	**1.59**	**5**	**0.83**	**0.07**	**0.12**	**1.98**	**4**	**0.67**	**0.05**	**0.10**	**1.59**	/	/	/	/	/	**6**	**0.55**	**0.08**	**0.14**	**1.30**	**5**	**0.71**	**0.07**	**0.12**	**1.70**				

[a] Recall **(Eq. 8 in the main text)**.

[b] Precision **(Eq. 9 in the main text)**.

[c] F1-score **(Eq. 12 in the main text)**.

[d] Gain in precision over *random classification* **(Eq. 11 in the main text)**.

Table S12: Distribution of *hot spots* regarding fractional solvent accessible surface areas.

fSASA	No of substitution sites	Class of *hot spots*																															
		X					XI					XII				XIII				XIV				VIII					IX				
		#	$p^{a,4}$	$p^{b,5}$	$F1^{c,6}$	$g/g^{d,M}$	#	$p^{a,4}$	$p^{b,5}$	$F1^{c,6}$	$g/g^{d,M}$	#	$p^{a,4}$	$p^{b,5}$	$F1^{c,6}$	$g/g^{d,A}$	#	$p^{a,4}$	$p^{b,5}$	$F1^{c,6}$	$g/g^{d,A}$	#	$p^{a,4}$	$p^{b,5}$	$F1^{c,6}$	$g/g^{d,M}$	#	$p^{a,4}$	$p^{b,5}$	$F1^{c,6}$	$g/g^{d,A}$
≥0–≤0.2	98	1	0.17	0.01	0.02	0.31	1	0.17	0.01	0.02	0.31	2	0.33	0.02	0.04	0.62	2	0.3	0.03	0.05	0.76	4	0.70	0.04	0.08	1.23	2	0.18	0.02	0.04	0.34	/	/	/	/	/
>0.2–≤0.4	31	/	/	/	/	/	1	0.17	0.03	0.05	0.97	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	3	0.27	0.10	0.14	1.95	3	0.43	0.10	0.16	2.59
>0.4–≤0.6	26	2	0.33	0.03	0.13	2.32	2	0.33	0.08	0.13	2.32	/	/	/	/	/	3	0.5	0.06	0.10	1.68	2	0.30	0.08	0.13	2.32	1	0.09	0.04	0.05	0.83	/	/	/	/	/
>0.6–≤0.8	18	3	0.50	0.17	0.25	5.03	2	0.33	0.11	0.17	3.35	4	0.67	0.22	0.33	6.70	1	0.2	0.05	0.07	1.44	/	/	/	/	/	4	0.36	0.22	0.28	3.66	3	0.43	0.17	0.24	4.31
>0.8–≤1.0	8	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	1	0.09	0.13	0.11	2.06	1	0.14	0.13	0.13	3.23

[a] Recall **(Eq. 8 in the main text)**.

[b] Precision **(Eq. 9 in the main text)**.

[c] F1-score **(Eq. 12 in the main text)**.

[d] Gain in precision over *random classification* **(Eq. 11 in the main text)**.

Table S13: Distribution of *hot spots* regarding physicochemical properties.

Physicochemical property	No of substitution sites	Class of *hot spots*																															
		X					XI					XII					XIII					XIV					VIII					IX				
		n	*r*[a]	*p*[b]	*F1*[c]	*gip*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gip*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gip*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gip*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gip*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gip*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gip*[d]
Aliphatic	**79**	2	0.33	0.03	0.05	0.76	1	0.17	0.01	0.02	0.38	2	0.33	0.03	0.05	0.76	3	0.50	0.04	0.07	1.15	3	0.50	0.04	0.07	1.15	2	0.18	0.03	0.04	0.42	2	0.29	0.03	0.05	0.55
Aromatic	**15**	2	0.33	0.13	0.19	4.02	1	0.17	0.07	0.10	2.01	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	1	0.09	0.07	0.08	1.10	1	0.14	0.07	0.09	1.72
Neutral	**54**	2	0.33	0.04	0.07	1.12	2	0.33	0.04	0.07	1.12	3	0.5	0.06	0.10	1.68	3	0.50	0.05	0.10	1.68	3	0.50	0.06	0.10	1.68	5	0.45	0.09	0.15	1.52	3	0.43	0.06	0.10	1.44
Charged (+)	**21**	/	/	/	/	/	1	0.17	0.05	0.07	1.44	1	0.17	0.05	0.07	1.44	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
Charged (-)	**12**	/	/	/	/	/	1	0.17	0.08	0.11	2.51	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	3	0.27	0.25	0.26	4.11	1	0.14	0.08	0.11	2.15

[a] Recall **(Eq. 8 in the main text)**.

[b] Precision **(Eq. 9 in the main text)**.

[c] F1-score **(Eq. 12 in the main text)**.

[d] Gain in precision over *random classification* **(Eq. 11 in the main text)**.

Table S14: Conservation of wt*Bs*LipA residues within bacterial lipases.

Residue[a]	Conservation	$N_{BsLipA;\,T}$	$N_{BsLipA;\,D}$ (SDS)	$N_{BsLipA;\,D}$ (CTAB)	$N_{BsLipA;\,D}$ (SB3-16)	$N_{BsLipA;\,D}$ (Tween 80)
A1	2	3	0	0	0	0
E2	2	4	2	1	1	1
H3	2	1	0	1	0	0
N4	2	0	1	0	0	0
P5	10	0	1	0	0	0
V6	9	0	1	0	1	0
V7	9	0	0	0	0	0
M8	7	1	0	0	0	0
V9	7	0	1	0	0	0
H10	3	1	0	0	0	0
G11	10	0	0	0	0	0
I12	5	4	3	1	0	3
G13	2	2	6	1	5	1
G14	4	2	0	0	0	0
A15	5	5	1	1	0	1
S16	0	2	1	0	1	0
F17	4	12	1	0	1	0
N18	0	3	1	0	0	0
F19	9	0	0	0	0	0
A20	3	7	5	0	0	1
G21	0	7	5	2	0	1
I22	0	0	0	0	0	0
K23	0	4	4	0	0	0
S24	0	6	8	0	0	0
Y25	1	0	0	0	0	0
L26	8	0	0	0	0	0
V27	2	4	3	0	1	0
S28	3	2	2	0	0	0
Q29	3	1	0	0	0	0
G30	10	0	0	0	0	0
W31	9	0	0	0	0	0
S32	3	9	1	0	0	0
R33	2	6	0	0	0	0
D34	0	1	12	0	0	0
K35	0	1	0	0	0	0
L36	0	0	0	0	0	0
Y37	0	0	1	0	0	2
A38	0	0	0	0	0	0
V39	0	0	0	0	1	0
D40	4	0	1	0	0	0
F41	7	0	0	0	0	0
W42	4	2	0	0	0	0
D43	2	0	4	1	4	1
K44	0	1	0	4	3	0
T45	0	1	11	1	4	1
G46	0	0	0	3	11	0
T47	0	3	8	5	4	0
N48	5	0	0	0	3	0
Y49	1	8	9	1	3	2
N50	2	1	7	4	6	0
N51	3	0	2	1	7	2
G52	3	1	0	0	3	0

Table S14 continued.

Residue[a]	Conservation	$N_{BsLipA;\,T}$	$N_{BsLipA;\,D}$ (SDS)	$N_{BsLipA;\,D}$ (CTAB)	$N_{BsLipA;\,D}$ (SB3-16)	$N_{BsLipA;\,D}$ (Tween 80)
P53	2	1	0	0	1	0
V54	3	1	1	1	2	1
L55	7	0	0	0	0	0
S56	3	0	0	0	0	0
R57	2	0	2	2	2	0
F58	3	0	8	0	0	3
V59	9	1	3	0	0	0
Q60	3	2	10	2	1	0
K61	3	0	11	4	0	0
V62	7	0	3	0	0	0
L63	5	0	8	1	0	0
D64	3	3	5	0	0	0
E65	2	0	14	3	1	1
T66	8	0	9	0	0	0
G67	9	1	10	2	0	0
A68	7	0	1	0	0	0
K69	2	1	4	1	0	0
K70	6	0	0	0	0	0
V71	7	0	3	0	0	0
D72	4	0	0	0	0	0
I73	9	0	1	0	0	0
V74	9	0	2	0	0	0
A75	8	0	0	0	0	0
H76	5	0	0	0	0	0
S77	10	0	0	0	0	0
M78	4	0	0	0	0	0
G79	10	0	0	0	0	0
G80	1	0	0	0	0	0
A81	4	0	0	1	1	0
N82	5	0	0	2	4	0
T83	4	0	0	0	1	1
L84	3	0	0	0	0	0
Y85	3	0	0	0	3	0
Y86	6	0	0	0	0	0
I87	7	0	0	0	0	0
K88	5	1	2	0	0	0
N89	3	4	0	0	1	0
L90	0	0	0	0	1	0
D91	0	0	0	0	0	0
G92	0	0	0	0	0	0
G93	0	0	0	0	0	0
N94	0	0	1	0	0	0
K95	0	0	0	0	0	0
V96	9	0	1	0	0	0
A97	2	0	0	0	0	0
N98	2	0	4	4	1	2
V99	5	0	0	5	5	5
V100	7	1	0	0	0	0
T101	0	2	0	0	0	0
L102	1	0	0	1	0	0
G103	1	0	0	0	0	0
G104	1	0	0	0	0	0

Table S14 continued.

Residue[a]	Conservation	$N_{BsLipA;\,T}$	$N_{BsLipA;\,D}$ (SDS)	$N_{BsLipA;\,D}$ (CTAB)	$N_{BsLipA;\,D}$ (SB3-16)	$N_{BsLipA;\,D}$ (Tween 80)
A105	0	0	0	0	0	1
N106	1	0	3	2	2	0
R107	1	0	0	0	1	1
L108	3	1	0	0	1	0
T109	5	0	1	0	0	0
T110	0	0	0	1	0	0
G111	0	4	1	0	0	0
K112	0	5	1	0	0	0
A113	0	0	0	0	0	1
L114	0	1	1	0	0	0
P115	0	0	0	0	0	1
G116	0	0	0	0	0	0
T117	1	0	0	1	1	1
D118	0	0	0	0	0	0
P119	0	0	0	0	0	0
N120	1	3	0	0	0	1
Q121	0	1	0	0	0	0
K122	1	0	0	0	0	0
L123	0	0	1	0	0	1
L124	2	0	0	0	1	0
Y125	1	0	0	0	0	1
T126	5	0	1	0	0	1
S127	5	0	1	0	0	3
I128	6	0	0	0	0	1
Y129	0	0	0	0	0	0
S130	7	0	0	0	0	0
S131	1	2	1	0	0	0
A132	2	3	5	0	0	0
D133	2	0	0	0	0	0
M134	0	9	4	4	1	3
I135	1	7	1	6	2	0
V136	0	0	0	0	0	0
M137	0	10	6	1	2	1
N138	2	1	0	0	0	0
Y139	0	7	1	1	2	0
L140	0	0	0	1	0	2
S141	0	0	1	0	0	0
R142	0	4	7	1	0	0
L143	0	0	0	0	0	0
D144	0	1	7	6	1	0
G145	0	0	0	0	1	0
A146	2	0	1	0	0	0
R147	1	0	0	0	0	0
N148	2	0	0	0	0	0
V149	1	1	1	0	0	0
Q150	0	0	0	3	1	2
I151	1	0	0	0	0	0
H152	0	1	0	0	0	1
G153	0	1	1	1	1	0
V154	0	0	0	0	0	0
G155	0	4	1	0	0	0
H156	5	0	0	0	0	0

Table S14 continued.

Residue[a]	Conservation	$N_{BsLipA;\,T}$	$N_{BsLipA;\,D}$ (SDS)	$N_{BsLipA;\,D}$ (CTAB)	$N_{BsLipA;\,D}$ (SB3-16)	$N_{BsLipA;\,D}$ (Tween 80)
I157	0	1	0	1	0	0
G158	0	2	0	1	1	0
L159	0	0	0	0	0	0
L160	0	0	0	0	0	0
Y161	0	7	2	1	0	0
S162	2	2	2	0	0	0
S163	0	1	0	0	0	1
Q164	0	0	1	0	0	0
V165	1	0	0	0	0	0
N166	0	3	0	0	1	0
S167	0	0	0	0	0	0
L168	0	0	0	0	0	0
I169	1	0	0	0	0	0
K170	0	1	0	0	0	0
E171	0	0	0	0	0	0
G172	1	1	0	0	0	0
L173	2	0	0	0	0	0
N174	0	7	0	0	0	0
G175	0	0	0	0	0	0
G176	0	0	0	0	0	0
G177	0	0	0	0	0	0
Q178	0	0	3	0	0	0
N179	0	0	0	0	0	0
T180	0	0	0	0	1	0
N181	0	0	1	0	0	

[a] Underlined substitution sites are identified as *hot spots*; substitution sites highlighted in bold are predicted as *weak spots* by CNA

Table S15: Amino acid types that lead to maximum changes in thermostability or detergent tolerance.

AA type	$\Delta T_{50;\,max}$ [K]	$\Delta D_{max;\,SDS}$	$\Delta D_{max;\,CTAB}$	$\Delta D_{max;\,SB3\text{-}16}$	$\Delta D_{max;\,Tween80}$
I	3.59	0.86	1.01	1.01	2.00
A	2.79	0.58	0.45	0.97	0.86
V	1.19	0.79	0.44	1.87	0.70
L	1.51	0.67	0.22	0.76	0.45
G	4.51	1.03	0.72	2.41	0.48
F	3.78	0.68	-0.19	1.07	0.39
Y	4.22	-1.37	0.35	1.44	0.51
W	1.68	-0.20	-0.19	0.67	-0.38
C	na[a]	na[a]	na[a]	na[a]	na[a]
P	0.54	0.35	-0.19	0.90	0.30
M	7.67	1.49	0.58	2.25	0.45
S	2.55	0.68	-0.19	0.74	2.29
T	0.94	1.29	1.63	1.90	1.98
N	2.65	0.76	0.55	2.10	0.51
Q	0.66	0.68	0.34	0.87	1.07
H	0.84	0.17	0.25	-0.67	0.31
K	1.87	0.80	1.04	2.26	-0.29
R	2.79	1.45	0.51	1.76	0.67
D	1.92	0.74	0.32	1.10	0.33
E	1.00	1.10	0.23	1.27	0.55

[a] Not available.

Table S16: *Hot spot* classes for amino acids located in loops, turns, or bends and being of neutral or aliphatic type.

Criterion	No of substitution sites	Classes of *hot spot*: X					XI					XII				
		n	*r*[a]	*p*[b]	*F1*[c]	*gi p*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gi p*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gi p*[d]
Loop or turn or bend and aliphatic or neutral	43	1	0.17	0.02	0.04	0.70	3	0.50	0.07	0.12	2.10	3	0.50	0.07	0.12	2.10

Criterion	No of substitution sites	Classes of *hot spot*: XIII					XIV					VIII					IX				
		n	*r*[a]	*p*[b]	*F1*[c]	*gi p*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gi p*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gi p*[d]	*n*	*r*[a]	*p*[b]	*F1*[c]	*gi p*[d]
Loop or turn or bend and aliphatic or neutral	43	1	0.17	0.02	0.04	0.70	/	/	/	/	/	2	0.18	0.05	0.07	0.77	2	0.29	0.05	0.08	1.20

[a] Recall (**Eq. 8 in the main text**).

[b] Precision (**Eq. 9 in the main text**).

[c] F1-score (**Eq. 12 in the main text**).

[d] Gain in precision over *random classification* (**Eq. 11 in the main text**).

Supplemental Figures

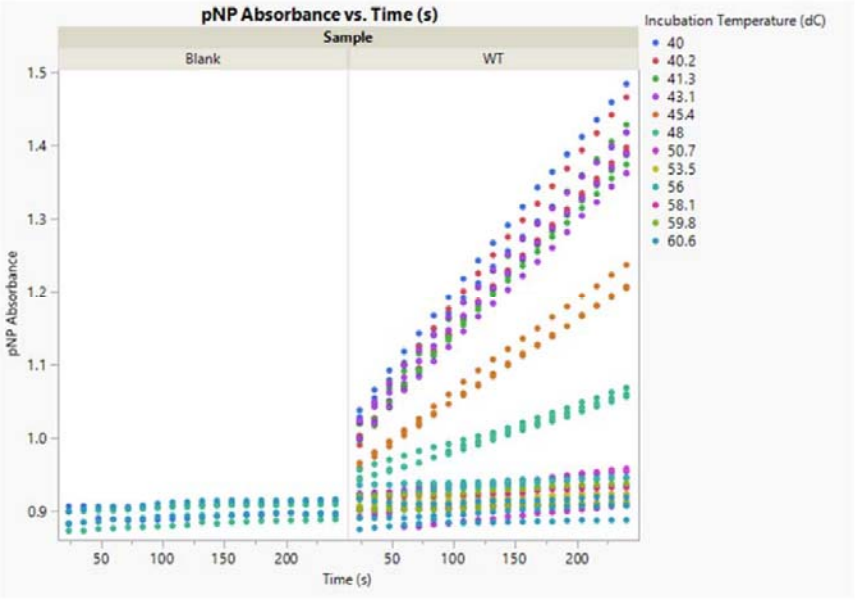


Figure S1: Control experiments regarding *p*NP absorption. The *p*NP absorption over time was measured at different temperatures between 40 and 60.6°C. On the left, the results of the control experiments, i.e., just *p*NPP in solution, are shown for temperatures up until 60.6°C. On the right, the results of *p*NP absorption in the presence of *Bs*LipA are shown. At temperatures above 48°C, the protein denatures; no increase in *p*NP absorption over time is observed then.

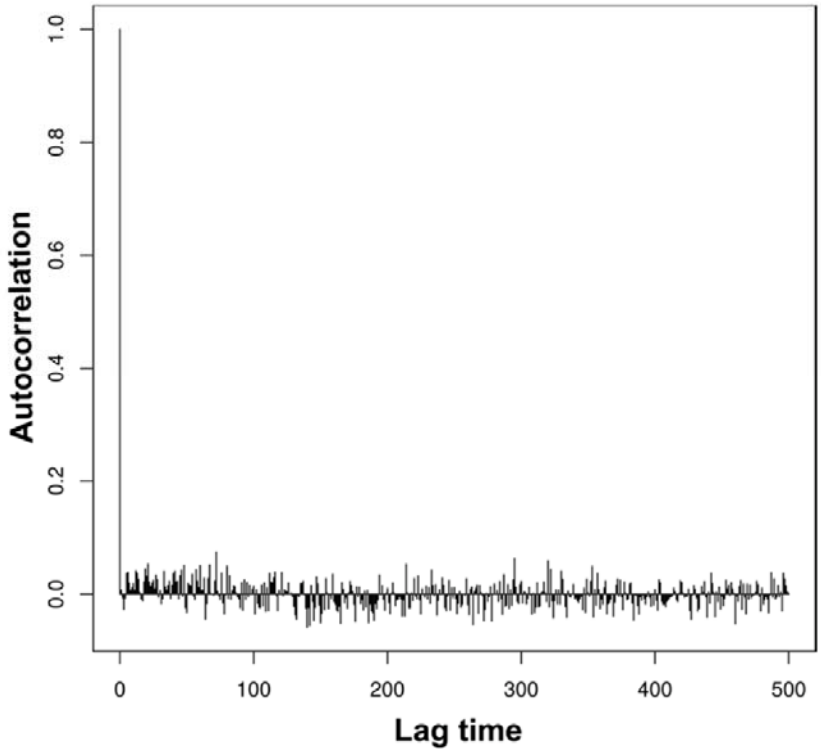


Figure S2: Autocorrelation function of the cluster configuration entropy H_{type2}. The snapshots were extracted at time intervals of 40 ps, and the lag time is in multiples of 40 ns.

Supplemental References

1. Bhairi, S, A Guide to the Properties and Uses of Detergents in Biology and Biochemistry. *Calbiochem-Novobiochem Corporation* San Diego, **1997**.

2. Fulton, A.; Frauenkron-Machedjou, V. J.; Skoczinski, P.; Wilhelm, S.; Zhu, L.; Schwaneberg, U.; Jaeger, K.-E., Exploring the Protein Stability Landscape Bacillus Subtilis Lipase a as a Model for Detergent Tolerance. *ChemBioChem* **2015**, 16, 930-936.

ORIGINAL PUBLICATION III

Promiscuous esterases counterintuitively are less flexible than specific ones

Nutschel, C., Coscolín, C., Mulnaes, D., David, B., Ferrer, M., Jaeger K.-E., Gohlke, H.

J Chem Inf Model. 2020, DOI: 10.1021/acs.jcim.1c00152.

Promiscuous esterases counterintuitively are less flexible than specific ones

Christina Nutschel[1], Daniel Mulnaes[2], Cristina Coscolín[3], Manuel Ferrer[3], Karl-Erich Jaeger[4,5], Holger Gohlke[1,2,*]

[1] John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) and Institute of Biological Information Processing (IBI-7: Structural Biochemistry), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[2] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

[3] Institute of Catalysis, Consejo Superior de Investigaciones Científicas, 28049 Madrid, Spain

[4] Institute of Molecular Enzyme Technology, Heinrich Heine University Düsseldorf, 52425 Jülich, Germany

[5] Institute of Bio- and Geosciences IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

Running title: Flexibility and promiscuity of esterases

Keywords: Esterase promiscuity, structural flexibility, thermostability, Constraint Network Analysis, TopModel, TopScore, conformational proofreading

*Corresponding author:
John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC), and Institute of Biological Information Processing (IBI-7: Structural Biochemistry)
Forschungszentrum Jülich GmbH
Wilhelm-Johnen-Straße
52425 Jülich
Germany
Email: gohlke@uni-duesseldorf.de or h.gohlke@fz-juelich.de

Abstract

Understanding mechanisms of promiscuity is increasingly important from a fundamental and application point of view. As to enzyme structural dynamics, more promiscuous enzymes generally have been recognized to also be more flexible. However, examples for the opposite received much less attention. Here, we exploit comprehensive experimental information on the substrate promiscuity of 147 esterases tested against 96 esters together with computationally efficient rigidity analyses to understand the molecular origin of the observed promiscuity range. Unexpectedly, our data reveal that promiscuous esterases are significantly less flexible than specific ones, are significantly more thermostable, and have a significantly increased specific activity. These results may be reconciled with a model according to which structural flexibility in the case of specific esterases serves for conformational proofreading. Our results signify that esterase sequence space can be screened by rigidity analyses for promiscuous esterases as starting points for further exploration in biotechnology and synthetic chemistry.

1. Introduction

Enzymes involved in primary metabolism typically exquisitely discriminate against other metabolites. Yet, evolution of specificity is only pushed by nature to the point at which 'unauthorized' reactions do not impair the fitness of the organism (1). As a result, the universe of promiscuous activities available in nature has been suggested to be enormous (2, 3). Understanding mechanisms of promiscuity has thus become increasingly important for the fundamental understanding of molecular recognition and how enzyme function has evolved over time(4) but also to optimize enzyme engineering applications (5). A particular challenge in the latter case is the ability to discover a suitable enzyme with 'sufficient' promiscuous activity to serve as a starting point for further exploration (1).

Enzyme structural dynamics, besides its role in catalysis (6, 7) and allosteric regulation (8-11), has been recognized as likely the single most important mechanism by which promiscuity can be achieved (5). Prominent examples are human cytochrome P450 (CYP) enzymes, for which crystallographic studies and molecular simulations demonstrated that more promiscuous CYPs show larger structural plasticity and mobility (12-14), or TEM-1 β-lactamase and a resurrected progenitor, for which molecular simulations show that the pocket of the ancestral, and more promiscuous, enzyme fluctuates to a greater extent (15). However, examples for the opposite, i.e., conformational changes selected in evolution such that they enhance specificity in molecular recognition (16), have received much less attention in the context of enzyme promiscuity.

A clear limitation for scrutinizing the link between enzyme structural dynamics and substrate promiscuity is the general lack of large-scale data on one enzyme (super)family tested against a multitude of ligands (17) (cf. ref. (1) for notable exceptions). Likewise, acquiring information on enzyme dynamics at the atomistic level by experimental techniques or classical molecular dynamics (MD) simulations is burdensome. Here, we exploit comprehensive experimental information on the substrate promiscuity (18) of esterases (abbreviated EHs, for "Ester Hydrolases") (19) together with computationally efficient rigidity analyses (20-23) of comparative models of EHs to understand the molecular origin of the observed promiscuity range. Enzyme rigidity, or its opposite flexibility, are static properties that denote the *impossibility*, or *possibility*, *of motions* in an enzyme under force, without giving information about directions and magnitudes of movements (23). Enzyme flexibility, thus, should not be confused with enzyme mobility, which describes *actual motions* in an enzyme. Rigidity analysis results do not rely on the correct description of the

time-dependency of processes (23), which makes them valuable in cases where timescales over multiple orders of magnitude may govern such processes, like in enzyme dynamics (6, 7).

In recent years, EHs have obtained much attention in basic research and industrial applications (24). EHs are widely distributed in nature within microbial communities (at least one EH is found in each bacterial genome), they have been extensively examined with state-of-the-art (meta)genomics techniques and investigated by functional screenings compared to many other classes of enzymes. They also possess outstanding properties in terms of stability, reactivity, and scalability that make them appropriate biocatalysts to improve competitiveness, innovation capacity, and sustainability in a modern circular bio-economy (25). Recently, a large-scale study on substrate promiscuity (P_{EH}, which denotes the number of esters hydrolyzed by an EH) of 147 phylogenetically, environmentally, and structurally diverse microbial EHs was described by Ferrer *et al.* (19), in which all EHs were functionally assessed against a customized library of 96 esters. As to mechanistic understanding, the authors related P_{EH} to a structural parameter, the active site effective volume. However, the impact of enzyme flexibility on P_{EH} was not assessed.

In our study, we thus ask the following questions: I) What is the relation between P_{EH} and EH flexibility? II) Does this relation hold if experimentally determined EH thermostabilities are used as proxies for enzyme flexibility? III) What is the relation between P_{EH} and EHs' specific activities? IV) Is there a preference of promiscuous or specific EHs for a particular type of esters. V) Can this preference be understood with respect to EHs flexibilities?

2. Materials and Methods

2.1. Definition of data sets

The present study builds on the study from Ferrer *et al.* (19). In order to assess P_{EH}, the authors experimentally investigated 147 phylogenetically, environmentally, and structurally diverse microbial EHs (termed *experimental data set*) against a customized library of 96 different esters. Two commercial lipases, which have found wide biotechnological applications, CalA and CalB from *Pseudozyma aphidis* (formerly *Candida antarctica*), were included for comparison. For details on determining and classifying P_{EH}, see **Supplemental Materials and Methods**.

As our computational approach involves extensive molecular dynamics (MD) simulations for generating conformational ensembles **(see section 2.3)**, we selected 35 EHs from the *volume data set* (termed *flexibility data set*) for comparative modeling **(see section 2.2)**. The criteria for choosing EHs of the *flexibility data set* are explained in **section 3.1**.

2.2. Comparative modelling and validations of the *flexibility data set*

Comparative models of the *flexibility data set* **(see section 2.1)** were generated using our in-house structure prediction meta-tool TopModel (26) that has been successfully applied in previous studies (27-30). TopModel uses multiple state-of-the-art threading and sequence/structure alignment tools to generate a large ensemble of models from different pairwise and multiple alignments of the top five highest ranked template structures. The TopModel software is available at https://cpclab.uni-duesseldorf.de/index.php/Software.

The quality of the homology models was assessed by our meta Model Quality Assessment Program (meta-MQAP) TopScore (31). TopScore uses deep neural networks (DNN) to combine scores from 15 different primary MQAP to predict accurate residue-wise and whole-protein error estimates. For details on model quality assessment by TopScore and validation, see **Supplemental Materials and Methods**.

To test whether CARs of the homology models are accessible for substrates, we applied the CAVER 3.0.3 PyMOL Plugin (32). Starting points for the computations were defined based on the Cartesian coordinates of the CARs' center of mass (COM). Default values were used for the probe radius (0.9 Å), shell radius (3.0 Å), and shell depth (4.0 Å).

2.3. Generation of structural ensembles

Structural ensembles of EHs were generated by all-atom MD simulations of in total 5 μs simulation time per EH. For details on starting structure preparation, parametrization, and equilibration see **Supplemental Materials and Methods**.

All minimization, equilibration, and production simulations were performed with the *pmemd.cuda* module (33) of Amber19 (34). During production simulations, we set the time step for the integration of Newton's equation of motion to 4 fs following the hydrogen mass repartitioning strategy (35). Coordinates were stored into a trajectory file every 200 ps. This resulted in 5000 configurations for each production run that were considered for subsequent analyses.

2.4. Constraint Network Analysis

The flexibility analyses were performed with the Constraint Network Analysis (CNA) software package (version 3.0) (20-23). CNA functions as front- and back-end to the graph theory-based software Floppy Inclusions and Rigid Substructure Topography (FIRST) (36). Applying CNA to biomolecules aims at identifying their composition of rigid clusters and flexible regions, which can aid in the understanding of biomolecular structure, stability, and function (21-23). As the mechanical heterogeneity of biomolecular structures is intimately linked to their diverse biological functions, biomolecules generally show a hierarchy of rigidity and flexibility (20). In CNA, biomolecules are modeled as constraint networks in a *body-and-bar* representation, which has been described in detail by Hesphenheide *et al*. (37). A fast combinatorial algorithm, the *pebble game*, counts the bond rotational degrees of freedom and floppy modes (internal, independent degrees of freedom) in the constraint network (38). In order to monitor the hierarchy of rigidity and flexibility of biomolecules, CNA performs thermal unfolding simulations by consecutively removing non-covalent constraints (hydrogen bonds, including salt bridges) from a network in increasing order of their strength (39-41). For details on thermal unfolding simulations, see **Supplemental Materials and Methods**. To improve the robustness and investigate the statistical uncertainty, we carried out CNA on ensembles of network topologies (ENT^{MD}) generated from MD trajectories **(see section 2.3)** (42).

The CNA software is available under academic license at https://cpclab.uni-duesseldorf.de/index.php/Software and the CNA web server is accessible at https://cpclab.uni-duesseldorf.de/cna.

2.5. Local and global indices

From the thermal unfolding simulations, CNA computes a comprehensive set of indices to quantify biologically relevant characteristics of the protein's stability. *Global* indices are used for determining the rigidity and flexibility at a macroscopic level; *local* indices determine the rigidity and flexibility at a microscopic level of bonds (43). The cluster configuration entropy H_{type2} is a *global* index that has been introduced by Radestock and Gohlke (20). As done previously, we applied H_{type2} as a measure for global structural stability of proteins (20, 41, 44-48). The stability map rc_{ij} is a *local* index that has been introduced by Radestock and Gohlke (20). We applied rc_{ij} as a measure for local structural stability of proteins in previous studies (45, 47, 48). For details on both indices, see **Supplemental Materials and Methods**.

2.6. Root mean square fluctuations

The per-residue root-mean-square fluctuations were calculated for each EH ($RMSF_{EH}$) and for its CARS ($RMSF_{CAR}$) based on the MD trajectories **(see section 2.3)**. Prior to the calculations, the structures of each trajectory were superimposed onto the average structure using the 90% least mobile residues of the respective EHs (49).

2.7. Torsion angles

For each of the 96 esters, the number of freely rotatable bonds (torsion angles, TA) was calculated based on the SMILES codes provided by Ferrer *et al.* (19).

To compare how many esters with a specific TA are hydrolyzed by each EH, we calculated the normalized proportion of ester hydrolysis with a specific TA ($Norm_{ester}$(TA)) as the number of hydrolyzed esters with a specific TA ($Ester_{hydrolysed}$ (TA)) divided by the total number of esters in the data set with this specific TA ($Ester_{library}$(TA)) **(Eq. 1)**.

$$Norm_{ester}(TA)\ [\%] = \frac{Ester_{hydrolyzed}(TA)}{Ester_{library}(TA)} * 100\% \qquad \textbf{Eq. 1}$$

2.8. Circular dichroism spectroscopy

Prior to analyses, soluble His-tagged proteins were produced and purified after binding to a Ni-NTA His-Bind resin as described by Ferrer *et al.* (19). Circular dichroism (CD) spectra were acquired between 190 and 270 nm with a Jasco J-720 spectropolarimeter equipped with a Peltier temperature controller, employing a 0.1 mm cell at 25°C. Spectra were analyzed, and denaturation temperatures were determined at 220 nm between 10 and 85°C at a rate of 30°C per hour, in 40 mM (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) buffer, pH 7.0. A protein concentration of 1.0 mg ml^{-1} was used. Denaturation temperatures were calculated by fitting the ellipticity (mdeg) at 220 nm at each of the different temperatures using a 5-parameters sigmoid fit with Sigma Plot 13.0.

3. Results

3.1. Definition of data sets

To understand the structural origin of and develop a method to predict P_{EH}, the present study builds on large-scale data from Ferrer *et al.* (19). The authors experimentally investigated P_{EH} of 147 EHs (termed *experimental data set*) **(see section 2.1)**. Additionally, they ranked (classified) P_{EH} of 96 EHs (termed *volume data set*) based on the active site effective volume **(see section 2.1) (Eq. S1)**, which will be used here as a reference to compare the power of P_{EH} predictions.

As our computational approach involves extensive MD simulations for generating conformational ensembles **(see section 2.3)**, we selected 35 EHs from the *volume data set* based on the following criteria; they constitute the *flexibility data set*. I) The data set contains eleven EHs with known crystal structures (including the commercial EHs CalA and CalB) **(Figure 1A, Table S1)** and 24 EHs for which no experimental structure is available but for which comparative models can be generated **(see section 3.2) (Figure 1A, Table S2)**. That way, we can probe to what extent the source of structural information influences the outcome of our results. II) The chosen EHs of the data set show high diversities as to P_{EH} and association to esterase families (F_{EH}, as defined by Arpigny and Jaeger (50)), similar to those found for the *volume data set* **(Figures S1 and S2, Tables S3 and S4)**. This resulted in P_{EH} ranging from 4 to 72 **(Figure 1A, Tables S1 and S2)**. In the following, we consider P_{EH} as *low* if the EH hydrolyzes ≤ 9 esters (11% of the data set), as *moderate* if the EH hydrolyzes between 10 and 29 esters (49%), and as *high* if the EH hydrolyzes ≥ 30 esters (40%) **(Figure S1, Table S3)**. The data set covers eleven F_{EH} of which F_{IV} (44% of the data set) and F_{V} (21%) are the best represented ones **(Figure S2, Table S4)**. This reflects the proportion of their presence in the *volume data set*. III) Only EHs with amino acid sequence identities ≥ 25% in comparison to an existing crystal structure were considered **(see section 2.1)** in order to ensure a sufficient quality of generated comparative models.

Finally, in order to uniformly depict the results across the present study, six EHs were selected as representatives of the *flexibility data set* based on P_{EH} (termed *representative data set*): EHs with the lowest (EH115) or highest P_{EH} (EH001) and known crystal structures, EHs with the lowest (EH127) or the highest P_{EH} (EH005) and unknown crystal structures, and commercial EHs with the lowest (CalA) or highest P_{EH} (CalB) **(Figure 1A-D, Tables S1 and S2)**.

3.2. Comparative models of EHs generated by TopModel show an overall and residue-wise good quality

To generate structural models of EHs as starting points for our investigations, we performed template-based modeling of the *flexibility data set* using TopModel (26) **(see section 2.2)**. In doing so, we also generated comparative models of the eleven EHs for which crystal structures are available. These structural models will be used to judge the quality of the comparative modeling.

The quality of the comparative models of the *flexibility data set* were assessed with TopScore (31), a meta Model Quality Assessment Program (meta-MQAP) **(see section 2.2)**. For the eleven Es with known crystal structure, the global TopScores range from 0.074 to 0.305 **(Figure 1A, Table S1)**. As the global TopScore describes whole-protein error estimates, this shows that the structures contain between 7.4 and 30.5% error. Notably, the global TopScores well and significantly correlate ($R^2 = 0.61$, $p = 0.004$) with values of 1 – lDDT˙ computed from comparisons of the comparative models of EHs with known crystal structure against these experimental reference structures, indicating that global TopScores are well suited to assess the model quality of EHs (**Figure S3, Table S5**). The global TopScore values of the comparative models of the other 24 EHs range from 0.087 to 0.269 **(Figure 1A, Table S2)**, indicating that these models are of equal quality than the ones for EHs with known crystal structure. The TopScore values of the *representative data set* lie in a comparable range **(Figure 1A, Tables S1 and S2)**. Moreover, the comparative models of the *flexibility data set* show low residue-wise TopScore values (31), indicating that all parts of a model are of good quality. We illustrate this for the residue-wise TopScore of the comparative models of the *representative data set* **(Figures 1B-D)**. This also applies to structural regions around CARs **(Figures 1B-D)**. That way, it was possible to confirm CARs in models of EHs with known crystal structures and to unambiguously identify CARs in models of EHs with unknown crystal structures **(Figures 1B-D, Tables S1 and S2)**.

Additionally, we validated that CARs in all models are accessible for substrates according to CAVER results (32) **(see section 2.2)**, i.e., that all models are in an open conformation: CARs are either located on the protein surface or are buried and connected with the surface by tunnels. We illustrate this for the comparative models of the *representative data set* **(Figure S4)**.

To conclude, comparative models were generated for 35 EHs of the *flexibility data set* using TopModel. The models showed both an overall and residue-wise good structural quality. Additionally, we validated that CARs in all models are accessible for substrates.

3.3. Promiscuous EHs are globally less flexible

Previous studies indicated that enzyme flexibility influences the substrate promiscuity of enzymes (12-14). For gaining insights into how the flexibility of EHs is linked to P_{EH}, we applied CNA (21, 23), a rigidity theory-based approach to analyze biomolecular statics (21-23), to the *flexibility data set* **(see sections 2.4)**. To improve the robustness and investigate the statistical uncertainty, for each of the comparative models we carried out CNA on ensembles of network topologies (ENT^{MD}) generated from five MD trajectories of 1 μs length each (44) **(see sections 2.3 and 2.4)**. In order to investigate if the global flexibility of the EHs influences P_{EH}, we predicted T_p, the phase transition temperature previously applied as a measure of structural stability of a protein (20, 41, 44-48), for each EH **(see section 2.5)**. T_p was averaged over five ensembles **(see sections 2.3 and 2.4)**, resulting in all but one case in SEM < 1.87 K **(Figure 2A, Tables S1 and S2)**.

T_p and P_{EH} of the *flexibility data set* are well and significantly correlated (R^2 = 0.60, p = $5.4*10^{-8}$) **(Figure 2A)**. To validate the consistency of our approach, we considered EHs with known or unknown crystal structures separately. In both cases, good and significant correlations between T_p and P_{EH} were revealed (known crystal structures: R^2 = 0.48, p = 0.019; unknown crystal structures: R^2 = 0.73, p = $1.1*10^{-7}$), lending support to the quality of comparative models predicted with TopModel and indicating that future predictions on EHs with unknown experimental structures should be promising. Notably, EHs with high P_{EH} have a high T_p and *vice versa*, i.e., promiscuous EHs are globally less flexible. Exemplarily, this is depicted for EHs of the *representative dataset* with known crystal structures and lowest (EH115) or highest P_{EH} (EH001), which showed T_p of 322.3 K and 357.2 K, with unknown crystal structures and lowest (EH127) or highest P_{EH} (EH005), which showed T_p of 318.6 K and 351.1 K, and CalA and CalB, which showed T_p of 346.2 K and 351.6 K **(Figure 2A, Tables S1 and S2)**. The differences in global structural stability of these EHs are illustrated by the rigid cluster decomposition at 332 K during the thermal unfolding simulations **(Figure 2B-D)**: promiscuous EHs are globally more structurally stable at the elevated temperature as indicated by fewer, but larger, rigid clusters.

The EH flexibility analyzed so far is a static property and describes the potential of motions in a biomolecule (23). Yet, direct information on mobility within EHs is available from the ensembles generated by MD simulations. We thus computed exemplarily $RMSF_{EH}$, a measure for protein mobility **(see section 2.6)**, across the ensembles of EHs from the *representative data set*. $RMSF_{EH}$, averaged over all residues and all five MD trajectories, and P_{EH} do not yield a significant correlation (p = 0.13) **(Figure S5A, Table S6)**, in contrast to T_p and P_{EH} (R^2 = 0.93, p = 1.8*10^{-3}) **(Figure S5B, Table S6)**. Still, as promiscuous EHs are globally less mobile, the same trend is obtained as in the case of the flexibility analysis.

To conclude, a good and significant correlation between T_p and P_{EH} was found for the *flexibility data set* (R^2 = 0.60, p = 5.4*10^{-8}). These findings demonstrate that promiscuous EHs are globally less flexible. $RMSF_{EH}$ is less predictive for P_{EH}, although again promiscuous EHs are characterized by a lower global mobility, mutually confirming either result.

3.4. Promiscuous EHs are more thermostable

Previous studies indicated that thermodynamically more thermostable proteins frequently have a higher structural stability (45, 48). In order to investigate if promiscuous EHs, which were predicted to be less flexible **(see section 3.3)**, are also more thermostable, CD spectroscopy was applied to determine the melting temperature T_d of the EHs **(see section 2.8)**. Note that only if the unfolding of a protein is reversible, CD spectroscopy provides true thermodynamic properties (51). However, even if the unfolding is irreversible, because the protein aggregates at high temperatures, the method can still give information about relative stabilities (51). Hence, to reduce the potential impact of different aggregation kinetics of structurally different proteins, we applied CD spectroscopy to one F_{EH} family only. In particular, we used F_{IV} because it is the largest F_{EH} **(Table S7)**.

Exemplarily, a CD spectrum for T_d determination is shown for EH001 **(Figure 3A)**; for each EH, T_d determination was performed in triplicates with STD < 0.62 K. T_d and P_{EH} yield a fair and significant correlation (R^2 = 0.40, p = 0.027) **(Figure 3B)**.

To conclude, promiscuous EHs are not only globally less flexible but also more thermostable.

3.5. Promiscuous EHs have less flexible catalytically active residues

The good correlation of P_{EH} and T_p encouraged us to investigate if local flexibility characteristics of CARs will provide an even better predictor of EH promiscuity. We thus computed $Flex_{CAR}$ for the *flexibility data set*, i.e., the stability of rigid contacts between CARs and other residues that are at most 5 Å apart from each other, based on the *local* index $rc_{ij,neighbor}$ **(see section 2.5)**. For each EH, $Flex_{CAR}$ was averaged over five ensembles **(see sections 2.3 and 2.4)**, resulting in SEM < 0.06 kcal mol^{-1} **(Figure S6A, Tables 1 and 2)**.

$Flex_{CAR}$ and P_{EH} of the *flexibility data set* yield a good and significant correlation (R^2 = 0.51, p = 1.7*10^{-6}) **(Figure S6A)**. To validate again the consistency of our approach, we considered EHs with known and unknown crystal structures separately. In both cases, good and significant correlations between $Flex_{CAR}$ and P_{EH} were found (known crystal structures: R^2 = 0.63, p = 3.7*10^{-3}; unknown crystal structures: R^2 = 0.47, p = 2.2*10^{-4}), again lending support to the quality of comparative models predicted with TopModel. Hence, EHs with high P_{EH} have low $Flex_{CAR}$ and *vice versa*, i.e., promiscuous EHs have less flexible CARs. Exemplarily, this is detailed for EHs of the *representative dataset* with known crystal structures and lowest (EH115) or highest P_{EH} (EH001), which showed $Flex_{CAR}$ of -0.74 kcal mol^{-1} and -1.91 kcal mol^{-1}, with unknown crystal structures and lowest (EH127) or highest P_{EH} (EH005), which showed $Flex_{CAR}$ of -1.10 kcal mol^{-1} and -1.86 kcal mol^{-1}, and CalA and CalB, which showed $Flex_{CAR}$ of -1.31 kcal mol^{-1} and -1.95 kcal mol^{-1} **(Figure S6A, Tables S1 and S2)**. The differences in local structural stability of these EHs are illustrated by rigid contacts between CARs and other residues that are at most 5 Å apart from each other **(Figure S6B-D)**: promiscuous EHs are locally more structurally stable as indicated by more stable rigid contacts.

Finally, we exemplarily computed $RMSF_{CAR}$, a measure for the mobility of a protein's CARs **(see section 2.6),** across the ensembles of EHs from the *representative data set*. Averaged $RMSF_{CAR}$ and P_{EH} correlate worse (R^2 = 0.74, p = 0.029) **(Figure S7A, Table S6)** than $Flex_{CAR}$ and P_{EH} (R^2 = 0.92, p = 2.4*10^{-3}) **(Figure S7B, Table S6)**, paralleling the above results for the global measures. Still, again, as promiscuous EHs have less mobile CARs, the same trend is obtained as in the case of the flexibility analysis.

To conclude, a good and significant correlation between $Flex_{CAR}$ and P_{EH} was found for the *flexibility data set* (R^2 = 0.51, p = 1.7*10^{-6}). Hence, promiscuous EHs have less flexible

CARs. $RMSF_{CAR}$ is less predictive for P_{EH}, although again promiscuous EHs are characterized by less mobile CARs, mutually confirming either result.

3.6. Promiscuous EHs have an increased specific activity

In the study by Ferrer *et al.* (19), the *experimental data set* was screened against 96 esters in a kinetic pH indicator assay **(see section 2.1)**. Besides the average specific activity $Act_{average}$ given in U / (g wet cells), also the average maximum specific activity Act_{max} was determined. Motivated by the reactivity-selectivity principle (RSP) initially introduced for organic chemistry reactions (52), which states that a more reactive chemical compound is less selective in chemical reactions, we intended to probe if P_{EH} is related to Act_{max}. For this, we established an approximate linear free-energy relationship (LFER) (53) by relating $\log(Act_{max})$ and $\log(P_{EH})$ **(Figure S8, Table S8)**. In this analysis, the CalA and CalB preparations were excluded because Act_{max} was given in U / (g total protein) there.

$\text{Log}(Act_{max})$ and $\log(P_{EH})$ of the *experimental data set* yield a good and significant correlation ($R^2 = 0.50$, $p = 4.6*10^{-23}$) **(Figure S8A)**. Likewise, $\log(Act_{max})$ and $\log(P_{EH})$ of the *flexibility data set* yield a fair and significant correlation ($R^2 = 0.22$, $p = 0.6*10^{-2}$) **(Figure S8B)**. To validate whether the same trend emerges for EHs with known and unknown crystal structures, we considered both types of EHs separately. In both cases, fair and significant correlations between $\log(Act_{max})$ and $\log(P_{EH})$ were found (known crystal structures: $R^2 = 0.34$, $p = 0.099$; unknown crystal structures: $R^2 = 0.23$, $p = 0.019$).

To conclude, good to fair and significant correlations between $\log(Act_{max})$ and $\log(P_{EH})$ of the *experimental data set* ($R^2 = 0.50$, $p = 4.6*10^{-23}$) and the *flexibility data set* ($R^2 = 0.22$, $p = 0.6*10^{-2}$) were found. Hence, promiscuous EHs have higher maximum specific activities.

3.7. Specific EHs prefer to hydrolyze large and flexible esters

Next, we investigated, which of the 96 esters was preferentially hydrolyzed by EHs with different P_{EH}. As a criterion, we chose the number of freely rotatable bonds of an ester, TA **(see section 2.7)**. We did so because TA is a combined measure for an ester's size and conformational dynamics (54). To account for the uneven distribution of esters in our data set with respect to TA, we calculated $Norm_{ester}(TA)$, i.e., the number of hydrolyzed esters with a

specific TA ($Ester_{hydrolysed}$ (TA)) divided by the total number of esters in the data set with this specific TA ($Ester_{library}$(TA)) **(see section 2.7) (Eq. 1)**.

According to TA, the esters were classified into 17 groups that ranged from small esters with no rotatable bond to large esters with 56 rotatable bonds **(Figure 4, Table S9)**. Esters with three (24% of the ester library) and four (16% of the ester library) rotatable bonds are most frequent. The analysis of the *experimental data set* revealed that promiscuous EHs have high $Norm_{ester}$ values irrespective of TA, i.e., promiscuous EHs accept a large variety of esters with different sizes and degrees of conformational dynamics **(Figure 4A, Table S9)**. In contrast, specific EHs only have high $Norm_{ester}$ values regarding esters with high TA, i.e., specific EHs preferentially hydrolyze (very) large and flexible esters **(Figure 4A, Table S9)**. The same tendency was observed for the *flexibility data set* **(Figure 4B, Table S9)**.

To conclude, promiscuous EHs accept a large variety of esters with different sizes and degrees of conformational dynamics whereas specific EHs preferentially hydrolyze (very) large and flexible esters.

4. Discussion

The main outcomes of our analyses are I) that promiscuous EHs are significantly globally less flexible and have less flexible catalytically active residues than specific ones, II) that promiscuous EHs are significantly more thermostable, III) that promiscuous EHs have a significantly increased specific activity, and IV) that specific EHs prefer to hydrolyze large and flexible esters.

We established these relations using one of the still few experimental large-scale datasets where a diverse set of EHs was functionally assessed against a customized library of dissimilar esters (19). Functional promiscuity may arise from several conditions, including the environment of the enzyme or the concentration of a substrate, which may complicate the analysis of the molecular mechanism underlying promiscuity (5). Still, functional promiscuity ultimately is a result of recognition promiscuity (5); here, we therefore focused on substrate promiscuity (18), i.e., an enzyme carries out its typical catalytic function using non-canonical substrates, in that experimental conditions had been kept constant for the assessment of the different esterase/ester combinations (19). Almost all of the EHs were unambiguously assigned to one of the F_{EH} of the Arpigny and Jaeger classification, which is based mainly on a comparison of amino acid sequences (50). Except for classes with a few members only (cyclase-like EHs and the yeast family), all other classes cover at least two of the three P_{EH} ranges such that P_{EH} cannot be assigned based on the EH's class affiliation **(Figure S9, Table S10)** and, hence, amino acid sequence information. Even family F_{IV}, which contains a higher proportion of substrate-promiscuous EHs, also contains EHs with a small substrate range.

For scrutinizing the mechanism underlying esterase promiscuity at the atomistic level, we needed to apply comparative models of EHs, since only for ~7% of the experimentally assessed EHs crystal structures were available. Restricting the generation of esterase models to sequence identifies ≥ 25% with respect to available targets yielded generally good structural models both globally and locally, as also validated against cases where crystal structures are known. Throughout our study, we probed for the consistency of our analyses between subsets of EHs for which either crystal structures are known or not; we only found quantitative differences, but no qualitative ones. One of the reasons is likely that rigidity analyses were based on structural ensembles generated by multiple and μs-long MD simulations, which markedly increases the robustness of the results (42). We furthermore showed that results are consistent irrespective of whether EH flexibility characteristics were assessed globally or only for CARs, and that mobility characteristics computed directly from

MD trajectories show the same trend, although the correlation with P_{EH} is insignificant. Finally, we used experimental melting temperatures of EHs as indicators for enzyme flexibility (45, 48), which yielded the same relation with P_{EH} as computed flexibility characteristics. Overall, these consistent and robust findings indicate that when applying this workflow to novel EHs, it should be possible to discover enzymes with 'sufficient' substrate promiscuity to serve as a starting point for further exploration in biotechnology and synthetic organic chemistry. In that respect, the flexibility characteristics of EHs analyzed here have a notably stronger predictive power than the active site effective volume introduced earlier (19) (**Figure S10, Tables S11 and S12**).

The finding that promiscuous EHs are significantly globally *less* flexible and have *less* flexible catalytically active residues than specific EHs is in stark contrast to the general view of the role of structural flexibility for promiscuity (4, 5): Besides the examples of CYP and β-lactamase mentioned above, the possibility of dynamically restructuring active sites has also been recognized for other systems as underlying their promiscuity (55-58). Finally, interactions between antibodies and antigens are likely the quintessential example of the canonical relationship between flexibility and binding promiscuity: As antibodies mature to become more specific, their flexibility is decreased (5).

It has been recognized that conformational changes may not always be necessary for promiscuity if a variety of substrates can be bound by partial recognition or the presence of multiple binding sites (5). However, these cases do not seem to be relevant reasons for EH promiscuity because partial recognition often is associated with catalytic inefficiency (1), which is contrary to our observation that P_{EH} correlates with EH activity, and the presence of multiple binding sites that could give rise to promiscuity is controverted by the finding that promiscuous EHs have large active site effective volumes (19), i.e., large pockets with few subpockets. Inversely, our findings of rigid promiscuous EHs may be consistent with the idea that multiple ligands can be accommodated in a single site by exploiting diverse interacting residues **(Figure 5)**.

Our results as to *specific but flexible* EHs may be reconciled with a model according to which conformational changes may have been selected in EH evolution for their ability to enhance specificity in recognition **(Figure 5)**, resulting in what has been termed conformational proofreading (16). In the case of specific EHs, flexibility may help to overcome a structural mismatch between the enzyme and its substrate existing when both are in their ground states, that way enhancing recognition specificity. This view is corroborated by our finding that

specific EHs prefer to hydrolyze large and flexible substrates: Larger substrates can form more interactions with the enzyme, that way helping to overcome the deformation energy required by the enzyme to optimizing the correct binding probability over the incorrect one; flexible substrates can tolerate higher strains and thus can be expected to participate in more binding events (59, 60) **(Figure 5)**.

In summary, the combined large-scale analysis of experimental EH promiscuity and computed EH flexibility reveals that promiscuous EHs are significantly less flexible than specific ones. This result is counterintuitive at first but may be reconciled with a model according to which multiple ligands can be accommodated in a single site of promiscuous EHs by exploiting diverse interacting residues, whereas structural flexibility in the case of specific EHs serves for conformational proofreading. Our results furthermore signify that EH sequence space, charted, e.g., by (meta)genomics studies, can be screened by rigidity analyses for promiscuous EHs that may serve as starting points for further exploration in biotechnology and synthetic organic chemistry.

5. Acknowledgements

CN is funded through a grant ("Vernetzungsdoktorand") provided by the Forschungszentrum Jülich. Parts of the study were supported by Bundesministerium für Bildung und Forschung (BMBF) through funding number 031B0837A "LipoBiocat" to HG and KEJ as well as the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through funding no. INST 208/704-1 FUGG to HG and INST 208/654-1 FUGG to KEJ. HG is grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich Heine University Düsseldorf. HG gratefully acknowledges the computing time granted by the John von Neumann Institute for Computing (NIC) and provided on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user IDs: HKF7; protil (project ID: 15956)) (61). MF acknowledges the grant 'INMARE' from the European Union's Horizon 2020 (grant agreement no. 634486) and BIO2017-85522-R from the Ministerio de Ciencia, Innovación y Universidades (MCIU), Agencia Estatal de Investigación (AEI), Fondo Europeo de Desarrollo Regional (FEDER) and European Union (EU). CC thanks the Ministerio de Economía y Competitividad and FEDER for a PhD fellowship (Grant BES-2015-073829). The authors are grateful to David Almendral and Ruth Matesanz for their support of CD analysis.

6. Authors contributions

HG and KEJ conceived the study. CN analyzed the experimental data, performed structure prediction, MD simulations and CNA computations, analyzed the computational data, and wrote the manuscript together with HG. DM initially contributed to the structure prediction. MF and CC measured and analyzed melting temperatures. HG supervised and managed the project. All authors reviewed and approved the manuscript.

7. Conflict of interest

The authors declare no financial and non-financial competing interests.

8. Figures

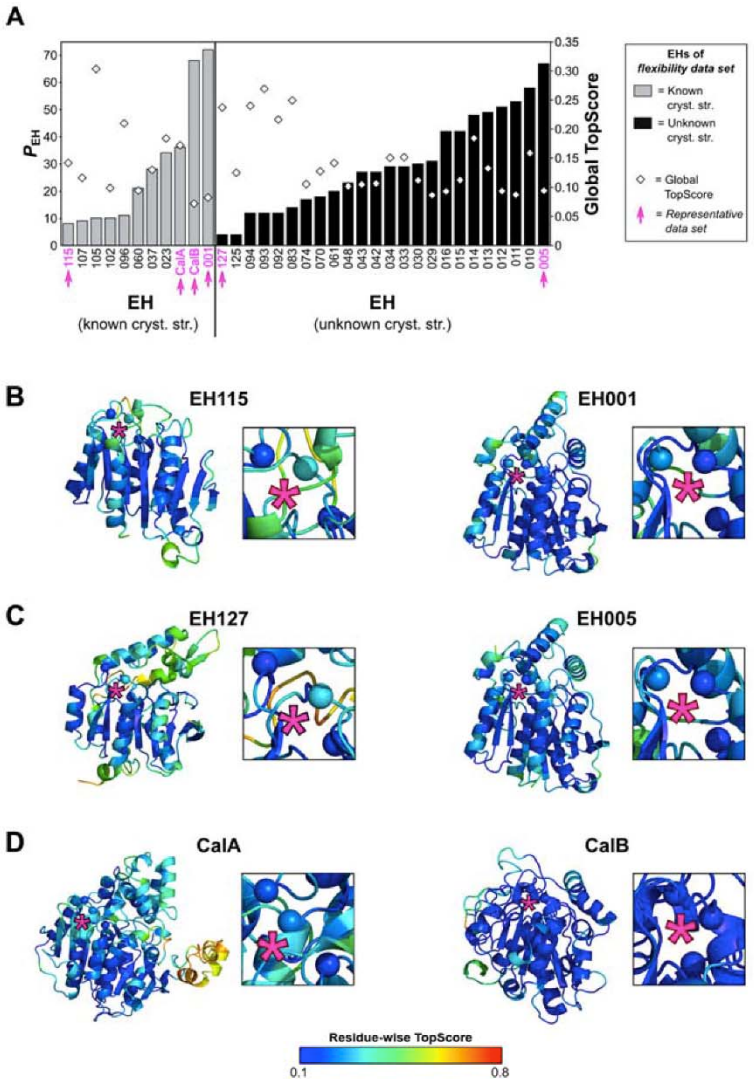


Figure 1: Comparative modeling of EHs. (A) Based on sequence data provided by a large-scale study from Ferrer *et al.* (19), comparative models were generated for 35 EHs with known (left, 11 EHs) and unknown (right, 24 EHs) crystal structures using TopModel (26). These EHs constitute the *flexibility data set*. The EHs vary in P_{EH} (left ordinate, bars) and global TopScores (right ordinate, diamonds). Six EHs were selected as representatives of the *flexibility data set* (termed *representative data set*) as indicated by magenta arrows. The quality of the comparative models of **(B)** EHs with known crystal structures and lowest (EH115) or highest P_{EH} (EH001), **(C)** EHs with unknown crystal structures and lowest (EH127) or highest P_{EH} (EH005), and **(D)** commercial EHs with highest (CalA) or lowest P_{EH} (CalB) was evaluated by TopScore (31). For each comparative model the residue-wise TopScore is shown: A good (bad) homology model shows a low (high) residue-wise TopScore (see color scale at the bottom). Insets depict CARs (spheres) within an EH. For clarity the position of CARs is indicated by magenta stars.

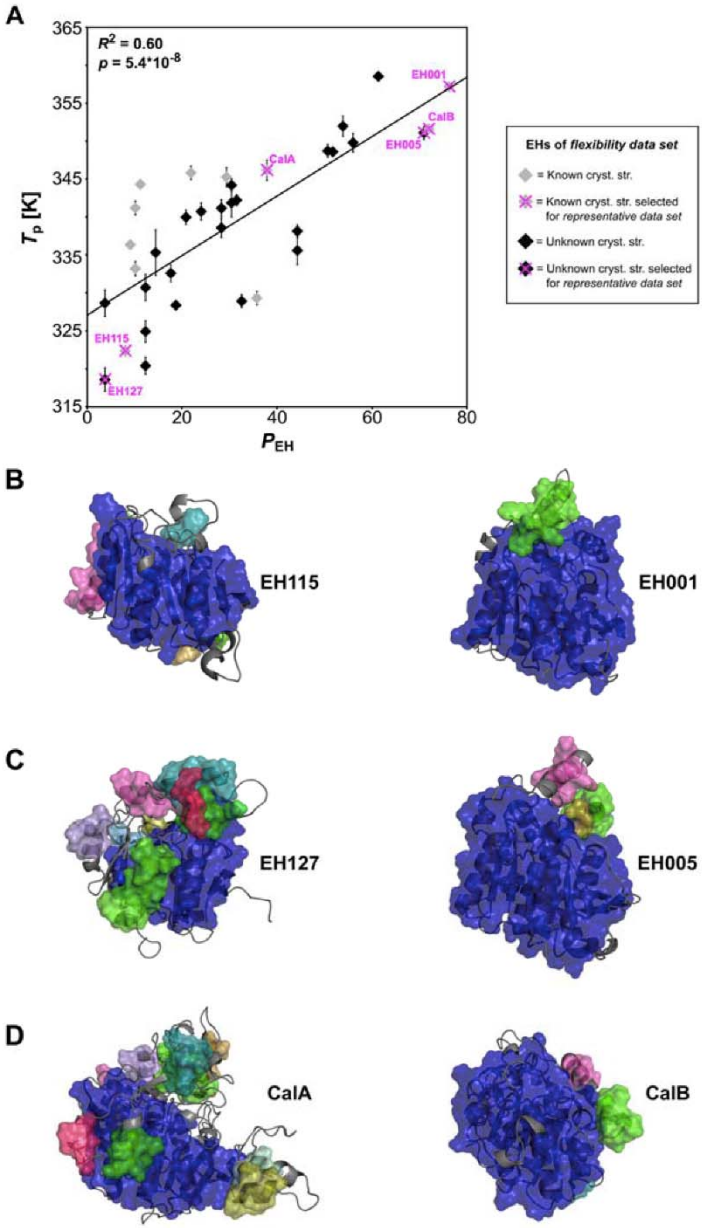


Figure 2: Correlation of T_p *versus* P_{EH}. (A) Correlation between predicted T_p based on the global index H_{type2} and P_{EH} for the *flexibility data set*. Data points colored grey (black) represent comparative models of EHs with (un)known crystal structures. The *representative data set* is indicated by magenta crosses. Error bars show the SEM over five independent MD simulations of 1 μs length each. Rigid cluster decomposition at 332 K during the thermal unfolding simulation of **(B)** EHs with known crystal structures and lowest (EH115) or highest P_{EH} (EH001), **(C)** EHs with unknown crystal structures and lowest (EH127) or highest P_{EH} (EH005), and **(D)** commercial EHs with lowest (CalA) or highest P_{EH} (CalB). Rigid clusters are represented as uniformly colored blue, green, pink, cyan, and magenta bodies in the descending order of their sizes.

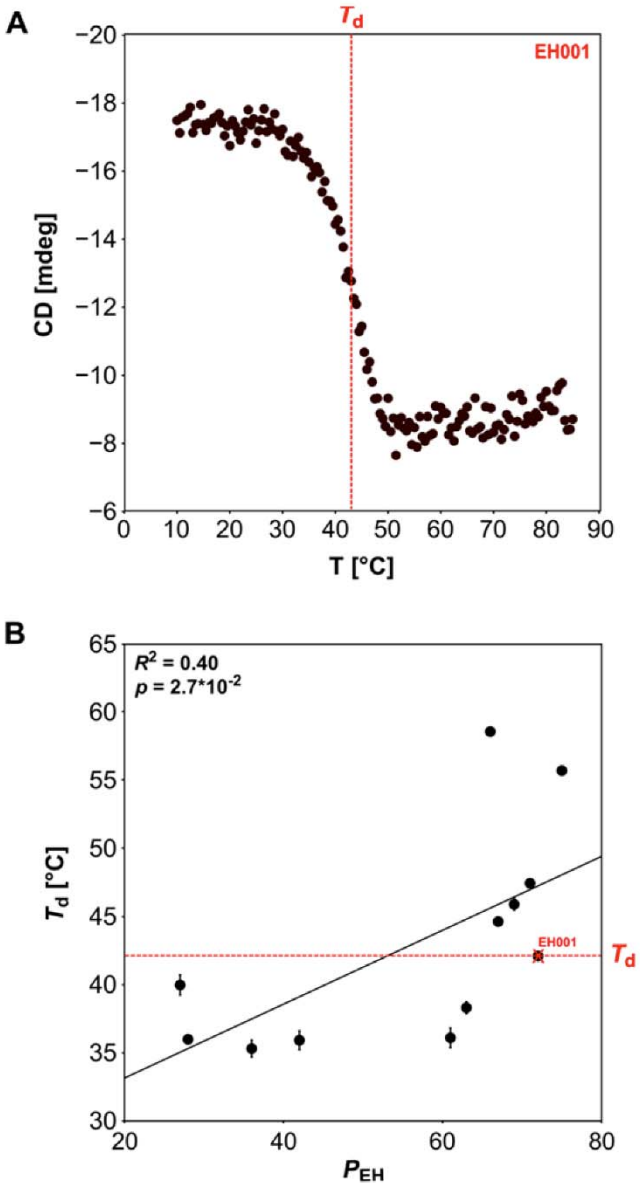


Figure 3: Determination of T_d via CD spectroscopy. (A) Exemplary CD spectrum of EH001. The ellipticity changes in mdeg at 220 nm was plotted against the temperature, resulting in a sigmoidal curve. The inflection point was used to obtain the T_d value (dotted line). **(B)** Correlation between T_d and P_{EH} for 12 EHs of F$_{IV}$.

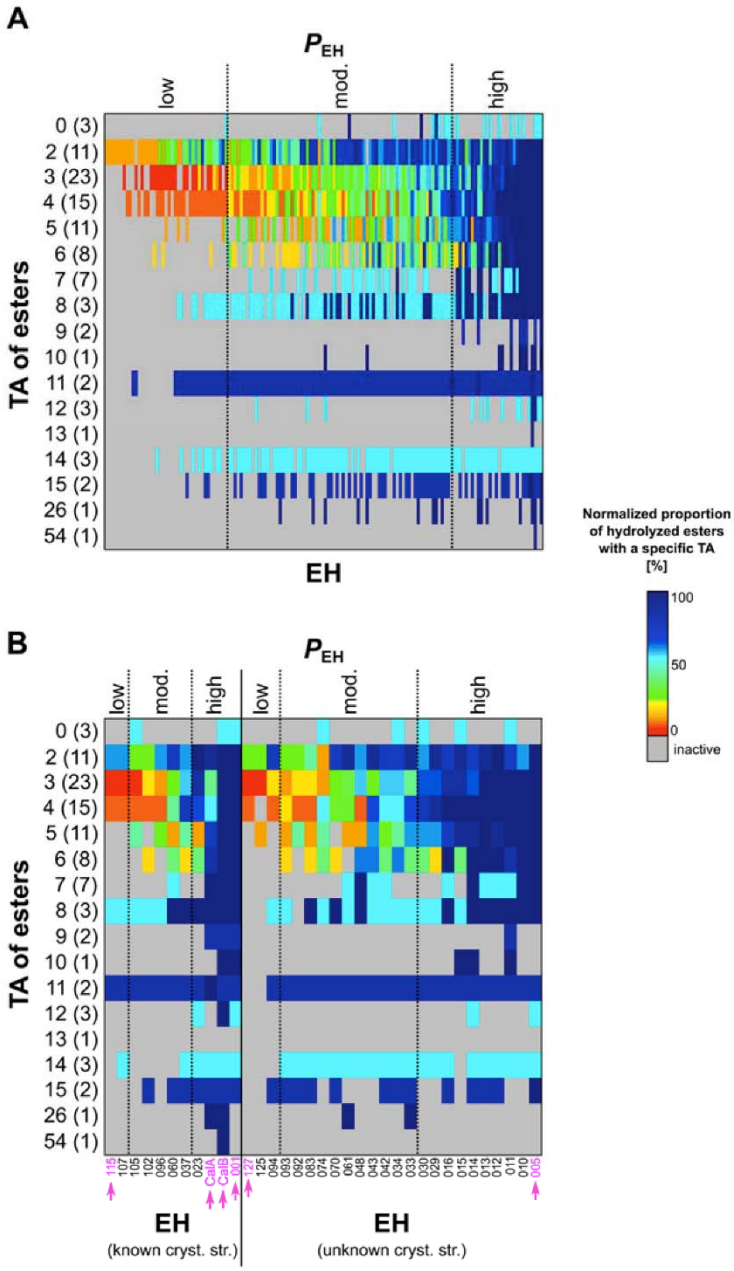


Figure 4: Relation between the number of esters' TA and P_{EH}. Relation between $Ester_{norm}$, i.e., the relative proportion of the number of hydrolyzed esters with a specific TA, and P_{EH} of **(A)** the *experimental data set* and **(B)** the *flexibility data set* containing EHs with known crystal structures (left), EHs with unknown crystal structures (right), and EHs constituting the *representative data set* (indicated by magenta arrows). TA was calculated based on SMILES codes of 96 esters provided by Ferrer *et al.* (19). A blue (red) color indicates that the EH hydrolyzes many (few) esters with a specific TA relative to the total number of esters in the data set with this specific TA (see color scale on the right); the total number of esters with a specific TA is given in brackets on the y-axis. P_{EH} is defined as *low* if the EH hydrolyzes ≤ 9 esters, as *moderate* if the EH hydrolyzes between 10 and 29 esters, and as *high* if the EH hydrolyzes ≥ 30 esters.

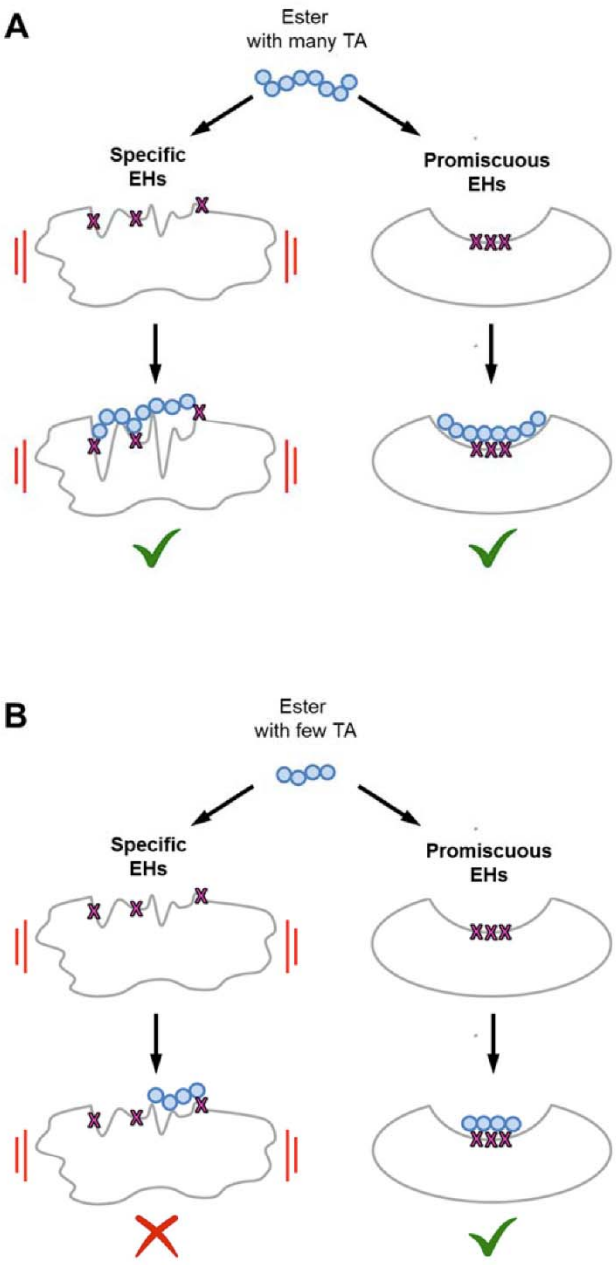


Figure 5: Mechanistic model of EH flexibility, ligand size and conformational dynamics affecting P_{EH}. Impact of esters with **(A)** many or **(B)** few TA on specific, and hence more flexible (left), and promiscuous, and hence more rigid (right), EHs. Ligand parts connected by TA are represented as blue circles. Specific EHs and large ligands with many TA can mutually adapt (panel **A**, left), and promiscuous EH can bind large ligands (panel **A**, right) and small ligands (panel **B**, right) exploiting different interaction partners. Small (and/or rigid) ligands are not able to lead to a structural adaptation of specific EHs (panel **B**, left), though, resulting in conformational proofreading. The red bars indicate the flexibility of the EHs. A green tick (red cross) indicates that ester cleavage is (not) catalyzed.

9. References

1. S. D. Copley, Shining a light on enzyme promiscuity. *Curr. Opin. Struct. Biol.* **47**, 167-175 (2017).
2. R. Chen *et al.*, Molecular insights into the enzyme promiscuity of an aromatic prenyltransferase. *Nat. Chem. Biol.* **13**, 226-234 (2017).
3. H. Huang *et al.*, Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1974-E1983 (2015).
4. O. Khersonsky, D. S. Tawfik, Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471-505 (2010).
5. I. Nobeli, A. D. Favia, J. M. Thornton, Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157-167 (2009).
6. E. Z. Eisenmesser, D. A. Bosco, M. Akke, D. Kern, Enzyme dynamics during catalysis. *Science* **295**, 1520-1523 (2002).
7. K. A. Henzler-Wildman *et al.*, A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **450**, 913-916 (2007).
8. M. J. Holliday, C. Camilloni, G. S. Armstrong, M. Vendruscolo, E. Z. Eisenmesser, Networks of dynamic allostery regulate enzyme function. *Structure* **25**, 276-286 (2017).
9. N. M. Goodey, S. J. Benkovic, Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.* **4**, 474-482 (2008).
10. H. G. Saavedra, J. O. Wrabl, J. A. Anderson, J. Li, V. J. Hilser, Dynamic allostery can drive cold adaptation in enzymes. *Nature* **558**, 324-328 (2018).
11. S. R. Tzeng, C. G. Kalodimos, Protein dynamics and allostery: an NMR view. *Curr. Opin. Struct. Biol.* **21**, 62-67 (2011).
12. J. Skopalík, P. Anzenbacher, M. Otyepka, Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *J. Phys. Chem. B.* **112**, 8165-8173 (2008).
13. T. Hendrychová *et al.*, Flexibility of human cytochrome P450 enzymes: molecular dynamics and spectroscopy reveal important function-related variations. *Biochim. Biophys. Acta.* **1814**, 58-68 (2011).
14. M. Ekroos, T. Sjögren, Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 13682-13687 (2006).
15. T. Zou, V. A. Risso, J. A. Gavira, J. M. Sanchez-Ruiz, S. B. Ozkan, Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. *Mol. Biol. Evol.* **32**, 132-143 (2015).
16. Y. Savir, T. Tlusty, Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition. *PloS One* **2**, e468 (2007).
17. M. Ferrer *et al.*, Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. *Microb. Biotechnol.* **9**, 22-34 (2016).
18. K. Hult, P. Berglund, Enzyme promiscuity: mechanism and applications. *Trends. Biotechnol.* **25**, 231-238 (2007).
19. M. Martinez-Martinez *et al.*, Determinants and prediction of esterase substrate promiscuity patterns. *ACS Chem. Biol.* **13**, 225-234 (2017).
20. S. Radestock, H. Gohlke, Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng. Life Sci.* **8**, 507-522 (2008).
21. C. Pfleger, P. C. Rathi, D. L. Klein, S. Radestock, H. Gohlke, Constraint Network Analysis (CNA): a Python software package for efficiently linking biomacromolecular structure, flexibility,(thermo-) stability, and function. *J. Chem. Inf. Model.* **53**, 1007-1015 (2013).

22. D. M. Krüger, P. C. Rathi, C. Pfleger, H. Gohlke, CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure,(thermo-) stability, and function. *Nucleic Acids Res.* **41**, W340-W348 (2013).
23. S. M. Hermans, C. Pfleger, C. Nutschel, C. A. Hanke, H. Gohlke, Rigidity theory for biomolecules: concepts, software, and applications. *Comput. Mol. Sci.* **7**, e1311 (2017).
24. K. de Godoy Daiha, R. Angeli, S. D. de Oliveira, R. V. Almeida, Are lipases still important biocatalysts? A study of scientific publications and patents for technological forecasting. *PloS One* **10**, e0131624 (2015).
25. M. Ferrer *et al.*, Biodiversity for biocatalysis: a review of the α/β-hydrolase fold superfamily of esterases-lipases discovered in metagenomes. *Biocatal. Biotranfor.* **33**, 235-249 (2015).
26. D. Mulnaes *et al.*, TopModel: Template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks. *J. Chem. Theory. Comput.* **16**, 1953-1967 (2020).
27. H. Gohlke *et al.*, Binding region of alanopine dehydrogenase predicted by unbiased molecular dynamics simulations of ligand diffusion. *J. Chem. Inf. Model.* **53**, 2493-2498 (2013).
28. N. Widderich *et al.*, Molecular dynamics simulations and structure-guided mutagenesis provide insight into the architecture of the catalytic core of the ectoine hydroxylase. *J. Mol. Biol.* **426**, 586-600 (2014).
29. Z. Zhang *et al.*, Determinants of FIV and HIV Vif sensitivity of feline APOBEC3 restriction factors. *Retrovirology* **13**, 46 (2016).
30. D. Milić *et al.*, Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1. *Sci. Rep.* **8**, 3890 (2018).
31. D. Mulnaes, H. Gohlke, TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment. *J. Chem. Theory Comput.* **14**, 6117-6126 (2018).
32. E. Chovancova *et al.*, CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.* **8**, e1002708 (2012).
33. R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, R. C. Walker, Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878-3888 (2013).
34. D. A. Case *et al.* (2019) AMBER 2019. (University of California, San Francisco).
35. C. W. Hopkins, S. Le Grand, R. C. Walker, A. E. Roitberg, Long-time-step molecular dynamics through hydrogen mass repartitioning. *J. Chem. Theory Comput.* **11**, 1864-1874 (2015).
36. D. J. Jacobs, A. J. Rader, L. A. Kuhn, M. F. Thorpe, Protein flexibility predictions using graph theory. *Proteins* **44**, 150-165 (2001).
37. B. Hespenheide, D. Jacobs, M. Thorpe, Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J. Condens. Matter Phys.* **16**, S5055 (2004).
38. D. J. Jacobs, M. F. Thorpe, Generic rigidity percolation: the pebble game. *Phys. Rev. Lett.* **75**, 4051 (1995).
39. A. J. Rader, B. M. Hespenheide, L. A. Kuhn, M. F. Thorpe, Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3540-3545 (2002).
40. D. R. Livesay, D. J. Jacobs, Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins* **62**, 130-143 (2006).
41. S. Radestock, H. Gohlke, Protein rigidity and thermophilic adaptation. *Proteins* **79**, 1089-1108 (2011).

42. C. Pfleger, H. Gohlke, Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. *Structure* **21**, 1725-1734 (2013).
43. C. Pfleger, S. Radestock, E. Schmidt, H. Gohlke, Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* **34**, 220-233 (2013).
44. P. C. Rathi, S. Radestock, H. Gohlke, Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* **159**, 135-144 (2012).
45. P. C. Rathi, K.-E. Jaeger, H. Gohlke, Structural rigidity and protein thermostability in variants of lipase A from Bacillus subtilis. *PloS One* **10**, e0130289 (2015).
46. M. Dick *et al.*, Trading off stability against activity in extremophilic aldolases. *Sci. Rep.* **6**, 17908 (2016).
47. P. C. Rathi, A. Fulton, K.-E. Jaeger, H. Gohlke, Application of rigidity theory to the thermostabilization of lipase A from Bacillus subtilis. *PLoS Comput. Biol.* **12**, e1004754 (2016).
48. C. Nutschel *et al.*, Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for Bacillus subtilis lipase A. *J. Chem. Inf. Model.* **60**, 1568-1584 (2020).
49. H. Gohlke, L. A. Kuhn, D. A. Case, Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* **56**, 322-337 (2004).
50. J. L. Arpigny, K.-E. Jaeger, Bacterial lipolytic enzymes: classification and properties. *Biochem. J.* **343**, 177-183 (1999).
51. N. J. Greenfield, Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat. Protoc.* **1**, 2527-2535 (2006).
52. H. Mayr, A. R. Ofial, The reactivity–selectivity principle: an imperishable myth in organic chemistry. *Angew. Chem. Int. Ed. Engl.* **45**, 1844-1854 (2006).
53. P. R. Wells, Linear Free Energy Relationships. *Chem. Rev.* **63**, 171-219 (1963).
54. H.-J. Böhm, The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **8**, 243-256 (1994).
55. C. M. Seibert, F. M. Raushel, Structural and catalytic diversity within the amidohydrolase superfamily. *Biochemistry* **44**, 6383-6391 (2005).
56. U. Oppermann *et al.*, Short-chain dehydrogenases/reductases (SDR): the 2002 update. *Chem. Biol. Interact.* **143-144**, 247-253 (2003).
57. S. Fushinobu, H. Nishimasu, D. Hattori, H.-J. Song, T. Wakagi, Structural basis for the bifunctionality of fructose-1, 6-bisphosphate aldolase/phosphatase. *Nature* **478**, 538-541 (2011).
58. J. Du, R. F. Say, W. Lü, G. Fuchs, O. Einsle, Active-site remodelling in the bifunctional fructose-1, 6-bisphosphate aldolase/phosphatase. *Nature* **478**, 534-537 (2011).
59. G. R. Stockwell, J. M. Thornton, Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* **356**, 928-944 (2006).
60. E. Perola, P. S. Charifson, Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **47**, 2499-2510 (2004).
61. D. Krause, JUWELS: Modular Tier-0/1 Supercomputer at the Jülich Supercomputing Centre. *JLSRF* **5**, A135 (2019).

ORIGINAL PUBLICATION III-SUPPORTING INFORMATION

Promiscuous esterases counterintuitively are less flexible than specific ones

Nutschel, C., Coscolín, C., Mulnaes, D., David, B., Ferrer, M., Jaeger K.-E., Gohlke, H.

J Chem Inf Model. 2020, DOI: 10.1021/acs.jcim.1c00152.

Supporting Information

Promiscuous esterases counterintuitively are less flexible than specific ones

Christina Nutschel[1], Daniel Mulnaes[2], Cristina Coscolín[3], Manuel Ferrer[3], Karl-Erich Jaeger[4,5], Holger Gohlke[1,2,*]

[1] John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) and Institute of Biological Information Processing (IBI-7: Structural Biochemistry), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

[2] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

[3] Institute of Catalysis, Consejo Superior de Investigaciones Científicas , 28049 Madrid, Spain

[4] Institute of Molecular Enzyme Technology, Heinrich Heine University Düsseldorf, 52425 Jülich, Germany

[5] Institute of Bio- and Geosciences IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

*Corresponding author:
John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC), and Institute of Biological Information Processing (IBI-7: Structural Biochemistry)
Forschungszentrum Jülich GmbH
Wilhelm-Johnen-Straße
52425 Jülich
Germany
Email: gohlke@uni-duesseldorf.de or h.gohlke@fz-juelich.de

Table of Contents

Supplemental Materials and Methods

Determining and classifying P_{EH}

Ferrer *et al.* (1) examined P_{EH} of all esterases (EHs) with a kinetic pH indicator assay (2-4), which unambiguously allows quantifying specific activities (*Act*) at pH 8.0 and 30 °C, using a substrate concentration above 0.5 mM. The specific activities were given in units (U) / (g wet cells); for CalA and CalB preparations, the specific activities were given in U / (g total protein). The assays were performed as triplicates, with the average specific activity ($Act_{average}$) given and standard deviation (STD) ≤ 1% in all cases. Additionally, the average maximum specific activity (Act_{max}) was determined for each EH.

In order to rank (classify) P_{EH}, the authors introduced a structural parameter, the active site effective volume (Vol_{eff}). Vol_{eff} represents the topology of the active site in terms of the active site cavity volume (Vol_{cav}) computed by Fpocket (5) divided by the relative solvent-accessible surface area ($SASA_{rel}$) using GetArea Web server (6) **(Eq. S1)**.

$$Vol_{eff}[\text{Å}^3] = \frac{Vol_{cav}\,[\text{Å}^3]}{SASA_{rel}} \qquad \textbf{Eq. S1}$$

Vol_{eff} was computed for 96 EHs (termed *volume data set*) for which the following four criteria were satisfied:

I. Eleven EHs with known crystal structures were included.

II. Homology models of 85 EHs with unknown crystal structures were generated using the Prime software from Schrödinger (7) (known crystal structures from EHs in I were used as templates).

III. EHs in II showed sequence identities ≥ 25% (in comparison to known crystal structures from EHs in I).

IV. Catalytically active residues (CARs) were unambiguously identified.

Model quality assessment by TopScore and validation

TopScore (8) predicts 1 – lDDT, with lDDT being the local Distance Difference Test (9), a measure for structural similarity that does not require superimpositioning of two structures. Therefore, the range of TopScore is [0, 1], with 0 (1) indicating low (high) estimated errors of the residues and models.

For validation, 1 - lDDT was also computed for EHs with known crystal structure and the respective comparative model, using the lDDT web server from Swiss-Model (9). Note

that in this case, the comparative model was generated by TopModel excluding the known crystal structure.

Starting structure preparation, parametrization, and equilibration

The EH structures were preprocessed with the Protein Preparation Wizard of Schrödinger's Maestro Suite (7). Of a crystal structure, we used only chain A and removed structurally resolved water molecules and ligands. Non-resolved termini were connected to acetyl (ACE) and *N*-methyl amide (NME) groups to avoid artificially charged termini. In order to match the experimental conditions of pH 8.0 **(see section 2.1)**, we used Epik (10) to calculate the p*K*a of relevant functional groups. All hydrogen atoms were then added according to the Amber ff14SB library (11). The prepared EH structures were solvated with OPC water (12), leaving at least 12 Å between the EH structure and the edges of the solvent box, by using LeaP of Amber19 (13). We also added sodium counter ions to ensure the neutrality of the system.

We used the Amber ff14SB force field (11) to parametrize the protein. Ion parameters were taken from Joung and Cheatham (14). The detailed minimization, thermalization, and equilibration protocol has been reported in ref. (15). In short, the system was initially subjected to three rounds of energy minimization to get rid of any bad contacts. The system was heated to 300 K and the pressure was adapted such that a density of 1 g cm^{-3} was obtained. During thermalization and density adaptation, we kept the solute fixed by positional restraints of 1 kcal mol^{-1} $Å^{-2}$, which were gradually removed. Subsequently, the system was subjected to five independent NPT production simulations of 1 µs length each using unbiased MD simulations. Therefore, the initial velocities were randomly assigned during the first step of the production simulations.

Thermal unfolding simulations

Therefore, a hydrogen bond energy E_{HB} is computed by a modified version of the potential by Mayo *et al*. (16). For a given network state $\sigma = f(T)$, hydrogen bonds with an energy $E_{HB} > E_{cut}(\sigma)$ are removed from the network at temperature T. In the present study, thermal unfolding simulations were carried out by decreasing E_{cut} from -0.1 kcal mol^{-1} to -6.0 kcal mol^{-1} with a step size of 0.1 kcal mol^{-1}. As E_{cut} can be converted to a temperature T using the linear equation introduced by Radestock *et al*. (17, 18) **(Eq. S2)**, the range of E_{cut} is equivalent to increasing the temperature from 302 K to 420 K with a step size of 2 K. Along the thermal unfolding simulations, hydrophobic interactions were not removed because they remain constant in strength or become even stronger with increasing temperature (19).

$$T = \frac{-20\ \mathrm{K}}{\mathrm{kcal} \cdot \mathrm{mol}^{-1}} E_{\mathrm{cut}} + 300\ \mathrm{K} \qquad \textbf{Eq. S2}$$

Cluster configuration entropy and stability map

The cluster configuration entropy H_{type2} was used to identify the phase transition temperature T_{p} of the EHs constituting the *flexibility data set* during the thermal unfolding simulation. At T_{p}, the protein switches from a rigid (structurally stable) to a floppy (unfolded) state. However, the percolation behavior of protein networks is usually more complex, and multiple phase transitions can be observed (17, 18, 20-24). Initially, the protein network is dominated by a giant rigid cluster, and H_{type2} is low because of the limited number of possible ways to configure a system with this cluster. When the giant rigid cluster starts to decay or stops to dominate the network, H_{type2} jumps. There, the network is in a partially flexible state with many ways to configure a system consisting of many small clusters. In order to determine T_{p}, a double sigmoid fit was applied to an H_{type2} *versus* $T(E_{\mathrm{cut}})$ curve as done previously (17, 18, 20-24), and T_{p} taken as that T value associated with the largest slope of the fit. The rigid cluster decomposition of the EHs was visually inspected by VisualCNA (25), an easy-to-use PyMOL plug-in that allows setting up CNA runs and analyzing CNA results linking data plots with molecular graphics representations. VisualCNA is available under an academic license from https://cpclab.uni-duesseldorf.de/index.php/Software.

During a thermal unfolding simulation, the stability map rc_{ij} indicates for all residue pairs the E_{cut} value at which a rigid contact rc between the two residues i and j (represented by their C_α atoms) is lost; rc exists as long as i and j belong to the same rigid cluster c of the set of rigid clusters $C^{E_{\mathrm{cut}}}$ (26). Thus, rc_{ij} contains information about the rigid cluster decomposition cumulated over all network states σ during the thermal unfolding simulation. The sum over all entries in rc_{ij} yields the chemical potential energy due to non-covalent bonding, obtained from the coarse-grained, residue-wise network representation of the underlying protein structure (21). In the present study, we applied the neighbor stability map $rc_{ij,neighbor}$ of each EH to investigate short-range rigid contacts. For this, as done previously (21, 24), rc_{ij} was filtered such that only rigid contacts between two residues that are at most 5 Å apart from each other were considered. Here, in particular, we focused on rigid contacts between CARs and other residues at most 5 Å apart and calculated the average over all such entries in $rc_{ij,neighbor}$ (termed $Flex_{\mathrm{CAR}}$).

Supplemental Tables

Table S1: Information about EHs of the *flexibility data set* with known crystal structures.

EH	PDB	P_{EH}[a]	F_{EH}[b]	Global Topscore[c]	CARs[d]	T_D [K][e]	$Flex_{CAR}$ [kcal/mol][f]
001	5JD4_A	72	IV	0.0849	S161; D256; H286	357.19 ± 0.46	-1.91 ± 0.05
CalB	4K6G_A	68	Yeast	0.0744	S107; D189; H226	351.60 ± 0.51	-1.95 ± 0.07
CalA147	3GUU_A	36	Yeast	0.1739	S205; D355; H387	346.17 ± 1.36	-1.31 ± 0.06
023	4Q3O_A	34	IV	0.1855	S194; D290; H320	329.31 ± 0.89	-1.23 ± 0.02
037	5JD5_A	28	IV	0.1318	S169; D265; H295	345.27 ± 1.29	-1.15 ± 0.05
060	4I3F_A	21	C-C MCPh	0.0968	S104; D230; H258	345.82 ± 0.87	-1.41 ± 0.04
096	4FBM_A	11	V	0.2114	S126; D227; H257	344.31 ± 0.57	-1.06 ± 0.06
102	5JD3_A	10	II	0.1010	S15; D192; H195	333.18 ± 0.93	-1.46 ± 0.07
105	5IBZ_A	10	Cyclase-like	0.3046	F84; R87; Q127; Q131; D133; H137; H286; E299	341.19 ± 0.91	-0.66 ± 0.03
107	4Q3L_A	9	V	0.1181	S97; D221; H249	336.34 ± 0.35	-1.36 ± 0.04
115	4Q3K_A	8	I	0.1436	S113; D169; H201	322.39 ± 0.65	-0.74 ± 0.05

[a] Experimentally measured substrate promiscuity level of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

[b] EH families based on the Arpigny and Jaeger classification (27) **(see section 2.1)**.

[c] Whole-protein error estimates predicted by TopScore for comparative models (8) **(see section 2.2)**.

[d] Catalytically active residues of EHs.

[e] Global flexibilities of EHs with SEM based on predicted H_{type2} **(see section 2.5)**.

[f] Local flexibilities of catalytically active residues of EHs with SEM based on predicted $rc_{ij,neighbor}$ **(see section 2.5)**.

Table S2: Information about EHs of the *flexibility data set* with unknown crystal structures.

EH	P_{EH}[a]	F_{EH}[b]	Global TopScore[c]	CARs[d]	T_p [K][e]	$Flex_{CAR}$ [kcal/mol][f]
005	67	IV	0.0945	S159; D254; H284	351.10 ± 0.88	-1.86 ± 0.05
010	58	IV	0.1586	S181; D279; H309	358.55 ± 0.38	-1.50 ± 0.05
011	53	IV	0.0878	S159; D254; H284	349.80 ± 1.25	-1.34 ± 0.05
012	51	IV	0.0939	S159; D254; H284	351.98 ± 1.35	-1.66 ± 0.06
013	49	IV	0.1325	S171; D268; H298	348.61 ± 0.45	-1.24 ± 0.06
014	48	IV	0.1840	S190; D290; H320	348.72 ± 0.88	-1.42 ± 0.05
015	42	IV	0.1123	S146; E240; H270	338.14 ± 0.84	-1.40 ± 0.06
016	42	IV	0.0934	S159; D254; H284	335.59 ± 1.87	-1.87 ± 0.05
029	31	IV	0.0869	S159; D254; H284	328.91 ± 0.89	-1.48 ± 0.06
030	30	VI	0.1118	S116; D164 ; H195	342.22 ± 0.75	-0.67 ± 0.06
033	29	C-C MCPh	0.1514	S104; D225; H253	344.20 ± 0.84	-1.26 ± 0.06
034	29	VI	0.1505	S119; D173; H204	341.87 ± 1.87	-1.01 ± 0.06
042	27	IV	0.1065	S125; E218; H248	341.16 ± 1.12	-1.21 ± 0.04
043	27	IV	0.1050	S144; E238; H268	338.57 ± 1.24	-0.72 ± 0.04
048	23	IV	0.1020	S146; E240; H270	340.73 ± 1.14	-1.13 ± 0.05
061	20	VI	0.1414	S118; D172; H203	339.97 ± 0.93	-0.66 ± 0.05
070	18	CE	0.1273	S189; D279; H308	328.37 ± 0.36	-1.09 ± 0.05
074	17	V	0.1056	S94; D203; H231	332.59 ± 1.12	-1.05 ± 0.05

083	14	V	0.2488	S120; D247; H275	335.31 ± 3.01	-0.68 ± 0.05
092	12	V	0.2161	S105; D233; H261	324.94 ± 1.44	-1.12 ± 0.05
093	12	VII	0.2687	S183; D311; H420	320.43 ± 1.13	-1.30 ± 0.06
094	12	V	0.2396	S127; D246; H279	330.70 ± 1.74	-0.70 ± 0.05
125	4	/	0.1248	S70; D149; H174	328.66 ± 1.73	-0.94 ± 0.03
127	4	V	0.2367	S101; D236; H263	318.60 ± 1.57	-1.10 ± 0.05

[a] Experimentally measured substrate promiscuity level of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

[b] EH families based on the Arpigny and Jaeger classification (27) **(see section 2.1)**.

[c] Whole-protein error estimates predicted by TopScore for the comparative models (8) **(see section 2.2)**.

[d] Catalytically active residues of EHs.

[e] Global flexibilities of EHs with SEM based on predicted H_{type2} **(see section 2.5)**.

[f] Local flexibilities of catalytically active residues of EHs with SEM based on predicted $rc_{ij,neighbor}$ **(see section 2.5)**.

Table S3: Comparison between the *volume data set* and the *flexibility data set* regarding P_{EH}.

P_{EH}[a]	#EHs of the *volume data set*[b]	#EHs of the *flexibility data set*[c]
Low	19 (19.79)	4 (11.43)
Moderate	51 (53.13)	17 (48.57)
High	26 (27.08)	14 (40.00)
#EHs	**96**	**35**

[a] Experimentally measured substrate promiscuity level of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**. P_{EH} is defined as *low* if the EH hydrolyzes ≤ 9 esters, as *moderate* if the EH hydrolyzes between 10 and 29 esters, and as *high* if the EH hydrolyzes ≥ 30 esters.

[b] Values in brackets represent the relative proportions of EHs in the *volume data set* in %.

[c] Values in brackets represent the relative proportions of EHs in the *flexibility data set* in %.

Table S4: Comparison between the *volume data set* and the *flexibility data set* regarding F_{EH}.

F_{EH}[a]	#EHs of the *volume data set*[b]	#EHs of the *flexibility data set*[c]
FI	6 (6.52)	1 (2.94)
FII	7 (7.61)	1 (2.94)
FIV	32 (34.78)	15 (44.12)
FV	23 (25.00)	7 (20.59)
FVI	5 (5.43)	3 (8.82)
FVII	4 (4.35)	1 (2.94)
CE	3 (3.26)	1 (2.94)
C-C MCPh	9 (9.78)	2 (5.88)
Cyclase-like	1 (1.09)	1 (2.94)
Yeast class	2 (2.17)	2 (5.88)
Unclassified	4 (4.35)	1 (2.94)
#EHs	**96**	**35**

[a] EH families based on the Arpigny and Jaeger classification (27) **(see section 2.1)**.

[b] Values in brackets represent the relative proportions of EHs in the *volume data set* in %.

[c] Values in brackets represent the relative proportions of EHs in the *flexibility data set* in %.

Table S5: TopScore performance on comparative models of EHs of the *flexibility data set* with known crystal structures.

EH	PDB ID[a]	Global TopScore[b]	1 - lDDT score[c]
001	5JD4_A	0.1501	0.2019
CalB	4K6G_A	0.0958	0.0998
CalA	3GUU_A	0.2219	0.1139
023	4Q3O_A	0.2578	0.1976
037	5JD5_A	0.2291	0.2622
060	4I3F_A	0.1644	0.195
096	4FBM_A	0.1997	0.146
102	5JD3_A	0.0954	0.1376
105	5IBZ_A	0.3519	0.6241
107	4Q3L_A	0.1798	0.2249
115	4Q3K_A	0.1788	0.2128

[a] PDB IDs that were used as references to calculate the 1 – lDDT (local Distance Difference Test) scores.

[b] Whole-protein error estimates predicted by TopScore (8) **(see section 2.2)**.

[c] Local Distance Difference Test computed by the Swiss-Model web server (9) **(see section 2.2)**.

Table S6: $RMSF_{EH}$ and $RMSF_{CAR}$ of the *representative data set.*

EH	$RMSF_{EH}$ [Å][a]	$RMSF_{CAR}$ [Å][b]
115	1.27 ± 0.03	1.60 ± 0.22
001	0.91 ±0.01	0.60 ± 0.03
127	1.54 ± 0.04	1.03 ± 0.06
005	1.13 ± 0.02	0.70 ± 0.04
CalA	1.76 ± 0.04	1.03 ± 0.07
CalB	0.90 ± 0.02	0.62 ± 0.02

[a] Average per-residue root-mean-square fluctuations for EHs **(see section 2.6)**.

[b] Average per-residue root-mean-square fluctuations for catalytically active residues of EHs **(see section 2.6)**.

Table S7: Melting temperatures of EHs determined by CD spectroscopy.

EH	P_{EH}[a]	T_d [°C][b]
000	75	55.70 ± 0.23
001	72	42.10 ± 0.20
002	71	47.45 ± 0.31
003	69	45.90 ± 0.43
004	67	44.63 ± 0.19
006	66	58.57 ± 0.24
008	63	38.31 ± 0.44
009	61	36.10 ± 0.73
016	42	35.92 ± 0.69
021	36	35.31 ± 0.62
037	28	35.99 ± 0.20
043	27	39.98 ± 0.74

[a] Experimentally measured substrate promiscuity level of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

[b] Melting temperatures of EHs ± STD (n = 3) determined by CD spectroscopy **(see section 2.8)**.

Table S8: P_{EH}, log(P_{EH}), Act_{max}, and log(Act_{max}) of EHs.

EH[a]	P_{EH}[b]	log(P_{EH})	Act_{max} [U / (g wet cells)][c]	log(Act_{max}) [log(U / (g wet cells))]
001	72	1.86	1326.63	3.12
002	71	1.85	113.26	2.05
003	69	1.84	106.35	2.03
CalB	68	1.83	69105.06 [U / (g total protein)]	n.d.[d]
004	67	1.83	262.23	2.42
005	67	1.83	23.52	1.37
006	66	1.82	338.55	2.53
007	64	1.81	77.72	1.89
008	63	1.80	2239.16	3.35
009	61	1.79	168.40	2.23
010	58	1.76	77.55	1.89
011	53	1.72	120.02	2.08
012	51	1.71	137.81	2.14
013	49	1.69	278.08	2.44
014	48	1.68	138.13	2.14
015	42	1.62	93.93	1.97
016	42	1.62	991.93	3.00
017	39	1.59	7787.23	3.89
018	38	1.58	304.25	2.48
019	37	1.57	5038.96	3.70
020	37	1.57	35.12	1.55
021	36	1.56	963.46	2.98
CalA	36	1.56	25224.17 [U / (g total protein)]	n.d.[d]
022	35	1.54	1366.25	3.14
023	34	1.53	6005.66	3.78
024	34	1.53	123.42	2.09
025	33	1.52	1441.93	3.16
026	32	1.51	50.93	1.71
027	32	1.51	90.19	1.96
028	31	1.49	667.07	2.82
029	31	1.49	7660.87	3.88
030	30	1.48	752.11	2.88
031	29	1.46	398.62	2.60
032	29	1.46	242.56	2.38
033	29	1.46	376.69	2.58
034	29	1.46	1207.80	3.08
035	29	1.46	32.65	1.51
036	28	1.45	311.41	2.49
037	28	1.45	746.72	2.87
038	28	1.45	193.26	2.29
039	28	1.45	39.50	1.60
040	27	1.43	81.25	1.91

041	27	1.43	1198.35	3.08
042	27	1.43	139.81	2.15
043	27	1.43	571.48	2.76
044	25	1.40	101.91	2.01
045	24	1.38	143.52	2.16
046	23	1.36	20.79	1.32
047	23	1.36	274.96	2.44
048	23	1.36	148.60	2.17
049	23	1.36	9.92	1.00
050	22	1.34	661.50	2.82
051	22	1.34	278.89	2.45
052	21	1.32	252.19	2.40
053	21	1.32	90.48	1.96
054	21	1.32	2665.38	3.43
055	21	1.32	20.54	1.31
056	21	1.32	19.52	1.29
057	21	1.32	440.22	2.64
058	21	1.32	348.91	2.54
059	21	1.32	17.96	1.25
060	21	1.32	240.62	2.38
061	20	1.30	621.61	2.79
062	20	1.30	243.06	2.39
063	20	1.30	197.90	2.30
064	20	1.30	34.38	1.54
065	20	1.30	18.26	1.26
066	19	1.28	101.69	2.01
067	18	1.26	124.41	2.09
068	18	1.26	49.73	1.70
069	18	1.26	114.76	2.06
070	18	1.26	22.16	1.35
071	18	1.26	89.21	1.95
072	18	1.26	189.15	2.28
073	17	1.23	677.95	2.83
074	17	1.23	25.25	1.40
075	16	1.20	93.59	1.97
076	16	1.20	131.29	2.12
077	16	1.20	349.92	2.54
078	15	1.18	195.57	2.29
079	14	1.15	16.12	1.21
080	14	1.15	120.56	2.08
081	14	1.15	40.62	1.61
082	14	1.15	8978.87	3.95
083	14	1.15	155.38	2.19
084	13	1.11	273.06	2.44
085	13	1.11	69.15	1.84
086	13	1.11	25.46	1.41
087	13	1.11	11.45	1.06
088	13	1.11	4646.55	3.67
089	13	1.11	62.73	1.80

090	13	1.11	15.24	1.18
091	13	1.11	243.86	2.39
092	12	1.08	41.08	1.61
093	12	1.08	466.48	2.67
094	12	1.08	41.73	1.62
095	11	1.04	98.26	1.99
096	11	1.04	5.65	0.75
097	11	1.04	191.44	2.28
098	11	1.04	24.79	1.39
099	11	1.04	498.13	2.70
100	11	1.04	241.38	2.38
101	11	1.04	17.89	1.25
102	10	1.00	3328.23	3.52
103	10	1.00	91.17	1.96
104	10	1.00	56.63	1.75
105	10	1.00	45.59	1.66
106	9	0.95	11.56	1.06
107	9	0.95	16.33	1.21
108	9	0.95	159.45	2.20
109	9	0.95	17.65	1.25
110	8	0.90	332.72	2.52
111	8	0.90	13.97	1.15
112	8	0.90	312.09	2.49
113	8	0.90	11.00	1.04
114	8	0.90	13.15	1.12
115	8	0.90	148.37	2.17
116	7	0.85	19.84	1.30
117	6	0.78	4.22	0.63
118	6	0.78	29.01	1.46
119	6	0.78	25.83	1.41
120	5	0.70	9.15	0.96
121	5	0.70	131.87	2.12
122	5	0.70	3.35	0.53
123	4	0.60	8.15	0.91
124	4	0.60	21.63	1.34
125	4	0.60	6.31	0.80
126	4	0.60	4.65	0.67
127	4	0.60	4.59	0.66
128	4	0.60	11.63	1.07
129	3	0.48	7.32	0.86
130	2	0.30	23.86	1.38
131	2	0.30	4.16	0.62
132	2	0.30	4.86	0.69
133	2	0.30	3.67	0.56
134	2	0.30	1.73	0.24
135	2	0.30	3.94	0.60
136	2	0.30	3.29	0.52
137	2	0.30	3.32	0.52
138	2	0.30	2.83	0.45

139	2	0.30	4.16	0.62
140	1	0.00	1.73	0.24
141	1	0.00	2.55	0.41
142	1	0.00	0.25	-0.59
143	1	0.00	1.31	0.12
144	1	0.00	1.80	0.25
145	1	0.00	2.48	0.39

[a] EHs highlighted in bold constitute the *flexibility data set*; for underlined EHs, no crystal structure is known.

[b] Experimentally determined substrate promiscuity level of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

[c] Experimentally determined average maximum specific activities of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

[d] Not determined.

Table S9: Ester library classified according to TA.

Ester	TA
ϒ-Valerolactone	0
D-Pantolactone	0
L-Pantolactone	0
1-Napthyl acetate	2
Ethyl acetate	2
Methyl 3-hydroxybenzoate	2
Methyl 2-hydroxybenzoate	2
Methyl benzoate	2
Vinyl acetate	2
Methyl glycolate	2
(+)-Methyl D-Lactate	2
(-)-Methyl L-Lactate	2
Phenyl acetate	2
Glyceryl trilaurate	2
Ethyl propionate	3
Ethyl benzoate	3
(1*R*)-(-)-Menthyl acetate	3
(1*S*)-(+)-Menthyl acetate	3
Methyl (*R*)-(-)-mandelate	3
Methyl (*S*)-(+)-mandelate	3
(+)-Ethyl D-Lactate	3
(−)-Ethyl L-lactate	3
(+)-Methyl (*S*)-3-hydroxybutyrate	3
(-)-Methyl (*R*)-3-hydroxybutyrate	3
(1R)-(+)-Neomenthyl acetate	3
(1S)-(+)-Neomenthyl acetate	3
Methyl butyrate	3
Methyl 2,5-dihydroxycinnamate	3
Methyl cinnamate	3
Vinyl propionate	3
Vinyl benzoate	3
Vinyl crotonate	3
Vinyl acrylate	3
Ethyl 2-chlorobenzoate	3
2,4-Dichlorophenyl 2,4-dichlorobenzoate [DCPDCB]	3
Propyl acetate	3
Phenyl propionate	3
1-Naphthyl butyrate	4
Ethyl butyrate	4
Propylparaben	4
(−)-Methyl (*R*)-3-hydroxyvalerate	4
(+)-Methyl (*S*)-3-hydroxyvalerate	4

Benzylparaben	4
Propyl propionate	4
Methyl ferulate	4
Vinyl butyrate	4
3-Methyl-3-buten-1-yl acetate	4
Ethyl 2-methylacetoacetate	4
Ethyl acetoacetate	4
Cyclohexyl butyrate	4
2,4-Dichlorobenzyl 2,4-dichlorobenzoate [DCBDCB]	4
Butyl acetate	4
N-Benzyl-L-proline ethyl ester	5
N-Benzyl-D-proline ethyl ester	5
Ethyl (*R*)-(+)-4-chloro-3-hydroxybutyrate [E(R)CHB]	5
Ethyl (*S*)-(−)-4-chloro-3-hydroxybutyrate [E(S)CHB]	5
Benzoic acid 4-formyl-phenylmethyl ester [BFPME]	5
Butylparaben	5
Methyl hexanoate	5
Propyl butyrate	5
Isobutyl cinnamate	5
Ethyl 2-ethylacetoacetate	5
Ethyl propionylacetate	5
Hexyl acetate	6
Ethyl hexanoate	6
Phthalic acid diethyl ester	6
Benzyl (*R*)-(+)-2-hydroxy-3-phenylpropionate [BHPP]	6
Phenylethyl cinnamate	6
Geranyl acetate	6
Ethyl 3-oxohexanoate	6
n-Pentyl benzoate	6
Methyl octanoate	7
Propyl hexanoate	7
Diethyl-2,6-dimethyl 4-phenyl-1,4-dihydro pyridine-3,5-dicarboxylate [DDPDPDC]	7
Glyceryl triacetate	8
Octyl acetate	8
Ethyl octanoate	8
Methyl decanoate	9
(1*R*)-(-)-dimenthyl succinate	9
Ethyl decanoate	10
Glyceryl tripropionate	11
Methyl dodecanoate	11
Dodecanoyl acetate	12
Ethyl dodecanoate	12
Vinyl laurate	12
Methyl myristate	13
Glyceryl tributyrate	14

Ethyl myristate	14
Vinyl myristate	14
Pentadecyl acetate	15
Glucose pentaacetate	15
Methyl oleate	16
Vinyl palmitate	16
Vinyl oleate	17
Glyceryl trioctanoate	26
Triolein	54

Table S10: Distribution of P_{EH} in F_{EH} of the *experimental data set*.

EH	F_{EH} [a]	P_{EH} [b]
026		32
040		27
041		27
046		23
071		18
072		18
075		16
077		16
090		13
097	F_{I}	11
101		11
108		9
110		8
113		8
115		8
118		6
131		2
132		2
142		1
145		1
051		22
073		17
088		13
098		11
102	F_{II}	10
116		7
136		2
138		2
140		1
001		72
002		71
003		69
004		67
005		67
006		66
008		63
009		61
010	F_{IV}	58
011		53
012		51
013		49
014		48
015		42
016		42
018		38
021		36
022		35

023		34
025		33
029		31
035		29
037		28
039		28
042		27
043		27
048		23
052		21
054		21
067		18
079		14
086		13
087		13
091		13
099		11
119		6
028		31
031		29
032		29
045		24
047		23
049		23
053		21
055		21
056		21
057		21
058		21
065		20
066		19
068		18
074		17
076		16
081	F_V	14
082		14
083		14
092		12
094		12
096		11
010		11
103		10
104		10
107		9
109		9
111		8
114		8
120		5
123		4
127		4
128		4

030	FVI	30
034		29
059		21
061		20
085		13
020	F_{VII}	37
064		20
084		13
093		12
112		8
139		2
007	FVIII (serine beta-lactamase like)	64
024		34
027		32
069		18
078		15
089		13
095		11
124		4
133		2
141		1
129		3
044	CE (carbohydrate esterase like)	25
070		18
126		4
134		2
135		2
137		2
017	C-C MCPh	39
019		37
033		29
036		28
038		28
050		22
060		21
062		20
063		20
105	Cyclase-like esterase	10
CalB	Yeast class	68
CalA		36
080	Unclassified	14
106		9
117		6
121		5
122		5
125		4
130		2
144		1
143		1

[a] EH families based on the Arpigny and Jaeger classification(27) **(see section 2.1)**.

[b] Experimentally measured substrate promiscuity level of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

Table S11: P_{EH} and Vol_{eff} of comparative models of EHs of the *flexibility data set* with known crystal structures.

EH	P_{EH}[a]	Vol_{eff} [Å^3][b]
001	72	166.667
CalB	68	200.000
CalA	36	1000.000
023	34	90.909
037	28	166.667
060	21	250.000
096	11	34.483
102	10	38.462
105	10	n.d.[c]
107	9	28.571
115	8	71.429

[a] Experimentally determined substrate promiscuity level of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

[b] Computed active site effective volumes of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

[c] Not determined.

Table S12: P_{EH} and Vol_{eff} of comparative models of EHs of the *flexibility data set* without known crystal structures.

EH	P_{EH}[a]	Vol_{eff} [Å^3][b]
005	67	200.000
010	58	200.000
011	53	83.333
012	51	333.333
013	49	333.333
014	48	200.000
015	42	166.667
016	42	333.333
029	31	500.000
030	30	66.667
033	29	166.667
034	29	32.258
042	27	200.000
043	27	66.667
048	23	111.111
061	20	111.111
070	18	43.478
074	17	58.824
083	14	58.824
092	12	41.667
093	12	37.037
094	12	45.455
125	4	19.231
127	4	55.556

[a] Experimentally determined substrate promiscuity level of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

[b] Computed active site effective volumes of EHs provided by Ferrer *et al.* (1) **(see section 2.1)**.

[c] Not determined.

Supplemental Figures

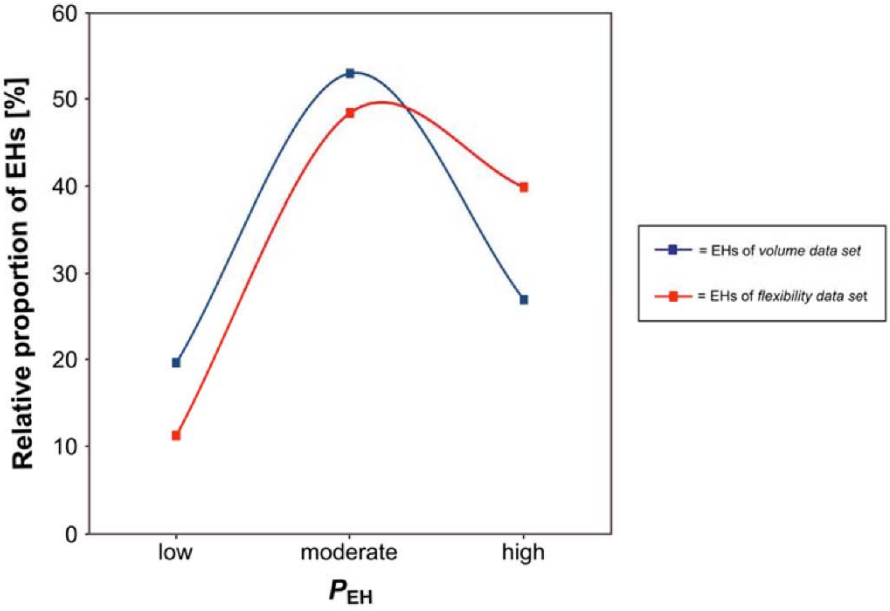


Figure S1: Comparison between the *volume data set* and the *flexibility data set* regarding P_{EH}. Relative proportions of EHs constituting the *volume data set* (red line) and the *flexibility data set* (blue line) regarding P_{EH} determined with a kinetic pH indicator assay (2-4) by Ferrer *et al.* (1). P_{EH} is defined as *low* if the EH hydrolyzes ≤ 9 esters, as *moderate* if the EH hydrolyzes between 10 and 29 esters, and as *high* if the EH hydrolyzes ≥ 30 esters.

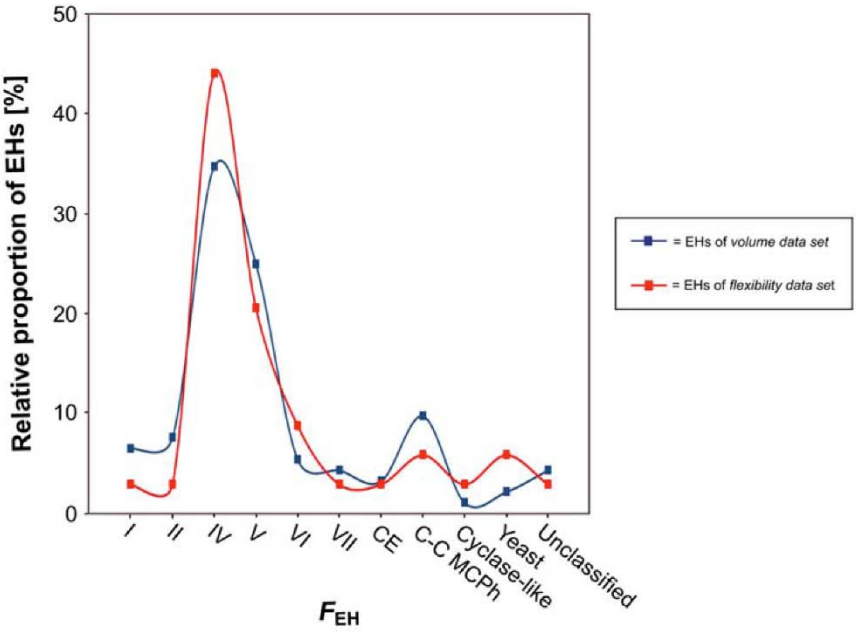


Figure S2: Comparison between the *volume data set* and the *flexibility data set* regarding F_{EH}. Relative proportions of EHs constituting the *volume data set* (red line) and the *flexibility data set* (blue line) regarding F_{EH} based on the Arpigny and Jaeger classification (27).

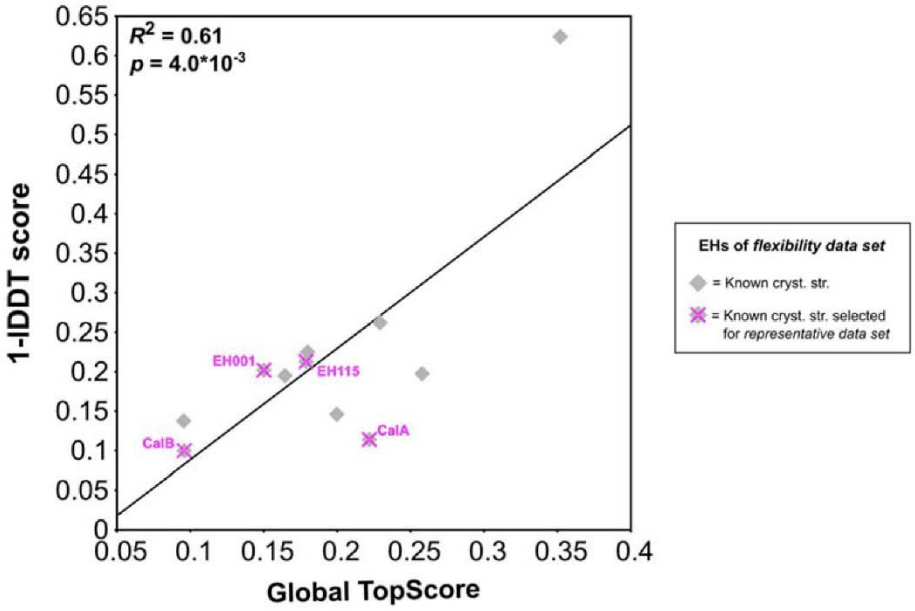


Figure S3: TopScore performance on comparative models of EHs of the *flexibility data set* with known crystal structures. Correlation between 1 - lDDT scores and global TopScores for comparative models of EHs of the *flexibility data set* with known crystal structures. The comparative models were generated by TopModel (28) (excluding the known crystal structures as templates) and evaluated by TopScore (8). The lDDT scores were computed by the lDDT web server from Swiss-Model (9) from comparisons of the comparative models of these EHs against the known crystal structures as experimental references. The EHs belonging to the *representative data set* are indicated by magenta crosses.

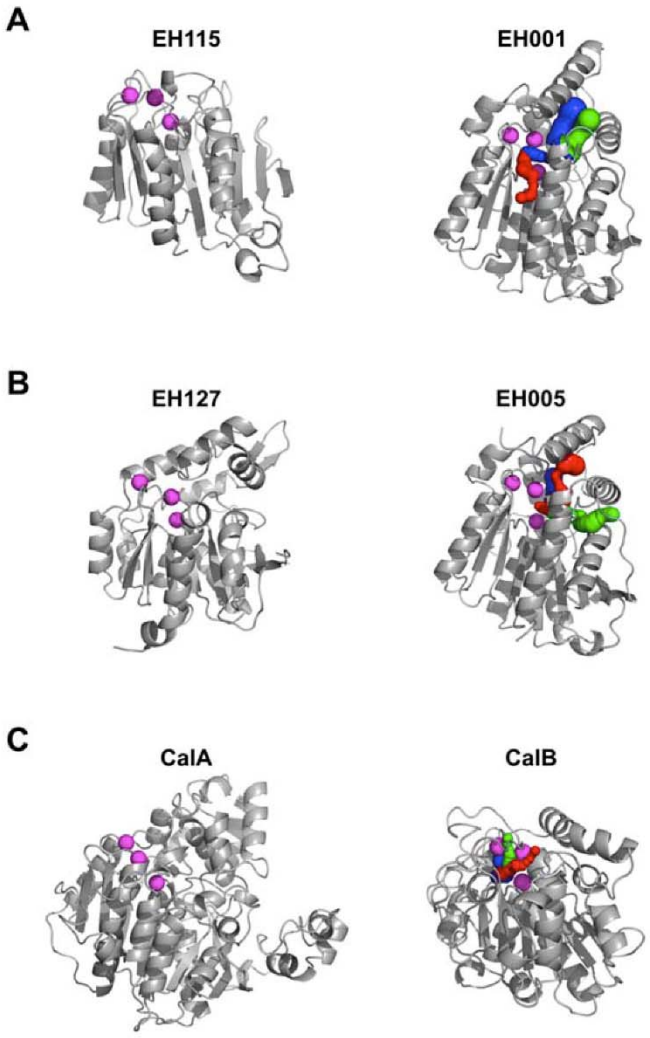


Figure S4: Substrate-accessibility of EHs of the *representative data set.* CAVER results (29) of comparative models of **(A)** EHs with known crystal structures and lowest (EH115) or highest P_{EH} (EH001), **(B)** EHs with unknown crystal structures and lowest (EH127) or highest P_{EH} (EH005), and **(C)** commercial EHs with lowest (CalA) or highest P_{EH} (CalB). CARs (magenta spheres) are either located on the protein surface or are buried and connected with the surface by tunnels (blue, green, and red spheres).

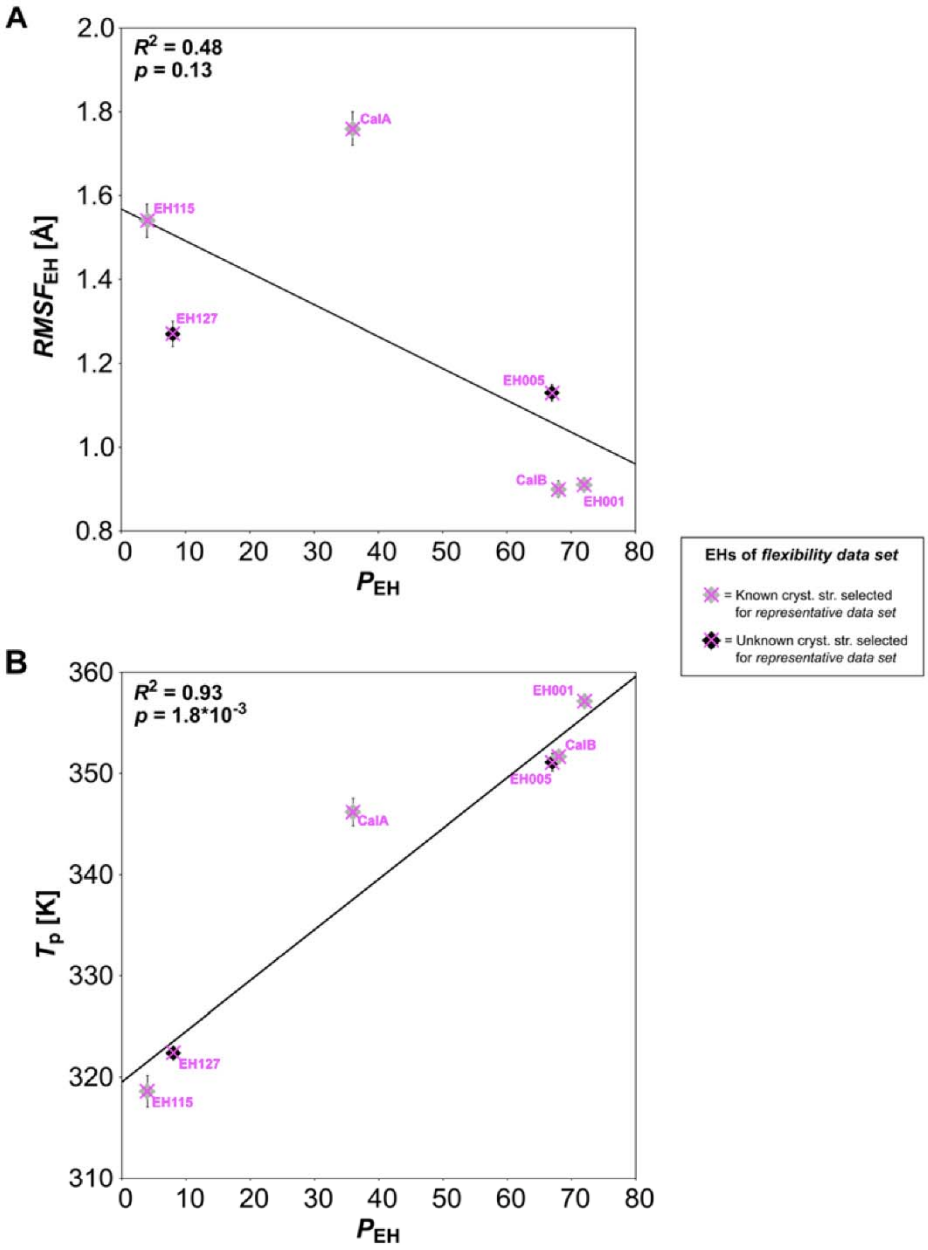


Figure S5: Correlation of $RMSF_{EH}$ or T_p *versus* P_{EH} of the *representative data set*. (A) Correlation between $RMSF_{EH}$ based on the MD trajectories and P_{EH} of the *representative data set*. **(B)** Correlation between T_p based on the *global* index H_{type2} (17) computed by CNA and P_{EH} of the *representative data set*. Data points colored grey (black) and indicated by magenta crosses represent comparative models of EHs with (un)known crystal structures. Error bars show the SEM over five independent MD simulations of 1 μs length each.

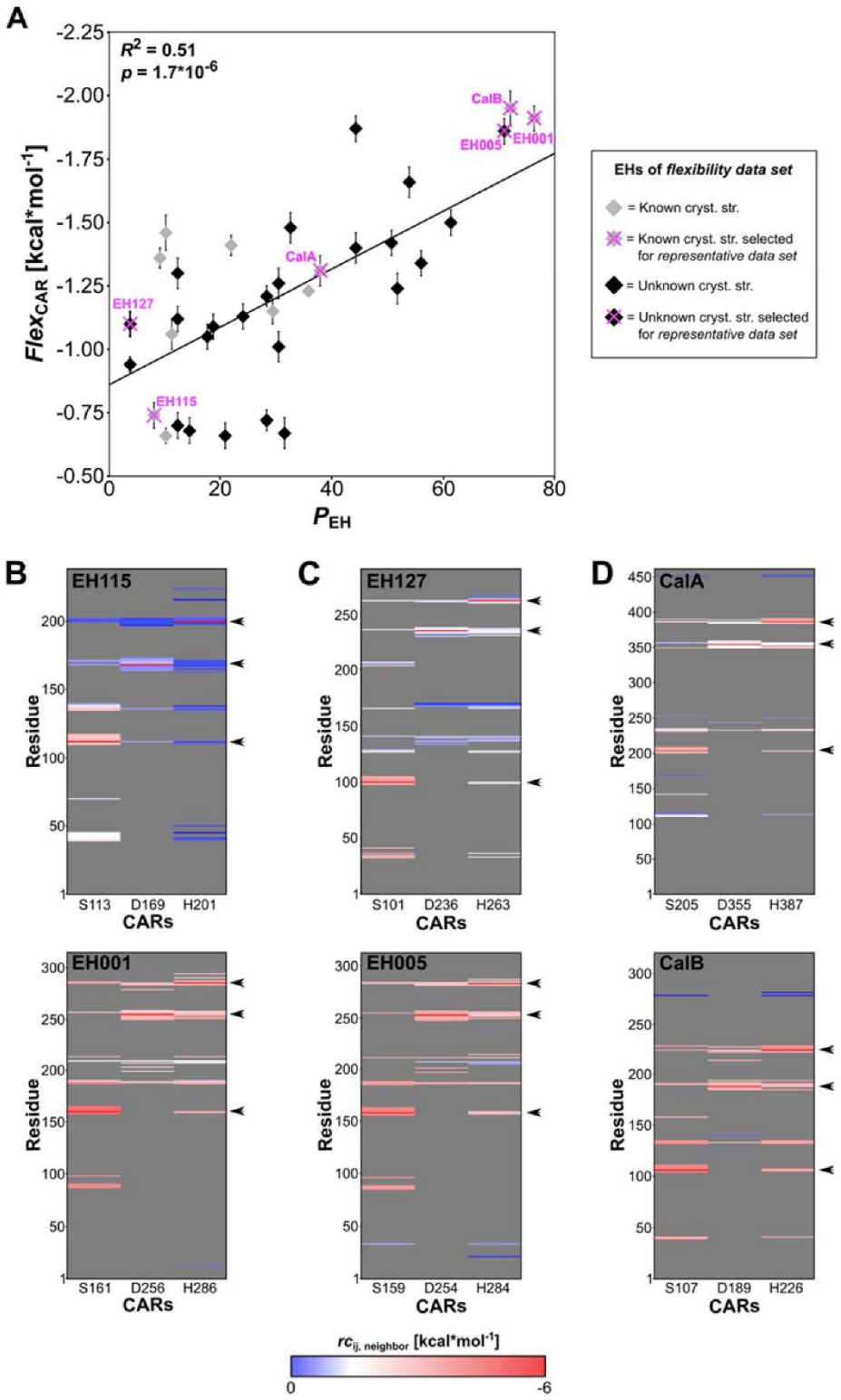


Figure S6: Correlation of *Flex*$_{CAR}$ *versus* P_{EH}. (A) Correlation between predicted *Flex*$_{CAR}$ based on the local index $rc_{ij;neighbor}$ and P_{EH} for the *flexibility data set*. Data points colored grey (black) represent homology models of EHs with (un)known crystal structures. The

representative data set is indicated by magenta crosses. Error bars show the SEM over five independent MD simulations of 1 μs length each.. $rc_{ij;neighbor}$ of CARs of **(B)** EHs with known crystal structures and lowest (EH115) or highest P_{EH} (EH001), **(C)** EHs with unknown crystal structures and lowest (EH127) or highest P_{EH} (EH005), and **(D)** commercial EHs with lowest (CalA) or highest P_{EH} (CalB). A red (blue) color indicates that a rigid contact between CARs and other residues within 5 Å distance is more (less) stable (see color scale at the bottom). The rigid contacts for all other residue pairs are colored grey. Black arrow heads indicate positions of CARs.

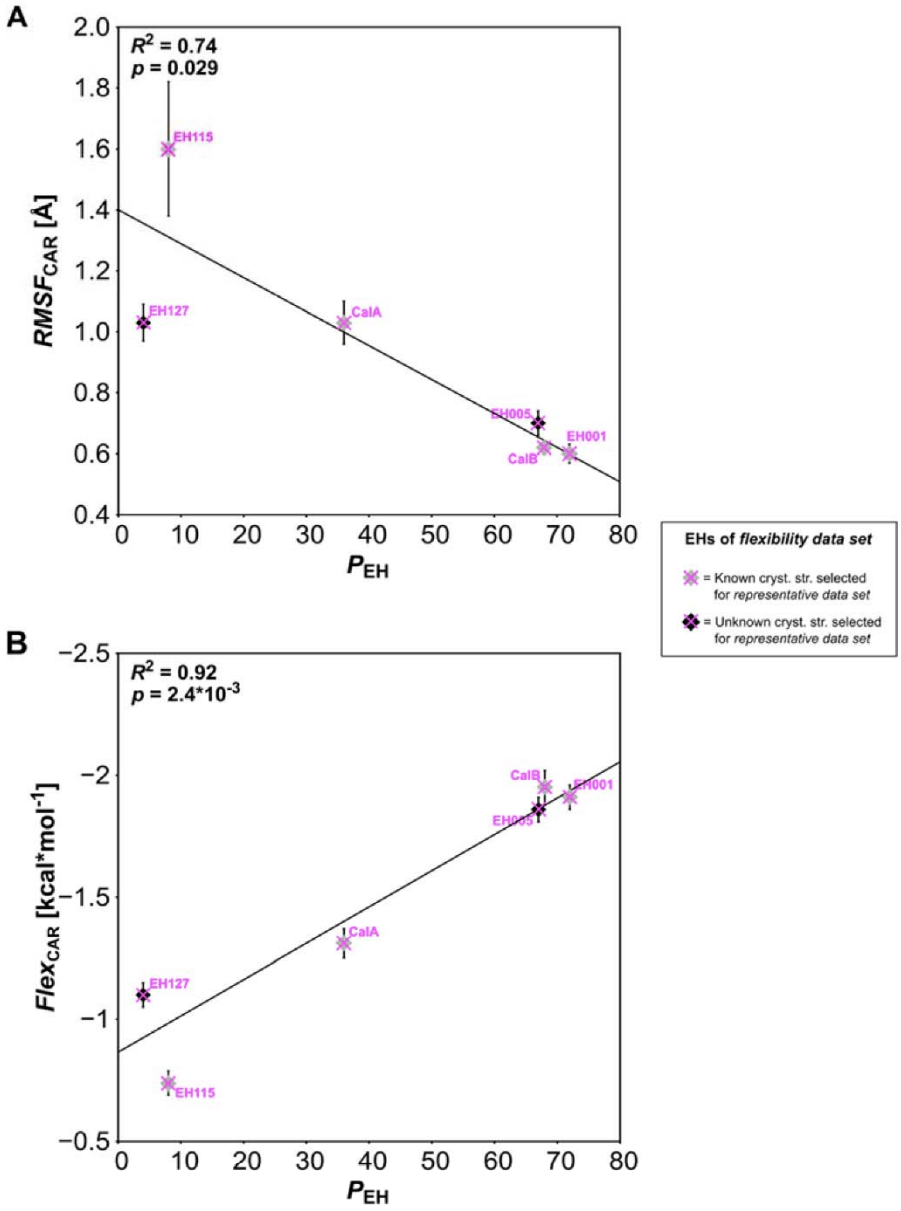


Figure S7: Correlation of *RMSF*$_{CAR}$ or *Flex*$_{CAR}$ *versus* P_{EH} of the *representative data set*. **(A)** Correlation between *RMSF*$_{CAR}$ based on the MD trajectories and P_{EH} of the *representative data set*. **(B)** Correlation between *Flex*$_{CAR}$ based on CNA and P_{EH} of the *representative data set*. Data points colored grey (black) and indicated by magenta crosses represent comparative models of EHs with (un)known crystal structures. Error bars show the SEM over five independent MD simulations of 1 μs length each.

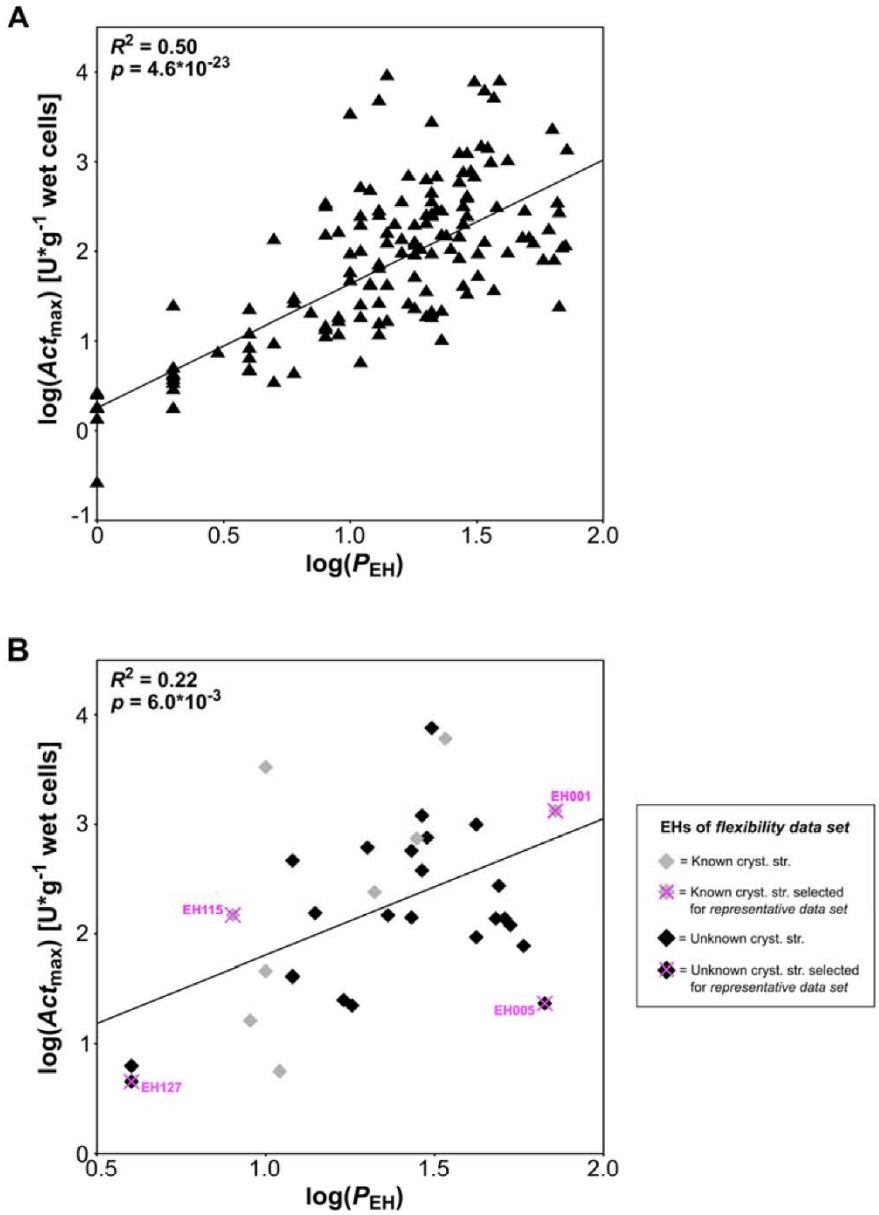


Figure S8: Correlation of log(Act_{max}) *versus* log(P_{EH}). Correlation between log(Act_{max}) and log(P_{EH}) for **(A)** the *experimental data set* and **(B)** the *flexibility data set* containing EHs with known crystal structures (grey data points), EHs with unknown crystal structures (black data points), and EHs constituting the *representative data set* (magenta crosses). The EHs were screened against 96 different esters in a kinetic pH indicator assay (2-4) that provided Act_{max} given in U (g wet cells)$^{-1}$. CalA and CalB preparations were excluded because Act_{max} was given in U (g total protein)$^{-1}$. The assays were performed as triplicates with STD ≤ 1%.

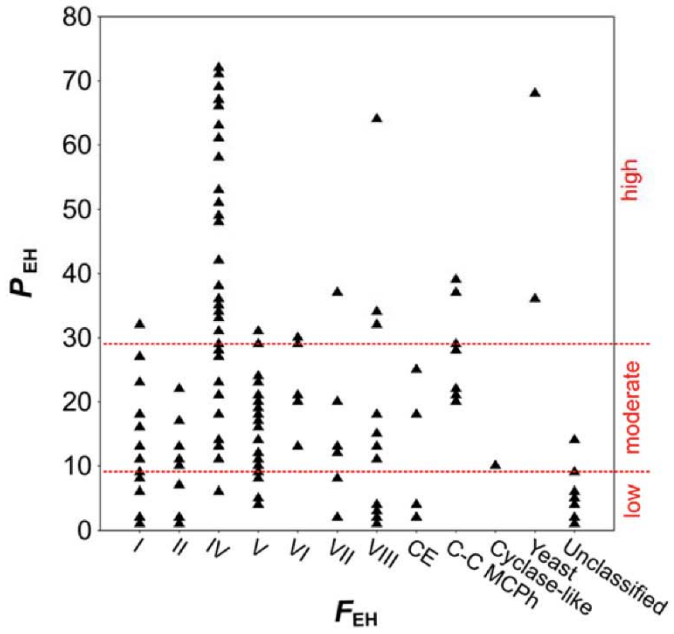


Figure S9: Distribution of P_{EH} in F_{EH} of the *experimental data set*. Distribution of P_{EH} determined with a kinetic pH indicator assay (2-4) by Ferrer *et al.* (1) in F_{EH} based on the Arpigny and Jaeger classification (27) of the *experimental data set*. P_{EH} is defined as *low* if the EH hydrolyzes ≤ 9 esters, as *moderate* if the EH hydrolyzes between 10 and 29 esters, and as *high* if the EH hydrolyzes ≥ 30 esters.

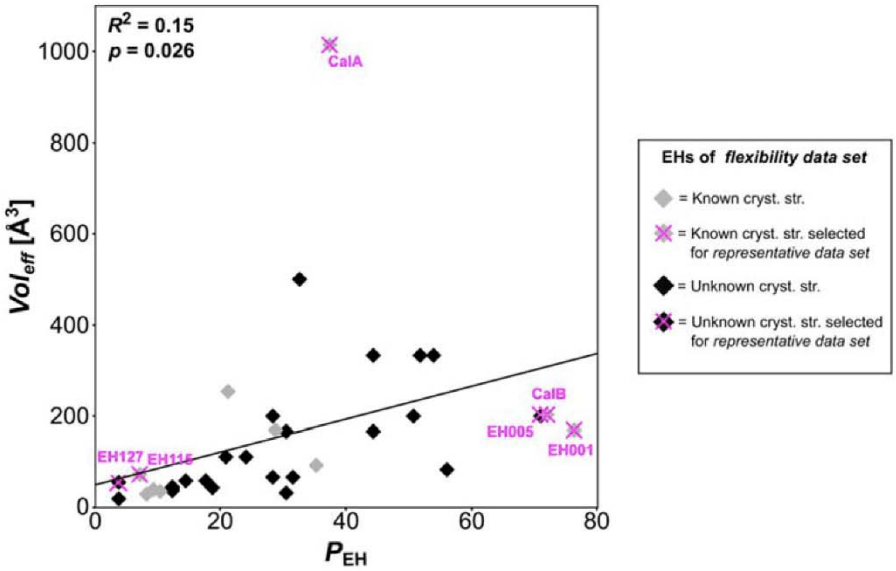


Figure S10: Correlation of Vol_{eff} *versus* P_{EH}. Correlation between Vol_{eff} and P_{EH} of the *flexibility data set*. Vol_{eff} represents the topology of the catalytic environment in terms of the active site cavity volume (Vol_{cav}) computed by Fpocket (5) per relative solvent-accessible surface area ($SASA_{rel}$) computed by GetArea webserver (6). Data points colored grey (black) represent comparative models of EHs with (un)known crystal structures. The *representative data set* is indicated by magenta crosses.

Supplemental References

1. M. Martinez-Martinez *et al.*, Determinants and prediction of esterase substrate promiscuity patterns. *ACS Chem. Biol.* **13**, 225-234 (2017).

2. M. Alcaide *et al.*, Single residues dictate the co-evolution of dual esterases: MCP hydrolases from the α/β hydrolase family. *Biochem. J.* **454**, 157-166 (2013).

3. L. E. Janes, A. C. Löwendahl, R. J. Kazlauskas, Quantitative screening of hydrolase libraries using pH indicators: identifying active and enantioselective hydrolases. *Chem. Eur.* **4**, 2324-2331 (1998).

4. M. Martinez-Martinez *et al.*, Biochemical diversity of carboxyl esterases and lipases from Lake Arreo (Spain): a metagenomic approach. *Appl. Environ. Microbiol.* **79**, 3553-3562 (2013).

5. V. Le Guilloux, P. Schmidtke, P. Tuffery, Fpocket: an open source platform for ligand pocket detection. *BMC Bioinform.* **10**, 168 (2009).

6. R. Fraczkiewicz, W. Braun, Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* **19**, 319-333 (1998).

7. (2016) Schrödinger. in *Maestro-Desmond Interoperability Tools*, D. E. Shaw Research (New York).

8. D. Mulnaes, H. Gohlke, TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment. *J. Chem. Theory Comput.* **14**, 6117-6126 (2018).

9. V. Mariani, M. Biasini, A. Barbato, T. Schwede, lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722-2728 (2013).

10. J. C. Shelley *et al.*, Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* **21**, 681-691 (2007).

11. J. A. Maier *et al.*, ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696-3713 (2015).

12. S. Izadi, R. Anandakrishnan, A. V. Onufriev, Building water models: a different approach. *J. Phys. Chem. Lett.* **5**, 3863-3871 (2014).

13. D. A. Case *et al.* (2019) AMBER 2019. (University of California, San Francisco).

14. I. S. Joung, T. E. Cheatham III, Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B.* **112**, 9020-9041 (2008).

15. B. Frieg *et al.*, Molecular mechanisms of glutamine synthetase mutations that lead to clinically relevant pathologies. *PLoS Comput. Biol.* **12**, e1004693 (2016).

16. B. I. Dahiyat, D. B. Gordon, S. L. Mayo, Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333-1337 (1997).

17. S. Radestock, H. Gohlke, Exploiting the link between protein rigidity and thermostability for data□driven protein engineering. *Eng. Life Sci.* **8**, 507-522 (2008).

18. S. Radestock, H. Gohlke, Protein rigidity and thermophilic adaptation. *Proteins* **79**, 1089-1108 (2011).

19. P. L. Privalov, S. J. Gill, Stability of protein structure and hydrophobic interaction. *Adv. Protein Chem.* **39**, 191-234 (1988).

20. P. C. Rathi, S. Radestock, H. Gohlke, Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J. Biotechnol.* **159**, 135-144 (2012).

21. P. C. Rathi, K.-E. Jaeger, H. Gohlke, Structural rigidity and protein thermostability in variants of lipase A from Bacillus subtilis. *PloS One* **10**, e0130289 (2015).

22. M. Dick *et al.*, Trading off stability against activity in extremophilic aldolases. *Sci. Rep.* **6**, 17908 (2016).

23. P. C. Rathi, A. Fulton, K.-E. Jaeger, H. Gohlke, Application of rigidity theory to the thermostabilization of lipase A from Bacillus subtilis. *PLoS Comput. Biol.* **12**, e1004754 (2016).

24. C. Nutschel *et al.*, Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for Bacillus subtilis lipase A. *J. Chem. Inf. Model.* **60**, 1568-1584 (2020).

25. P. C. Rathi, D. Mulnaes, H. Gohlke, VisualCNA: a GUI for interactive constraint network analysis and protein engineering for improving thermostability. *Bioinformatics* **31**, 2394-2396 (2015).

26. C. Pfleger, S. Radestock, E. Schmidt, H. Gohlke, Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* **34**, 220-233 (2013).

27. J. L. Arpigny, K.-E. Jaeger, Bacterial lipolytic enzymes: classification and properties. *Biochem. J.* **343**, 177-183 (1999).

28. D. Mulnaes *et al.*, TopModel: Template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks. *J. Chem. Theory. Comput.* **16**, 1953-1967 (2020).

29. E. Chovancova *et al.*, CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.* **8**, e1002708 (2012).

ORIGINAL PUBLICATION IV

Contribution of single amino acid and codon substitutions to the production and secretion of a lipase by *Bacillus subtilis*

Skoczinski, P., Volkenborn, K., Fulton, A., Bhadauriya, A., Nutschel, C., Gohlke, H., Knapp, A., Jaeger, K.-E.

Microb. Cell Fact. 2017, 16, 160.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5613506/

Skoczinski *et al. Microb Cell Fact (2017) 16:160*
DOI 10.1186/s12934-017-0772-z

Microbial Cell Factories

RESEARCH

Open Access

Contribution of single amino acid and codon substitutions to the production and secretion of a lipase by *Bacillus subtilis*

Pia Skoczinski[1,5], Kristina Volkenborn[1], Alexander Fulton[1,6], Anuseema Bhadauriya[2], Christina Nutschel[2], Holger Gohlke[2,3], Andreas Knapp[1] and Karl-Erich Jaeger[1,4*]

Abstract

Background: *Bacillus subtilis* produces and secretes proteins in amounts of up to 20 g/l under optimal conditions. However, protein production can be challenging if transcription and cotranslational secretion are negatively affected, or the target protein is degraded by extracellular proteases. This study aims at elucidating the influence of a target protein on its own production by a systematic mutational analysis of the homologous *B. subtilis* model protein lipase A (LipA). We have covered the full natural diversity of single amino acid substitutions at 155 positions of LipA by site saturation mutagenesis excluding only highly conserved residues and qualitatively and quantitatively screened about 30,000 clones for extracellular LipA production. Identified variants with beneficial effects on production were sequenced and analyzed regarding *B. subtilis* growth behavior, extracellular lipase activity and amount as well as changes in lipase transcript levels.

Results: In total, 26 LipA variants were identified showing an up to twofold increase in either amount or activity of extracellular lipase. These variants harbor single amino acid or codon substitutions that did not substantially affect *B. subtilis* growth. Subsequent exemplary combination of beneficial single amino acid substitutions revealed an additive effect solely at the level of extracellular lipase amount; however, lipase amount and activity could not be increased simultaneously.

Conclusions: Single amino acid and codon substitutions can affect LipA secretion and production by *B. subtilis*. Several codon-related effects were observed that either enhance *lipA* transcription or promote a more efficient folding of LipA. Single amino acid substitutions could improve LipA production by increasing its secretion or stability in the culture supernatant. Our findings indicate that optimization of the expression system is not sufficient for efficient protein production in *B. subtilis*. The sequence of the target protein should also be considered as an optimization target for successful protein production. Our results further suggest that variants with improved properties might be identified much faster and easier if mutagenesis is prioritized towards elements that contribute to enzymatic activity or structural integrity.

Keywords: *Bacillus subtilis*, Lipase, Protein production, Secretion, Optimization

Background

The Gram-positive soil bacterium *Bacillus subtilis* secretes up to 20 g/l of produced proteins directly into the culture supernatant [1, 2]. Therefore, it has become more and more important in industrial applications for the production of homologous and heterologous proteins in large-scale fermentation processes [1]. Due to this fact, *B. subtilis* has been intensively studied and optimized as a protein production host in the last decades, establishing it as a 'microbial cell factory' [3, 4].

Optimization strategies have targeted several bottlenecks for heterologous protein production in *B. subtilis*.

*Correspondence: karl-erich.jaeger@fz-juelich.de
[4] Institute of Bio- and Geosciences IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
Full list of author information is available at the end of the article

Examples include optimization of transcription efficiency by using strong promoters such as the constitutive promoter P_{aprE} or an arabinose-inducible promoter [3]. Fine-tuning of translation [5] can be achieved by either using optimized ribosome binding sites to improve ribosome binding of the mRNA [3] or by introducing translational pauses using 'slow-translating' codons, as previously shown for heterologous protein production in *E. coli* [5, 6].

The majority of secretory proteins in *B. subtilis* are targeted to the Sec translocon and translocated via the cotranslational Sec-SRP pathway [7–10]. To optimize the protein secretion step as a prospective bottleneck, several studies assayed for the optimal signal peptide necessary for secretion. Screening a set of 173 Sec-specific signal peptides of *B. subtilis* [11] or the additional screening of heterologous signal peptides from *B. licheniformis* [12] successfully identified signal peptides for improved secretion of the *Fusarium solani pisii* cutinase [11] and the *B. amyloliquefaciens* subtilisin BPN' [13] in *B. subtilis*. Maturation and folding of secreted proteins are increased by the overexpression of regulatory factors, e.g. the lipoprotein PrsA, which resulted in increased secretion rates of α-amylase of *B. stearothermophilus* by *B. subtilis* [14]. Furthermore, strains lacking the majority of the major extracellular proteases have been constructed, e.g. the *B. subtilis* strain WB800 lacking all eight extracellular proteases (AprE, NprE, NprB, Vpr, Bpr, Mpr, Epr, WprA), resulting in strongly decreased degradation of extracellular target proteins [2]. A few studies with Gram-negative bacteria indicated that the target protein itself can also influence its production and secretion, e.g. by interactions with the translocation machinery [15, 16]. However, no systematic study has yet been reported on the role of each amino acid of a secreted protein for its production and secretion. Here, we have systematically analyzed single amino acids and their respective codons of *B. subtilis* lipase A (LipA) to understand beneficial and detrimental effects of amino acid and codon substitutions on LipA production and secretion.

The extracellular lipase LipA is one of the smallest known lipases showing a minimal α/β-hydrolase fold consisting of six β-sheets and six α-helices [17]. Compared to the classical α/β-hydrolase fold, two β-sheets are missing, the αD-helix is substituted by a small 3_{10}-helix, and the αE-helix contains only four amino acids [17]. LipA features a surface-exposed active site consisting of amino acids S77, D133 and H156, which is accessible for the substrate without conformational change; the oxyanion hole is formed by I12 and M78 [17, 18]. LipA is secreted cotranslationally via the Sec-SRP pathway. The N-terminal signal peptide is cleaved off by a signal peptidase resulting in the mature enzyme with 181 amino acids and a molecular weight of 19.34 kDa [8, 19].

LipA was subjected to a nearly complete site saturation mutagenesis targeting 155 of 181 residues with a conservation < 95% within the *Firmicutes* phylum. The resulting library was screened for extracellular lipase production both qualitatively and quantitatively. Our results indicate that both single amino acid and codon substitutions significantly affect production and secretion of the target protein and suggest that optimization studies should aim primarily at structural elements that contribute to enzymatic activity or structural integrity.

Methods

Bacterial strains and plasmids

Bacterial strains and plasmids used in this study are listed in Table 1. *E. coli* DH5α was used for cloning and plasmid amplification. *B. subtilis* TEB1030 was used as the secretory expression host.

Growth of *B. subtilis*

Escherichia coli and *B. subtilis* were grown in LB medium (10 g/l tryptone, 10 g/l NaCl, 5 g/l yeast extract) with 100 µg/ml ampicillin or 50 µg/ml kanamycin,

Table 1 Bacterial strains and plasmids

Bacterial strains and plasmids	Genotype	References
Bacterial strains		
E. coli DH5α	*supE44 Δ(lacZYA-argF) U196 (phi80ΔlacZM15) hsdR17 recA1 endA1 gyrA96 thi-1 relA1*	[20]
B. subtilis TEB1030	*trpC2 his nprE aprE bpf ispI lipA lipB*	[19]
Plasmids		
pBSMul1	*E. coli–B. subtilis* shuttle vector, ribosome binding site, P_{Hpall}, secretion (*sslipA*) and purification (C-terminal 6x-His-tag); ColE1 *repB* Km^r Amp^r	[21]
pET22lipA	pET22b (Novagen, USA) containing a 557 bp *Eco*RV/*Sac*I fragment of *B. subtilis lipA* gene fused to *pelB* signal peptide sequence, P_{T7lac}	[18]
pBSlipA	pBSMul1 containing a 568 bp *Eco*RI/*Hin*dIII fragment of *B. subtilis lipA* gene; additionally deleted *Eco*RI restriction site	This study

respectively, at 37 °C. Culture volumes, agitation speed and preparation of supernatants at different cultivation conditions are described below.

96-well microtiter plate cultivation

For the two-step screening procedure, *B. subtilis* was pre-cultivated in 150 µl LB medium in 96-well microtiter plates (Greiner Bio-one, Germany) at 37 °C, 900 rpm for 6 h (TiMix 5, Edmund Bühler GmbH, Germany). These pre-cultures were used to inoculate expression cultures in 150 µl fresh LB medium in 96-well microtiter plates (Greiner Bio-one, Germany) to an $O.D._{580nm}$ of 0.05 with a TECAN® robotic system freedom evo (Tecan Group Ltd., Germany). Expression cultures were cultivated at 25 °C, 900 rpm for 16 h (TiMix 5, Edmund Bühler GmbH, Germany). The cells were harvested by centrifugation (4 °C, 5000 × *g*, 30 min) and the culture supernatant was immediately used for analysis.

Microfermentation in 48-well FlowerPlate® and online biomass measurement

Bacillus subtilis clones were pre-cultivated in 1100 µl LB medium in 48-well Flowerplates (FlowerPlate® 48 well MTP without optodes, m2p-labs, Germany) at 37 °C, 1100 rpm for 16 h (TiMix 5, Edmund Bühler GmbH, Germany). Expression cultures were inoculated to an $O.D._{580nm}$ of 0.05 in 1100 µl LB medium in 48-well Flowerplates and cultivated at 37 °C, 1100 rpm for 6 h. For cell harvest, 50 µl of each culture were transferred into a 96-well microtiter plate (Greiner Bio-one, Germany) and centrifuged as described above.

Transformation of *E. coli* and *B. subtilis*

Electrocompetent *E. coli* DH5α cells were prepared as previously described [22]. *E. coli* DH5α was transformed by electroporation in a MicroPulser (BioRad, Germany). *B. subtilis* TEB1030 cells were transformed by protoplast formation as previously described [23].

Construction of the *lipA* expression vector pBSlipA, site saturation mutagenesis and library construction

The *lipA* gene (KEGG Accession Number BSU02700) without its native signal sequence was amplified from the *E. coli* expression vector pET22lipA [18] using the oligonucleotides *Eco*RI_fw (5′ cgcggaattcgctgaacac 3′) and *Hin*dIII_rev (5′ agtgcggccgcaagcttgtcgacgtaatgttcattaattcgtatt 3′). The resulting 568 bp *Eco*RI/*Hin*dIII fragment was cloned in frame with the native *lipA* signal sequence (*sslipA*) under the control of the strong constitutive promoter $P_{Hpa\text{II}}$ in the *E. coli–B. subtilis* shuttle vector pBSMul1 [21] previously used for analysis of secretory protein production [11, 13]. The additional six base pair linker of the *Eco*RI restriction site between the *sslipA* and the *lipA* gene was subsequently deleted by QuikChange® PCR [24] using the primer pair Δ*Eco*RI_fw (5′ agcaaaagccgctgaacacaatc 3′) and Δ*Eco*RI_rev (5′ gattgtgttcagcggcttttgct 3′). The generated expression vector pBSlipA harbors a native full-length *lipA* gene and was used for *lipA* expression and mutagenesis.

Oligonucleotide design and site saturation PCR were performed as previously described [25]. In short, the vector was amplified with degenerated 'NNS' oligonucleotides (Additional file 1: Table S1) by QuikChange® PCR [24]. The remaining template vector DNA in the PCR product was hydrolyzed using *Dpn*I, and the site saturation PCR product was desalted and concentrated by PCR Purification Kit (Analytik Jena, Germany). First, *E. coli* DH5α was transformed by electroporation, and the mutagenesis vectors were isolated from 2000 to 4000 *E. coli* clones. Subsequently, the secretory protein production strain *B. subtilis* TEB1030 was transformed with 20 ng of vector DNA by protoplast formation.

To achieve a library coverage of about 99.9%, 192 clones are necessary for each position, i.e. six-times the number of codons (32 via 'NNS') as described in [26]. Thus, a library for the site saturation mutagenesis of a certain position was distributed to two 96-well plates. However, we reduced the clone number to 184 *B. subtilis* TEB1030 transformants allowing to add 8 wild-types and negative controls. Taking into account that mutagenesis could also re-introduce the wild-type codon, a set of 184 transformants per residue leads to a full coverage probability of 93.87% calculated with TopLib (http://stat.haifa.ac.il/~yuval/toplib/) [27] and a supposed mutagenesis yield of 90%.

Double mutants were constructed by site directed PCR following the procedure described above for site saturation PCR. Oligonucleotides for site directed mutagenesis are listed in Additional file 1: Table S2.

Lipase activity assay with *B. subtilis* culture supernatant

Extracellular lipase activity was determined in 96-well microtiter plates (Greiner Bio-one, Germany). The *B. subtilis* culture supernatant obtained by centrifugation was mixed with *para*-nitrophenyl palmitate (*p*NPP) substrate solution as previously described [11], and hydrolysis of *p*NPP was measured spectrophotometrically ($\lambda_{abs} = 410$ nm) at 37 °C for 15 min using the plate reader SpectraMax 250 (Molecular Devices, Germany). Lipolytic volume activity was calculated using a molar extinction coefficient of 15,000 M^{-1} cm^{-1}. Specific lipase activity (U/mg) was calculated by the volume activity (U/ml) per protein amount (mg/ml). The LipA protein amount was quantified as described in the next paragraph. Unless stated otherwise, a two-tailed t-test was

performed with a significance level of $p < 0.05$ to determine significant activity changes.

Enzyme-linked immunosorbent assay with *B. subtilis* culture supernatant

For quantitative detection of extracellular LipA protein, an enzyme-linked immunosorbent assay (ELISA) using a specific polyclonal LipA antibody (Eurogentec, Germany) was performed. 15.6 µl twofold prediluted *B. subtilis* culture supernatant obtained by centrifugation was diluted in 100 µl bicarbonate buffer (100 mM; pH 9.6) and transferred into Polysorp® 96-well microtiter plates (Nunc-Immuno™ MicroWell™ 96-Well Plate) using the TECAN® robot system. After coating of proteins onto the plastic surface at 4 °C, 100 rpm for 22 h and three times washing with PBS (10 mM phosphate-buffered saline; pH 7.4), blocking with 1% (w/v) bovine serum albumin (BSA) diluted in PBS was performed at 22 °C, 150 rpm for 2.5 h. Plates were washed two-times with PBS and polyclonal rabbit anti-LipA antibody diluted 1:5000 in PBS was added and incubated at 22 °C, 150 rpm for 2 h, followed by four times washing with PBS. After another 3 h incubation with the goat anti-rabbit horseradish peroxidase antibody (diluted 1:5000 in PBS; BioRad, Germany), Polysorp® 96-well microtiter plates (Nunc-Immuno™ MicroWell™ 96-Well Plate) were finally washed four times with PBS.

LipA was quantified by determination of horseradish peroxidase activity measured using the 1-step TMB ELISA substrate (3,3′,5,5′-tetramentylbenzidine; Thermo Fisher Scientific, Germany) at 25 °C for 15 min in the SpectraMax 250-plate reader (Molecular Devices, Germany). The amount of extracellular LipA was calculated using a standard curve determined with purified LipA. A two-tailed t-test was performed with a significance level of $p < 0.05$ to determine significant changes in LipA protein amount.

Real-time quantitative PCR for determination of *lipA* transcripts

Cell cultures were harvested after 6 h of growth, and RNA was prepared using the NucleoSpin® RNA Kit (Macherey–Nagel, Germany). cDNA synthesis of 1 µg RNA was performed with the Maxima First Strand cDNA Synthesis Kit for RT-qPCR Kit (Thermo Fisher Scientific, Germany). 50 ng cDNA and 50 ng of RNA (NoRT controls) were applied for RT-qPCR using the Maxima SYBR/ROX qPCR Master Mix (Thermo Fisher Scientific, Germany) and the primer pairs *lipA*_fw: 5′gcttccgggaacagatccaa 3′ and *lipA*_rev: 5′ acagaaggccgatgtgtcca 3′. The *sigA* gene was used as a reference and amplified using the primers *sigA*_fw: 5′ atcgcctgtctgatccacca 3′ and *sigA*_rev: 5′ ggtatgtcggacgcggtatg 3′. Gene expression analysis was performed with the REST 2009 software (Qiagen, Germany) using the $2^{-\Delta\Delta CT}$ method with an assumed PCR efficiency of 100% [28, 29]. Here, expression of *lipA* in three biological replicates (each analyzed three-times by RT-qPCR) is first normalized to the expression level of the reference gene *sigA* in the same culture, which encodes for the major sigma factor in *B. subtilis* and is equally expressed in all cells with less than 5% deviation in all analyzed samples. In a second step, the resulting value is compared to the corresponding value derived from a control culture, here *B. subtilis* expressing the wild-type *lipA* gene, resulting in an x-fold change in expression level.

To obtain information about the reliability and reproducibility of the RT-qPCR data, the relative change of normalized *lipA* transcript amount among all wt*lipA* expressions was determined using the REST 2009 software (Qiagen, Germany). 33 replicates were analyzed twice and revealed a standard error for the wt*lipA* transcript amount of 0.6 or 1.2 (lower and upper standard error, respectively). Therefore, only changes of transcript amounts lower than 0.4 or higher than 2.2 with a p value < 0.05 (calculated by REST 2009) were defined as significantly changed.

Sequence analysis

Protein sequences were obtained from the Pfam database of protein families [30] to determine the degree of amino acid conservation with respect to *B. subtilis* LipA. 64 lipase (Class 2) sequences out of 41 species from the *Firmicutes* phylum were aligned using Clustal Omega [31]. The number of amino acids in this alignment identical to the amino acid in the *B. subtilis* LipA sequence was counted for each position. This position-dependent conservation of each *B. subtilis* LipA amino acid within the *Firmicutes* phylum in percent is shown in Additional file 1: Table S3. The hydropathy index of Kyte and Doolittle was used as hydrophobicity scale [32] and changes > 1 were assumed to be significant.

Constraint network analysis

The X-ray crystal structure (PDB ID: 1ISP) with the highest resolution (1.3 Å) of *B. subtilis* LipA was used as the wtLipA structure, as well as a template to generate structures for LipA variants. All buffer ions and crystallization solvents were removed from the crystal structure. The models of the single variant structures were generated by the SCWRL4 program [33]. With the help of a rotamer library, SCWRL4 constructs variant models by predicting backbone-dependent side-chain conformations, while coordinates of backbone atoms stay unchanged. For enabling a local structural relaxation around the mutated residue, conformations of side chains of all residues

within 8 Å of the mutated residue were re-predicted. Hydrogen atoms were added, and side chains of Asn, Gln and His were flipped by the REDUCE program [34] for all variant structures. All structures were minimized by 100 steps of steepest descent followed by 5000 steps of conjugate gradient minimization or until the root mean-square gradient of the energy was < $1.0 * 10^{-4}$ kcal mol^{-1} $Å^{-1}$. The energy minimization was carried out with Amber14 using the ff99SB force field [35] and the GB^{OBC} Generalized Born model [36].

Thermal unfolding simulations by constraint network analysis (CNA) were performed as described previously [37–39]. In order to improve the robustness of CNA but without comprising CNA's high computational efficiency, CNA was carried out on an ensemble of network topologies generated from a single input structure by using fuzzy non-covalent constraints [40]. Here, the number and distribution of non-covalent constraints are modulated by random components within the ranges described in the Additional file 2: Methods, thus simulating thermal fluctuations of a biomacromolecule without actually moving atoms. An ensemble of 1000 network configurations was generated for wtLipA and all LipA variants. For the thermal unfolding simulations, the hydrogen bond energy cutoff E_{cut} was varied between −0.1 to −6.0 kcal mol^{-1} with a step size of 0.1 kcal mol^{-1}, equivalent to increasing the temperature from 302 to 380 K in steps of 2 K [41]. The number of hydrophobic constraints was kept constant during the thermal unfolding simulations.

A neighbor stability map [42] averaged over all 1000 conformations was computed from the thermal unfolding trajectories, and its median ($\widetilde{rc}_{ij,\,neighbor}$) was used to compare the thermostabilities of wtLipA and LipA variants, as done previously [43]. See Additional file 2: Methods for more information.

Results

Construction of the *lipA* site saturation mutagenesis library

The expression vector pBSlipA (see "Methods" section) encoding the native LipA of *B. subtilis* was used for site saturation mutagenesis (Fig. 1). In total, 155 amino acid residues of LipA with a conservation < 95% within the *Firmicutes* phylum (Pfam database entry: PF01674) [30] were used to generate the screened 29,199 clones as described in the "Methods" section.

Two-step screening of the *lipA* site saturation mutagenesis library

The LipA clones were cultivated in 96-well microtiter plates and analyzed with a two-step screening procedure including determination of extracellular volume activity and amount of LipA (Fig. 2a).

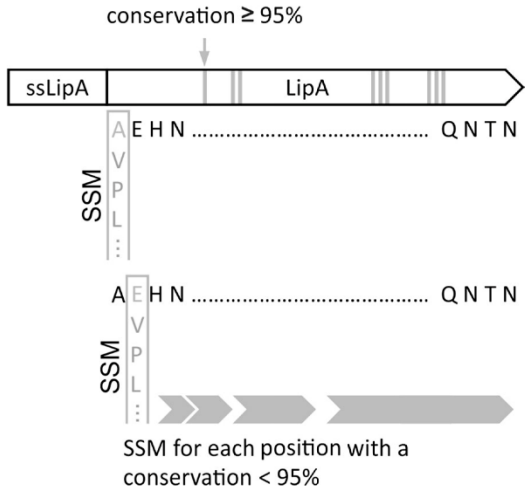


Fig. 1 Schematic strategy to construct the *lipA* site saturation mutagenesis library. The expression vector pBSlipA, where the native *lipA* gene is under control of the constitutive $P_{\,lipA\,II}$ promoter, was used for construction of the *lipA* library by site saturation mutagenesis (SSM). The gene encodes the secretion signal (ssLipA) and the mature lipase (LipA). Each codon encoding an amino acid of the mature LipA with a conservation < 95% within the *Firmicutes* was considered for SSM using QuikChange® PCR with a degenerated "NNS" codon to randomly introduce every possible amino acid. Amino acids with a conservation ≥ 95% (|) were not considered for substitution and screening. In total, 155 out of 181 residues were substituted and the resulting variants subjected to screening

In the first step, extracellular lipase activity was determined with *p*NPP as the substrate. In total, 5444 clones (19%) were inactive with the majority located at amino acid positions 26, 35, 41, 49, 101, 102, 104, 156, 160 and 181. To calculate a mean wtLipA lipase activity, 384 wtLipA clones were analyzed allowing to separate clones with significantly increased or decreased extracellular lipase activity from those with wtLipA activities. The volume activity and the corresponding standard deviation (σ) of wtLipA were 0.57 ± 0.12 U/ml. Compared to this, 4230 clones (14%) showed a significant decrease in extracellular lipase activity with amino acid substitutions at positions 19, 22, and 40. Furthermore, 66% (19,350) of all 29,199 screened clones showed activities similar to that of wtLipA and were therefore discarded.

Only 175 clones (1%) produced LipA variants with volume activities that were larger than wtLipA volume activity with its standard deviation (LipA variant U/ml > wtLipA U/ml + σ). Sequencing of the respective inserts revealed 26 clones as false-positive harboring the *lipA* wild-type sequence, 65 clones as duplicates with the identical codon exchange, and four LipA clones with multiple amino acid substitutions. The resulting 80 LipA

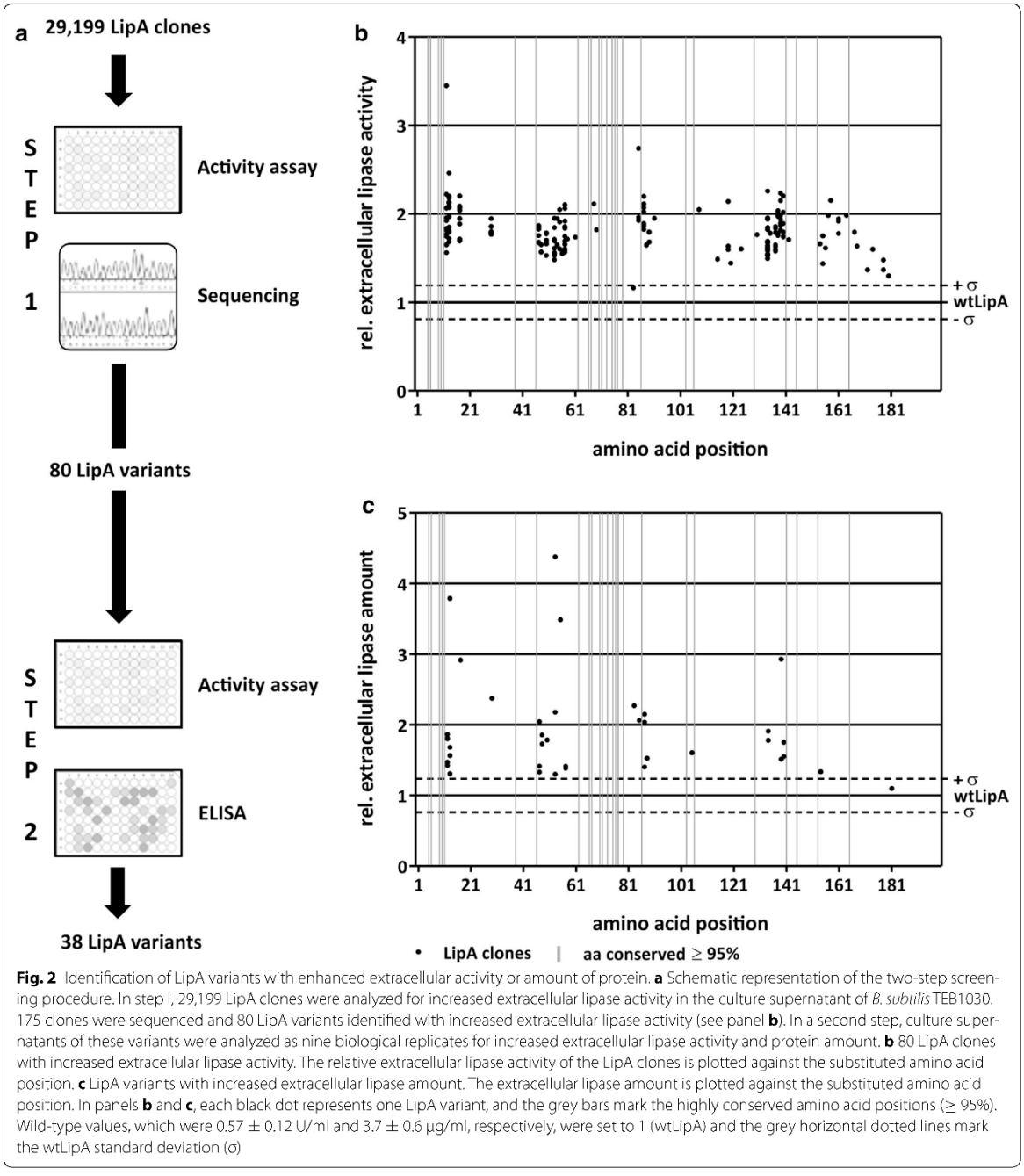


Fig. 2 Identification of LipA variants with enhanced extracellular activity or amount of protein. **a** Schematic representation of the two-step screening procedure. In step I, 29,199 LipA clones were analyzed for increased extracellular lipase activity in the culture supernatant of *B. subtilis* TEB1030. 175 clones were sequenced and 80 LipA variants identified with increased extracellular lipase activity (see panel **b**). In a second step, culture supernatants of these variants were analyzed as nine biological replicates for increased extracellular lipase activity and protein amount. **b** 80 LipA clones with increased extracellular lipase activity. The relative extracellular lipase activity of the LipA clones is plotted against the substituted amino acid position. **c** LipA variants with increased extracellular lipase amount. The extracellular lipase amount is plotted against the substituted amino acid position. In panels **b** and **c**, each black dot represents one LipA variant, and the grey bars mark the highly conserved amino acid positions (≥ 95%). Wild-type values, which were 0.57 ± 0.12 U/ml and 3.7 ± 0.6 µg/ml, respectively, were set to 1 (wtLipA) and the grey horizontal dotted lines mark the wtLipA standard deviation (σ)

variants (Fig. 2b) showed single amino acid substitutions distributed over 38 amino acid positions and an increase in extracellular lipase activity from 1.2- to 3.4-fold in comparison to wtLipA.

Beneficial substitutions mainly accumulated between N-terminal amino acid positions 11–18, in the middle part of LipA between positions 46–59, and in the C-terminal part between positions 129–143 and 151–169, but

a clear pattern regarding amino acid position or property was not obvious.

In a second step, the 80 LipA variants from step 1 (Fig. 2) exhibiting increased extracellular lipase activity were analyzed as nine biological replicates in a 96-well microtiter plate. Extracellular lipase activity was determined and extracellular lipase amount was quantified with an enzyme-linked immunosorbent assay (Fig. 2c). 31 variants turned out to be false-positives in this verification step and did not show improved activity or amount compared to wtLipA. Additional eleven variants exhibited increased lipase activity but not protein amount. The remaining 38 variants showed an increased lipase amount with increased or similar activity compared to wtLipA. These 38 variants included 34 different amino acid substitutions and four variants with a substitution caused by a synonymous codon. Their extracellular protein amount ranged from 1.3-fold (a substitution at the C-terminal amino acid position 134) to 3.8-fold (N-terminal position 13) higher than that of wtLipA, which is produced at 3.7 ± 0.6 µg/ml (Fig. 2c).

The extracellular activity and amount of the 38 LipA variants could be affected at different stages including transcription, translation, and secretion (which are coupled for LipA), and/or improved maturation, folding, and activity. We produced these LipA variants by cultivating *B. subtilis* TEB1030 in a microfermentation system linked to online biomass measurement and analyzed transcription, activity, and protein amount after 6 h when production and secretion of wtLipA had reached their optimum (Additional file 3: Figure S1). Furthermore, online biomass measurements were performed for 24 h to exclude differences in growth of variant-producing *B. subtilis* clones, which was, however, not observed (Additional file 3: Figure S2).

Twelve LipA variants did not show increased extracellular enzyme activity or protein amount and were therefore discarded as false positives (Additional file 1: Table S4). Six LipA variants were identified as more active with an up to 2.4-fold increase in specific lipolytic activity in comparison to wtLipA with 64 ± 13 U/mg (Fig. 3a; Additional file 1: Table S4). In total, 21 variants (including

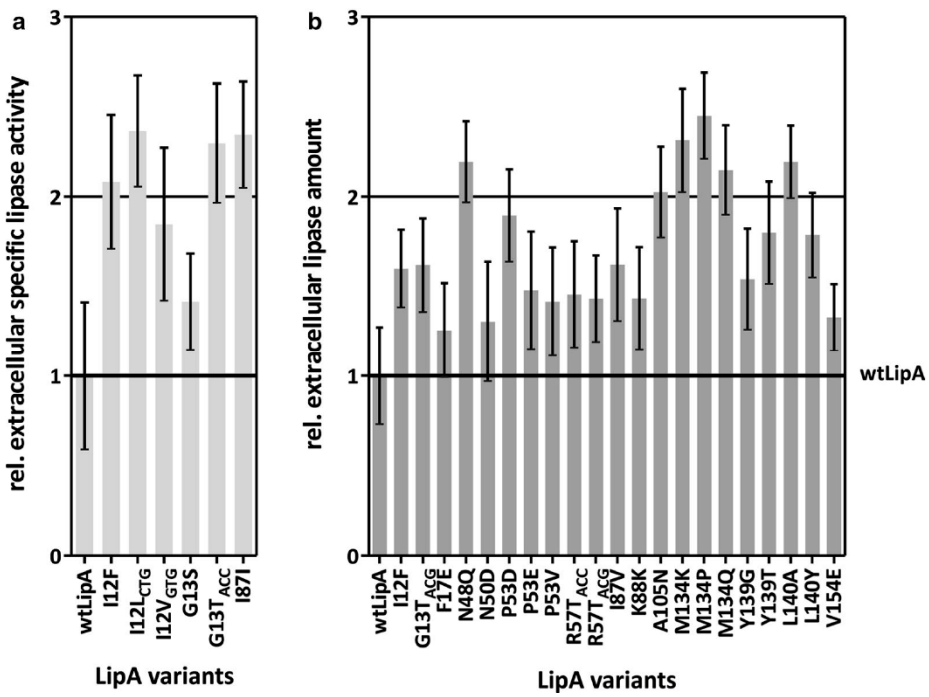


Fig. 3 LipA variants showing increased extracellular lipase activity (**a**) or increased extracellular lipase amount (**b**). *B. subtilis* TEB1030 producing the different LipA variants were cultivated for 6 h in a 48 well Flowerplate®, and the relative extracellular specific lipase activity (U/mg) in the culture supernatant was calculated by normalizing the volume activity (U/ml) to the determined protein amount (mg/ml). The extracellular lipase activity was measured using *p*NPP as a substrate, and the extracellular protein amount was determined by ELISA using a specific polyclonal LipA antibody. LipA variants with significant ($p < 0.05$) increase in extracellular specific lipase activity ($n = 9$) are shown relative to wtLipA. wtLipA values were set to 1 (thick black line)

one that also showed increased activity) showed an up to 2.3-fold increase in extracellular lipase amount (Fig. 3b; Additional file 1: Table S4).

Interestingly, the increase in extracellular LipA amount and/or activity of these 26 variants is unrelated to a change in hydrophobicity of the respective amino acid: 11 LipA variants carry a substitution to a significantly less hydrophobic amino acid, amino acid substitutions of 12 LipA variants do not or only slightly change hydrophobicity, and 3 LipA variants carry substitutions to more hydrophobic amino acids (see "Methods" section).

LipA variants with improved extracellular specific activity

Three out of the six variants with increased specific activity carry a substitution at amino acid I12 to phenylalanine, leucine, or valine, leading to a twofold increase in extracellular specific activity (Fig. 3a; Additional file 1: Table S4). LipA variants $I12L_{CTG}$, $I12V_{GTG}$, and $G13T_{ACC}$ were identified as more active, whereas identical amino acid substitutions encoded by different codons either showed no effect on LipA specific activity or LipA amount ($I12L_{TTG}$, $I12V_{GTC}$, see Additional file 1: Table S4) or resulted in an increased LipA amount ($G13T_{ACG}$, see Fig. 3b and Additional file 1: Table S4). Variant I87I with a silent mutation showed a twofold increase in extracellular specific activity but also a 3.6-fold significant change in *lipA* transcript level (Additional file 1: Table S4). This indicates, in all four cases, a codon- and not an amino acid-specific effect on LipA specific activity.

LipA variants with increased extracellular lipase amount

21 LipA variants showed a 1.3- to 2.3-fold increase in extracellular LipA protein amount at predominantly similar or decreased levels of extracellular specific activity compared to wtLipA (Fig. 3b; Additional file 1: Table S4) with the exception of variant I12F, which also showed a significant twofold increase in extracellular specific lipase activity (Fig. 3a; Additional file 1: Table S4). Only the mutations $G13T_{ACG}$ and I87I showed a significant 2.7- or 3.6-fold change in *lipA* transcript amount, respectively, while the transcript amount of all other 19 LipA variants was not significantly changed compared to wt*lipA* transcript (Additional file 1: Table S4).

We identified two LipA variants with the identical amino acid substitution R57T, which were encoded by the codons ACC and ACG (Fig. 3b; Additional file 1: Table S4). Both variants showed a similar increase in the extracellular LipA amount of ca. 1.4-fold compared to wtLipA level, indicating that this effect is caused by the introduced amino acid and not by the codon.

Seven LipA variants (N50D, P53D, P53E, P53 V, $R57T_{ACC}$, $R57T_{ACG}$ and M134Q) with increased extracellular LipA amount have amino acid substitutions located either in the αB-helix of LipA or carry a substitution to glutamine at position 134 (M134Q) (Fig. 3b; Additional file 1: Table S4). Since position M134 is known to contribute to thermostability [44] and the αB-helix also plays a role in tolerance towards detergents and ionic liquids [25, 45], (thermo)stability simulations were performed to probe for changes on LipA's (thermo)stability.

Thermal unfolding simulations of LipA variants

In order to determine to what extent an increase in LipA (thermo)stability could contribute to an increased extracellular LipA amount, the five variants N50D, P53D, P53E, P53V, and R57T with amino acid substitutions in the αB-helix and variant M134Q were subjected to thermal unfolding simulations by constraint network analysis [38]. CNA is a rigidity theory-based approach that models proteins as networks of constraints, where the constraints are defined from covalent and non-covalent (hydrogen bonds and hydrophobic interactions) bonds in the protein. Thermal unfolding of the protein is then simulated by removing hydrogen bond constraints in a step-wise manner in the order of increasing strength [41], and the influence on protein structural stability is monitored by global and local rigidity indices [42]. Here, as done previously for LipA [39, 46], the thermodynamic thermostability of LipA variants is compared to wtLipA in terms of a local index, the median of the neighbor stability map $\widetilde{rc}_{ij,\,neighbor}$. This $\widetilde{rc}_{ij,\,neighbor}$ has been shown to be related to the experimental melting temperature (T_m) and to be robust if variants follow different unfolding pathways [46]. Compared to the wtLipA $\widetilde{rc}_{ij,\,neighbor}$ value of 316.1 K, the variants N50D, P53E, P53V, R57T and M134Q show a decrease in thermodynamic thermostability by about 1.5 K on average (Table 2).

Combination of single amino acid substitutions

Single beneficial amino acid substitutions with different effects were combined to analyze putative synergistic

Table 2 Constraint network analysis (CNA) of wtLipA and LipA variants

LipA variants	$\widetilde{rc}_{ij,\,neighbor}$ (K)[a]	$\Delta\widetilde{rc}_{ij,\,neighbor}$ (K)[b]
wtLipA	316.1	–
N50D	312.1	−4.0
P53D	316.2	0.1
P53E	315.8	−0.3
P53 V	315.8	−0.3
R57T	314.9	−1.2
M134Q	314.7	−1.4

[a] The $\widetilde{rc}_{ij,\,neighbor}$ values were converted to a temperature scale according to equation 4 in Ref. [46]

[b] Difference of $\widetilde{rc}_{ij,\,neighbor}$ values of LipA variants minus wtLipA, respectively

effects on extracellular lipase activity and amount, or additive effects at the level of extracellular lipase amount. To do so, single amino acid substitutions with an increasing effect on either activity (G13S) or amount (A105N and Y139T) were chosen (Fig. 3; Additional file 1: Table S4), and double mutants were generated by site-directed mutagenesis. The corresponding single variants and the wild-type were produced and analyzed again as controls in this experiment confirming the beneficial effects of these substitutions with only slight differences in the absolute numbers.

No synergistic effect was observed when combining G13S with either A105N or Y139T (Additional file 1: Table S4). When G13S was combined with A105N, the extracellular specific lipase activity of the double mutant G13S/A105N was significantly increased by 2.9-fold compared to wtLipA with 42.7 ± 9.1 U/mg (Fig. 4a; Additional file 1: Table S4), reaching similar levels as the G13S variant. However, the extracellular lipase amount of this double mutant was only slightly increased compared to wtLipA but reduced compared to the A105N variant. The *lipA* transcript amount of the double mutant is not significantly changed compared to wtLipA (Additional file 1: Table S4). This indicates that the G13S substitution, affecting the extracellular lipase activity, largely abolishes the influence of the A105N substitution on protein amount.

The second double mutant G13S/Y139T was unaffected on the level of extracellular specific lipase activity (Fig. 4a; Additional file 1: Table S4) compared to wtLipA and 2.5-fold reduced compared to the G13S single variant. The extracellular lipase amount was 1.4-fold increased compared to wtLipA at similar levels of *lipA* transcript amount, but reduced compared to the single A105N variant (Fig. 4a; Additional file 1: Table S4). Here, both beneficial single amino acid substitutions compensate each other, thus preventing a synergistic beneficial effect when being combined.

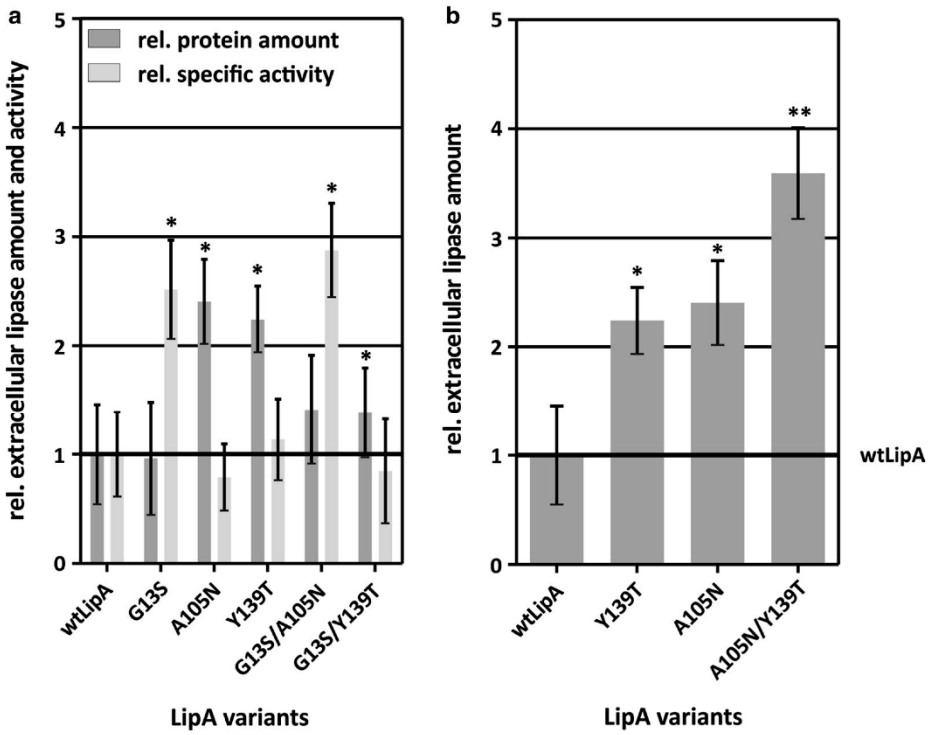


Fig. 4 Combination of beneficial single amino acid mutations in LipA and effects on (**a**) extracellular lipase activity and amount, and (**b**) additive effect of two amino acid substitutions on amount of extracellular lipase. *B. subtilis* TEB1030 harboring LipA variants were cultivated for 6 h in a 48 well Flowerplate®, and the relative extracellular specific lipase activity (U/mg) in the culture supernatant was calculated by normalizing the volume activity (U/ml) to the protein amount (mg/ml). The extracellular lipase activity was measured using *p*NPP as a substrate, and the extracellular protein amount was determined by ELISA using a specific polyclonal LipA antibody. The relative extracellular lipase amount (dark grey bars) and relative extracellular specific lipase activity (light grey bars) are given relative to wtLipA. Respective wtLipA values were set to 1 (thick black line). Significant changes ($p < 0.05$) compared to wtLipA are marked with a *, significant changes compared to the single variants ($p < 0.05$) are marked with **

However, we also observed an additive effect of two beneficial single mutations in LipA. Variants A105N and Y139T showed a significant increase in extracellular LipA amount of up to 2.4-fold compared to wtLipA with 3.5 ± 0.8 µg/ml at similar levels of extracellular specific lipase activity and similar levels of *lipA* transcript amount (Fig. 4b; Additional file 1: Table S4). The corresponding double mutant LipA A105N/Y139T showed a significant 3.6-fold increase in extracellular LipA amount compared to wtLipA as well as a significant increase of 1.2-fold when compared to the LipA single variants (Fig. 4b; Additional file 1: Table S4).

Discussion

In this study, we have interrogated the role of single amino acid substitutions of the extracellular lipase LipA from *B. subtilis* with respect to increasing the activity and amount of secreted enzyme. LipA consists of 181 amino acids of which 26 were identified as strictly conserved in 64 lipase sequences within the *Firmicutes* phylum. The remaining 155 amino acids, which are less than 95% conserved, were subjected to a complete site saturation mutagenesis resulting in a library of about 30,000 clones. This library was analyzed to identify clones producing LipA with an increased extracellular activity or an increased amount of lipase protein. The plasmid-based *lipA* expression system increased the extracellular lipase activity from about 0.02 U/ml for the wild-type strain *B. subtilis* 168 [47] to ca. 0.6 U/ml with LipA yields at a mg/l-scale. This is below the g/l-yields obtained under optimized production conditions reported in literature [1, 2], however, it allows measurements also of small effects caused by beneficial substitutions.

Codon-specific effects

Several LipA variants seem to be affected by the changed codon, but not by the changed amino acid, namely $I12L_{CTG}$, $I12V_{GTG}$, $G13T_{ACC}$, and I87I. A codon substitution can obviously result in a changed amino acid, but can also alter the amount of mRNA, change the transcription rate or the transcript stability as well as the co-translational folding of a protein. We have performed RT-qPCRs to determine the amount of *lipA* transcripts. A mean transcript level of 33 biological and two systematic replicates of wtLipA were calculated resulting in a standard error ranging from 0.4 to 2.2 with the mean value arbitrarily set to 1. Only variants with a changed transcript level below or above this standard error range were assumed to be significantly changed and are discussed here.

An increased amount of transcript may result in an increased protein amount in the supernatant as observed for variant $G13T_{ACG}$ (Fig. 3b; Additional file 1: Table S4) whereas the specific activity remained unaffected. However, the synonymous amino acid substitution in $G13T_{ACC}$ interestingly did not affect the transcript amount but increased the specific activity. Since the same amino acid is introduced, the effect must be caused by the substituted T_{ACC} codon, which is less frequent than the wtLipA codon and the above mentioned T_{ACG} codon (Additional file 1: Table S4). Rare codons can decelerate the translation velocity, that way enabling a more efficient folding of the protein [48], which may explain the increased specific activity of $G13T_{ACC}$. Contrarily, variant I87I also showed an increased specific activity although it contains a more frequent codon (Fig. 3a; Additional file 1: Table S4). The impact of the introduced codon is also illustrated by different I12 variants (Fig. 3a; Additional file 1: Table S4).

Amino acid substitutions within and near the oxyanion hole can increase specific lipase activity

Five out of the six identified amino acid substitutions increasing extracellular specific lipase activity are located at position 12, forming part of the oxyanion hole [17], or nearby at position 13 (Fig. 3a; Additional file 1: Table S4). This supports former suggestions [49] that optimization approaches should focus on mutations near the substrate-binding site. Substitution of isoleucine by the larger aromatic phenylalanine in variant I12F could lead to a local conformational change, thereby shifting the NH group of the residue at position 12, which could improve the stabilization of the transition state and cause the observed twofold increase in specific activity. Surprisingly, we did not identify substitutions at position M78, the other amino acid forming part of the oxyanion hole [17]. In contrast to I12 and G13, which are located in a flexible turn of LipA, M78 is located in the αC-helix [17]. It is thus possible that substitutions in the αC-helix do not have an effect on LipA activity because conformational changes are sterically hindered. The substitution of glycine with serine in the G13S variant could also lead to a local structural change of LipA in the oxyanion hole region and/or stabilize this region by potential hydrogen bond interactions between the side chains of S13 and R44, that way positively affecting the stabilization of the transition state, which could explain the 1.4-fold increase in specific activity (Fig. 3a; Additional file 1: Table S4).

Amino acid substitutions improving LipA secretion and stability

In total, 21 LipA variants were identified with amino acid substitutions increasing extracellular LipA amount up to twofold. Six of these variants carry substitutions within the αB-helix of LipA (N50D, P53D, P53E, P53V, $R57T_{ACC}$ and $R57T_{ACG}$; Fig. 3b; Additional file 1: Table S4). Amino

acid positions in this helix are known to contribute to detergent tolerance, when substituted to amino acids with charges opposite to the tested detergent [25], and to ionic liquid resistance, when charged and/or polar residues are introduced [45]. Therefore, it is possible that the higher extracellular LipA amount of these variants is not due to a more efficient secretion, but due to an increased stability in the culture supernatant of *B. subtilis*. This stability issue could also underlie the twofold higher extracellular LipA amount of variant M134Q (Fig. 3b; Additional file 1: Table S4). To probe this hypothesis, differences in the thermodynamic thermostability of the LipA variants with respect to wtLipA were predicted by thermal unfolding simulations using CNA; this approach has been previously applied successfully to retro- and prospectively analyze the thermodynamic thermostability of LipA variants [39, 43]. While for three variants (P53D, P53E, P53V) marginal changes in the predicted thermostability compared to wtLipA were found, a pronounced decrease in the thermostability was predicted for the other three variants (N50D, R57T, M134Q). The magnitude of this decrease is in the same ballpark as the magnitude of the median increase in the melting temperature found for 93 cases of engineered proteins, most of which contain more than one mutation [50]. Thus, the results of the CNA analyses do not support the hypothesis that increased *thermodynamic* thermostability of the six variants led to a higher LipA amount in the culture supernatant of *B. subtilis*. However, it should be noted that CNA does not consider time-dependency of processes; hence, our analyses do not rule out an increase in *kinetic* thermostability as a cause for higher extracellular LipA amount.

For the 13 LipA variants I12F, F17E, N48Q, I87V, K88K, A105N, M134K, M134P, Y139G, Y139T, L140A, L140Y, and V154E (Fig. 3b; Additional file 1: Table S4) no stabilizing effects have been described in literature so far. Noteworthy exceptions are amino acid positions N48 and A105, which have been previously identified during thermal unfolding simulations by CNA as structural 'weak spots', where mutations could particularly enhance LipA's thermostability [39].

The identified amino acid positions affecting extracellular protein amount are located in the N- (12, 17, 48), the middle (87, 88, 105), and the C- (134, 139, 140, 154) terminal part of LipA and show no preference regarding the charge of the introduced amino acid. Such randomly distributed mutations within the mature part of an enzyme can affect its secretion as shown for a lipase from *Pseudomonas aeruginosa* [15]. Furthermore, it was demonstrated that N-terminally located amino acids of the mature LamB protein are required for efficient transport in *E. coli* [51]. This could also explain the effect of the three substitutions I12F, F17E, and N48Q in the N-terminal part of LipA. The substitutions identified within the middle (I87V, K88K, A105N) and the C-terminal part of LipA (M134K, M134P, Y139G, Y139T, L140A, L140Y, and V154E) may confer a higher affinity to or allow for a better interaction with components of the translocation machinery such as Sec ATPase or SecYEG translocon [7–10].

Rational combination of LipA substitutions

In order to answer the question whether a synergistic effect can be achieved by combining single amino acid substitutions that themselves have led to increased specific activity or protein amount, we chose a single amino acid substitution beneficial for extracellular specific lipase activity (G13S; Fig. 3a; Additional file 1: Table S4) and two single amino acid substitutions increasing the extracellular lipase amount (A105N and Y139T; Fig. 3b; Additional file 1: Table S4). The combination of substitutions G13S/A105N and G13S/Y139T (Fig. 4a; Additional file 1: Table S4) resulted in either improved activity, or the effects of the single mutations were abrogated resulting in wild-type level specific activity and protein amount. Apparently, a beneficial mutation can affect e.g. RNA or protein structure or stability. Such effects may thus reinforce or neutralize each other when combined in a double mutant. However, the combination of amino acid substitutions A105N and Y139T, which both individually increased the extracellular protein amount 1.4-fold, resulted in a further increase to 3.6-fold in extracellular protein as compared to the single variants (Fig. 4b; Additional file 1: Table S4), demonstrating in this case an additive effect. Similar additive effects were already described for amino acid substitutions improving thermostability, where 12 amino acid substitutions were introduced by several rounds of in vitro evolution resulting in an increase of the LipA temperature optimum by ~ 30 °C [52]. It should be mentioned that many of such combination experiments need to be carried out before a general conclusion can be drawn.

Conclusions

In this study, we have systematically analyzed the role of single amino acid and codon substitutions for the secretory production of the model protein LipA in *B. subtilis*. In addition to single amino acid substitutions increasing LipA specific activity and protein amount, we also observed multiple codon-related effects on *lipA* transcription which apparently also influence LipA specific activity. We have identified six LipA variants with increased extracellular specific lipase activity (I12F, $I12L_{CTG}$, $I12V_{GTG}$, G13S, $G13T_{ACC}$, and I87I), of which one also showed an increased extracellular lipase amount

(I12F), and a double mutant (A105N/Y139T) which showed an additive effect of the single mutations on the level of extracellular protein amount. The fact that silent mutations can alter the LipA translation rate and thus promote more or less efficient LipA folding is expected to contribute to discussions on the importance of codon bias and abundance in *B. subtilis*, as previously remarked [53]. In summary, we have identified 26 in about 30,000 LipA variants that showed an increase in either amount or specific activity of extracellular lipase. The low success rate and the fact that the most pronounced increases were about twofold only indicate that nature has already optimized production and secretion very well for this lipase in *B. subtilis*. Nevertheless, our results also suggest that optimization campaigns aiming at increased enzyme production may also consider the target protein itself. Variant generation with improved properties might be particularly successful if prioritized towards 'sensitive' structural elements, as we find that mutations in the vicinity of the active site on the αB-helix, or at structural 'weak spots' showed a higher propensity for improved protein amount and/or activity.

Additional files

Additional file 1. Additional tables.

Additional file 2. Additional methods.

Additional file 3. Additional figures.

Authors' contributions

PS performed most of the biological experiments and drafted the manuscript. KV constructed various expression plasmids. AF designed and coordinated parts of the study. AB and CN performed the constraint network analysis and drafted the corresponding parts in the manuscript. AK supervised parts of the study and drafted the manuscript. HG and KEJ conceived the project, participated in the design and coordination and edited the manuscript. All authors read and approved the final manuscript.

Author details

[1] Institute of Molecular Enzyme Technology, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany. [2] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany. [3] John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) & Institute for Complex Systems - Structural Biochemistry (ICS6), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany. [4] Institute of Bio- and Geosciences IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany. [5] Present Address: Macromolecular Chemistry and New Polymeric Materials, Zernike Institute of Advanced Materials, University of Groningen, Nijenborgh 4, 9747AG Groningen, The Netherlands. [6] Present Address: Novozymes A/S, Krogshoejvej 36, 2880 Bagsvaerd, Denmark.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All data analyzed during this study are included in this published article and its Additional files 1 (tables), 2 (methods), 3 (figures).

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

Part of this work was funded by the Bioeconomy Science Center, which is financially supported by the Ministry of Innovation, Research and Science of North-Rhine Westphalia, Germany, in the framework of the NRW Strategieprojekt BioSC (No. 313/323-400-00213). PS and AB were funded by a scholarship from the CLIB^{2021} Graduate Cluster "Industrial Biotechnology", and AF was funded by the German Research Foundation (DFG) within research training group 1166 "Biocatalysis using Non-Conventional Media—BioNoCo. We further acknowledge support by the DFG for financial contribution to the liquid handling platform Tecan Freedom evo 200 (INST 208/654-1 FUGG to KEJ) and the hybrid compute cluster (INST 208//04-1 FUGG to HG).

Publisher's Note

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

Received: 23 May 2017 Accepted: 13 September 2017
Published online: 25 September 2017

References

1. Schallmey M, Singh A, Ward OP. Developments in the use of *Bacillus* species for industrial production. Can J Microbiol. 2004;50:1–17. doi:10.1139/w03-076.
2. Westers L, Westers H, Quax WJ. *Bacillus subtilis* as cell factory for pharmaceutical proteins: a biotechnological approach to optimize the host organism. Biochim Biophys Acta. 2004;1694:299–310. doi:10.1016/j.bbamcr.2004.02.011.
3. Nijland R, Kuipers O. Optimization of protein secretion by *Bacillus subtilis*. Recent Pat Biotechnol. 2008;2:79–87. doi:10.2174/187220808784619694.
4. van Dijl JM, Hecker M. *Bacillus subtilis*: from soil bacterium to super-secreting cell factory. Microb Cell Fact. 2013;12:3. doi:10.1186/1475 2859 12 3.
5. Fedyunin I, Lehnhardt L, Boehmer N, Kaufmann P, Zhang G, Ignatova Z. tRNA concentration fine tunes protein solubility. FEBS Lett. 2012;586:3336–40. doi:10.1016/j.febslet.2012.07.012.
6. Hess A-K, Saffert P, Liebeton K, Ignatova Z. Optimization of translation profiles enhances protein expression and solubility. PLoS ONE. 2015;10:e0127039. doi:10.1371/journal.pone.0127039.
7. Tjalsma H, Antelmann H, Jongbloed JDH, Braun PG, Darmon E, Dorenbos R, Dubois JF, Westers H, Zanen G, Quax WJ, Kuipers OP, Bron S, Hecker M, Van Dijl JM. Proteomics of protein secretion by *Bacillus subtilis*: separating the "secrets" of the secretome. Microbiol Mol Biol Rev. 2004;68:207–33.
8. Zanen G, Antelmann H, Meima R, Jongbloed JDH, Kolkman M, Hecker M, van Dijl JM, Quax WJ. Proteomic dissection of potential signal recognition particle dependence in protein secretion by *Bacillus subtilis*. Proteomics. 2006;6:3636–48. doi:10.1002/pmic.200500560.
9. Fekkes P, Driessen AJ. Protein targeting to the bacterial cytoplasmic membrane. Microbiol Mol Biol Rev. 1999;63:161–73.
10. Ton-That H, Marraffini LA, Schneewind O. Protein sorting to the cell wall envelope of Gram-positive bacteria. Biochim Biophys Acta. 2004;1694:269–78. doi:10.1016/j.bbamcr.2004.04.014.
11. Brockmeier U, Caspers M, Freudl R, Jockwer A, Noll T, Eggert T. Systematic screening of all signal peptides from *Bacillus subtilis*: a powerful strategy in optimizing heterologous protein secretion in Gram-positive bacteria. J Mol Biol. 2006;362:393–402. doi:10.1016/j.jmb.2006.07.034.
12. Caspers M, Brockmeier U, Degering C, Eggert T, Freudl R. Improvement of sec-dependent secretion of a heterologous model protein in *Bacillus subtilis* by saturation mutagenesis of the N-domain of the AmyE signal peptide. Appl Microbiol Biotechnol. 2010;86:1877–85. doi:10.1007/s00253-009-2405-x.
13. Degering C, Eggert T, Puls M, Bongaerts J, Evers S, Maurer K-H, Jaeger K-E. Optimization of protease secretion in *Bacillus subtilis* and *Bacillus*

licheniformis by screening of homologous and heterologous signal peptides. Appl Environ Microbiol. 2010;76:6370–6. doi:10.1128/AEM.01146-10.
14. Vitikainen M, Hyyryläinen H-L, Kivimäki A, Kontinen VP, Sarvas M. Secretion of heterologous proteins in *Bacillus subtilis* can be improved by engineering cell components affecting post-translocational protein folding and degradation. J Appl Microbiol. 2005;99:363–75. doi:10.1111/j.1365-2672.2005.02572.x.
15. Hausmann S, Wilhelm S, Jaeger K-E, Rosenau F. Mutations towards enantioselectivity adversely affect secretion of *Pseudomonas aeruginosa* lipase. FEMS Microbiol Lett. 2008;282:65–72. doi:10.1111/j.1574-6968.2008.01107.x.
16. Altman E, Emr SD, Kumamoto CA. The presence of both the signal sequence and a region of mature LamB protein is required for the interaction of LamB with the export factor SecB. J Biol Chem. 1990;265:18154–60.
17. van Pouderoyen G, Eggert T, Jaeger K-E, Dijkstra BW. The crystal structure of *Bacillus subtilis* lipase: a minimal α/β hydrolase fold enzyme. J Mol Biol. 2001;309:215–26. doi:10.1006/jmbi.2001.4659.
18. Eggert T. Die lipolytischen Enzyme LipA und LipB von *Bacillus subtilis*: Charakterisierung und Optimierung mit gerichteter Evolution. Dissertation, Ruhr-Universität Bochum. 2001.
19. Eggert T, van Pouderoyen G, Dijkstra BW, Jaeger KE. Lipolytic enzymes LipA and LipB from *Bacillus subtilis* differ in regulation of gene expression, biochemical properties, and three-dimensional structure. FEBS Lett. 2001;502:89–92. doi:10.1016/S0014-5793(01)02665-5.
20. Woodcock DM, Crowther PJ, Doherty J, Jefferson S, DeCruz E, Noyer-Weidner M, Smith SS, Michael MZ, Graham MW. Quantitative evaluation of *Escherichia coli* host strains for tolerance to cytosine methylation in plasmid and phage recombinants. Nucleic Acids Res. 1989;17:3469–78.
21. Brockmeier U, Wendorff M, Eggert T. Versatile expression and secretion vectors for *Bacillus subtilis*. Curr Microbiol. 2006;52:143–8. doi:10.1007/s00284-005-0231-7.
22. Sambrook J, Russell DW. Molecular cloning: a laboratory manual. 3rd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2001.
23. Chang S, Cohen S. High frequency transformation of *Bacillus subtilis* protoplasts by plasmid DNA. Mol General Genet. 1979;115:111–5.
24. Edelheit O, Hanukoglu A, Hanukoglu I. Simple and efficient site-directed mutagenesis using two single-primer reactions in parallel to generate mutants for protein structure function studies. BMC Biotechnol. 2009;9:61. doi:10.1186/1472-6750-9-61.
25. Fulton A, Frauenkron-Machedjou VJ, Skoczinski P, Wilhelm S, Zhu L, Schwaneberg U, Jaeger K-E. Exploring the protein stability landscape: *Bacillus subtilis* lipase A as a model for detergent tolerance. ChemBioChem. 2015;16:930–6. doi:10.1002/cbic.201402664.
26. Reetz MT, Kahakeaw D, Lohmer R. Addressing the numbers problem in directed evolution. ChemBioChem. 2008;9:1797–804. doi:10.1002/cbic.200800298.
27. Nov Y. When second best is good enough: another probabilistic look at saturation mutagenesis. Appl Environ Microbiol. 2012;78:258–62. doi:10.1128/AEM.06265-11.
28. Pfaffl MW, Horgan GW, Dempfle L. Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res. 2002;30:e36. doi:10.1093/nar/30.9.e36.
29. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. Nat Protocol. 2008;3:1101–8. doi:10.1038/nprot.2008.73.
30. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222–30. doi:10.1093/nar/gkt1223.
31. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539. doi:10.1038/msb.2011.75.
32. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157:105–32. doi:10.1016/0022-2836(82)90515-0.
33. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637.
34. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol. 1999;285:1735–47.
35. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins Struct Funct Bioinform. 2006;65:712–25. doi:10.1002/prot.21123.
36. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. Proteins Struct Funct Bioinform. 2004;55:383–94. doi:10.1002/prot.20033.
37. Hermans Pfleger C, Nutschel C, Hanke CA, Gohlke HSMA. Rigidity theory for biomolecules: concepts, software, and applications. WIREs Comput Mol Sci. 2017;. doi:10.1002/wcms.1311.
38. Pfleger C, Rathi PC, Klein DL, Radestock S, Gohlke H. Constraint network analysis (CNA): a Python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. J Chem Inf Model. 2013;53:1007–15. doi:10.1021/ci400044m.
39. Rathi PC, Fulton A, Jaeger K E, Gohlke H. Application of rigidity theory to the thermostabilization of lipase A from *Bacillus subtilis*. PLoS Comput Biol. 2016;12:e1004754. doi:10.1371/journal.pcbi.1004754.
40. Pfleger C, Gohlke H. Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. Structure. 2013;21:1–10. doi:10.1016/j.str.2013.07.012.
41. Radestock S, Gohlke H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng Life Sci. 2008;8:507–22. doi:10.1002/elsc.200800043.
42. Pfleger C, Radestock S, Schmidt E, Gohlke H. Global and local indices for characterizing biomolecular flexibility and rigidity. J Comput Chem. 2013;34:220–33. doi:10.1002/jcc.23122.
43. Rathi PC, Jaeger K, Gohlke H. Structural rigidity and protein thermostability in variants of lipase A from *Bacillus subtilis*. PLoS ONE. 2015;10:1–24. doi:10.1371/journal.pone.0130289.
44. Kamal MZ, Ahmad S, Yedavalli P, Rao NM. Stability curves of laboratory evolved thermostable mutants of a *Bacillus subtilis* lipase. Biochim Biophys Acta. 2010;1804:1850–6. doi:10.1016/j.bbapap.2010.06.014.
45. Frauenkron-Machedjou VJ, Fulton A, Zhu L, Anker C, Bocola M, Jaeger K-E, Schwaneberg U. Towards understanding directed evolution: more than half of all amino acid positions contribute to ionic liquid resistance of *Bacillus subtilis* lipase A. ChemBioChem. 2015;16:937–45. doi:10.1002/cbic.201402682.
46. Radestock S, Gohlke H. Protein rigidity and thermophilic adaptation. Proteins Struct Funct Bioinform. 2011;79:1089–108. doi:10.1002/prot.22946.
47. Eggert T, Brockmeier U, Droge MJ, Quax WJ, Jaeger K-E. Extracellular lipases from *Bacillus subtilis*: regulation of gene expression and enzyme activity by amino acid supply and external pH. FEMS Microbiol Lett. 2003;225:319–24. doi:10.1016/S0378-1097(03)00536-6.
48. Spencer PS, Siller E, Anderson JF, Barral JM. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. J Mol Biol. 2012;422:328–35. doi:10.1016/j.jmb.2012.06.010.
49. Morley KL, Kazlauskas RJ. Improving enzyme properties: when are closer mutations better? Trends Biotechnol. 2005;23:231–7. doi:10.1016/j.tibtech.2005.03.005.
50. Wijma HJ, Floor RJ, Janssen DB. Structure- and sequence analysis inspired engineering of proteins for enhanced thermostability. Curr Opin Struct Biol. 2013;23:588–94. doi:10.1016/j.sbi.2013.04.008.
51. Rasmussen BA, Silhavy TJ. The first 28 amino acids of mature LamB are required for rapid and efficient export from the cytoplasm. Genes Dev. 1987;1:185–96. doi:10.1101/gad.1.2.185.
52. Kamal MZ, Ahmad S, Molugu TR, Vijayalakshmi A, Deshmukh MV, Sankaranarayanan R, Rao NM. In vitro evolved non-aggregating and thermostable lipase: structural and thermodynamic investigation. J Mol Biol. 2011;413:726–41. doi:10.1016/j.jmb.2011.09.002.
53. Ogasawara N. Markedly unbiased codon usage in *Bacillus subtilis*. Gene. 1985;40:145–50.

ORIGINAL PUBLICATION IV-SUPPORTING INFORMATION

Contribution of single amino acid and codon substitutions to the production and secretion of a lipase by *Bacillus subtilis*

Skoczinski, P., Volkenborn, K., Fulton, A., Bhadauriya, A., Nutschel, C., Gohlke, H., Knapp, A., Jaeger, K.-E.

Microb. Cell Fact. 2017, 16, 160.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5613506/

Skoczinski *et al.*, 2017: LipA single substitutions

Additional Tables

Table S1 Oligonucleotide sequences for generation of LipA site saturation mutagenesis library
The forward and reverse oligonucleotide sequence is shown for each mutated codon positions. Codon positions highlighted in light grey coding for amino acids with a conservation ≥ 95 % (among the Firmicutes) and were not considered for mutagenesis.

position	forward primer	reverse primer
1	GACTGGATTGTGTTCSNNGGCTTTTGCTGACGG	CCGTCAGCAAAAGCCNNSGAACACAATCCAGTC
2	AACGACTGGATTGTGSNNAGCGGCTTTTGCTGA	TCAGCAAAAGCCGCTNNSCACAATCCAGTCGTT
3	CATAACGACTGGATTSNNTTCAGCGGCTTTTGC	GCAAAAGCCGCTGAANNSAATCCAGTCGTTATG
4	AACCATAACGACTGGSNNGTGTTCAGCGGCTTT	AAAGCCGCTGAACACNNSCCAGTCGTTATGGTT
5	-	-
6	-	-
7	AATACCGTGAACCATSNNGACTGGATTGTGTTC	GAACACAATCCAGTCNNSATGGTTCACGGTATT
8	TCCAATACCGTGAACSNNAACGACTGGATTGTG	CACAATCCAGTCGTTNNSGTTCACGGTATTGGA
9	-	-
10	-	-
11	-	-
12	GAATGATGCCCCTCCSNNACCGTGAACCATAAC	GTTATGGTTCACGGTNNSGGAGGGGCATCATTC
13	ATTGAATGATGCCCCSNNAATACCGTGAACCATAAC	GTTATGGTTCACGGTATTNNSGGGGCATCATTCAAT
14	AAAATTGAATGATGCSNNTCCAATACCGTGAAC	GTTCACGGTATTGGANNSGCATCATTCAATTTT
15	CGCAAAATTGAATGASNNCCCTCCAATACCGTG	CACGGTATTGGAGGGNNSTCATTCAATTTTGCG
16	TCCCGCAAAATTGAASNNTGCCCCTCCAATACC	GGTATTGGAGGGGCANNSTTCAATTTTGCGGGA
17	AATTCCCGCAAAATTSNNTGATGCCCCTCCAATAC	GTATTGGAGGGGCATCANNSAATTTTGCGGGAATT
18	CTTAATTCCCGCAAASNNGAATGATGCCCCTCC	GGAGGGGCATCATTCNNSTTTGCGGGAATTAAG
19	GCTCTTAATTCCCGCSNNATTGAATGATGCCCC	GGGGCATCATTCAATNNSGCGGGAATTAAGAGC
20	GCTCTTAATTCCSNNAAAATTGAATGATGC	GCATCATTCAATTTTNNSGGAATTAAGAGC
21	ATAGCTCTTAATSNNCGCAAAATTGAATGATGC	GCATCATTCAATTTTGCGNNSATTAAGAGCTAT
22	-	-
23	AGATACGAGATAGCTSNNAATTCCCGCAAAATTG	CAATTTTGCGGGAATTNNSAGCTATCTCGTATCT
24	CTGAGATACGAGATASNNCTTAATTCCCGC	GCGGGAATTAAGNNSTATCTCGTATCTCAG
25	GCCCTGAGATACGAGSNNGCTCTTAATTCCCGC	GCGGGAATTAAGAGCNNSCTCGTATCTCAGGGC
26	CCAGCCCTGAGATACSNNATAGCTCTTAATTCC	GGAATTAAGAGCTATNNSGTATCTCAGGGCTGG
27	CGACCAGCCCTGAGASNNGAGATAGCTCTTAATTC	GAATTAAGAGCTATCTCNNSTCTCAGGGCTGGTCG
28	CCGCGACCAGCCCTGSNNTACGAGATAGCTC	GAGCTATCTCGTANNSCAGGGCTGGTCGCGG
29	GTCCCGCGACCAGCCSNNAGATACGAGATAGCT	AGCTATCTCGTATCTNNSGGCTGGTCGCGGGAC
30	CTTGTCCCTCGACCASNNCTGAGATACGAGATAG	CTATCTCGTATCTCAGNNSTGGTCGAGGGACAAG
31	CAGCTTGTCCCGCGASNNGCCCTGAGATACGAG	CTCGTATCTCAGGGCNNSTCGCGGGACAAGCTG
32	ATACAGCTTGTCCCGSNNCCAGCCCTGAGATAC	GTATCTCAGGGCTGGNNSCGGGACAAGCTGTAT
33	TGCATACAGCTTGTCSNNCGACCAGCCCTGAG	CTCAGGGCTGGTCGNNSGACAAGCTGTATGCA
34	AACTGCATACAGCTTSNNCCGCGACCAGCCCTG	CAGGGCTGGTCGCGGNNSAAGCTGTATGCAGTT
35	ATCAACTGCATACAGSNNGTCCCGCGACCAGCC	GGCTGGTCGCGGGACNNSCTGTATGCAGTTGAT
36	AAAATCAACTGCATASNNCTTGTCCCGCGACCAG	CTGGTCGCGGGACAAGNNSTATGCAGTTGATTTT
37	CCAAAAATCAACTGCSNNCAGCTTGTCCCGCGAC	GTCGCGGGACAAGCTGNNSGCAGTTGATTTTTGG
38	-	-
39	CTTGTCCCAAAAATCSNNTGCATACAGCTTGTC	GACAAGCTGTATGCANNSGATTTTTGGGACAAG
40	TGTCTTGTCCCAAAASNNAACTGCATACAGCTT	AAGCTGTATGCAGTTNNSTTTTGGGACAAGACA
41	GCCTGTCTTGTCCCSNNATCAACTGCATACAG	CTGTATGCAGTTGATNNSGGGACAAGACAGGC
42	TGTGCCTGTCTTGTCSNNAAAATCAACTGCATA	TATGCAGTTGATTTTNNSGACAAGACAGGCACA
43	ATTTGTGCCTGTCTTSNNCCAAAAATCAACTGC	GCAGTTGATTTTTGGNNSAAGACAGGCACAAAT
44	ATAATTTGTGCCTGTSNNGTCCCAAAAATCAAC	GTTGATTTTTGGGACNNSACAGGCACAAATTAT
45	GTTATAATTTGTGCCSNNCTTGTCCCAAAAATC	GATTTTTGGGACAAGNNSGGCACAAATTATAAC

Skoczinski *et al.*, 2017: LipA single substitutions

46	-	-
47	TCCATTGTTATAATTSNNGCCTGTCTTGTCCCAAAAATC	GATTTTTGGGACAAGACAGGCNNSAATTATAACAATGGA
48	CGGTCCATTGTTATASNNTGTGCCTGTCTTGTC	GACAAGACAGGCACANNSTATAACAATGGACCG
49	TACCGGTCCATTGTTSNNATTTTGTGCCTGTCTT	AAGACAGGCACAAAATNNSAACAATGGACCGGTA
50	TAATACCGGTCCATTSNNATAATTTGTGCCTGT	ACAGGCACAAATTATNNSAATGGACCGGTATTA
51	TGATAATACCGGTCCSNNGTTATAATTTGTGCC	GGCACAAATTATAACNNSGGACCGGTATTATCA
52	TCGTGATAATACCGGSNNATTGTTATAATTTGTG	CACAAATTATAACAATNNSCCGGTATTATCACGA
53	AAATCGTGATAATACSNNTCCATTGTTATAATT	AATTATAACAATGGANNSGTATTATCACGATTT
54	CACAAATCGTGATAASNNCGGTCCATTGTTATA	TATAACAATGGACCGNNSTTATCACGATTTGTG
55	TTGCACAAATCGTGASNNTACCGGTCCATTG	CAATGGACCGGTANNSTCACGATTTGTGCAA
56	CTTTTGCACAAATCGSNNTAATACCGGTCCATTG	CAATGGACCGGTATTANNSCGATTTGTGCAAAAG
57	AACCTTTTGCACAAASNNTGATAATACCGGTCC	GGACCGGTATTATCANNSTTTGTGCAAAAGGTT
58	TAAAACCTTTTGCACSNNTCGTGATAATACCGG	CCGGTATTATCACGANNSGTGCAAAAGGTTTTA
59	ATCTAAAACCTTTTGSNNAAATCGTGATAATAC	GTATTATCACGATTTNNSCAAAAGGTTTTAGAT
60	TTCATCTAAAACCTTSNNCACAAATCGATAA	TTATCGATTTGTGNNSAAGGTTTTAGATGAA
61	CGTTCCATCTAAAACSNNTTGCACAAATCGTG	CACGATTTGTGCAANNSGTTTTAGATGGAACG
62	-	-
63	CGCACCCGTTTCATCSNNAACCTTTTGCACAAATC	GATTTGTGCAAAAGGTTNNSGATGAAACGGGTGCG
64	TTTCGCACCCGTTTCSNNTAAAACCTTTTGCAC	GTGCAAAAGGTTTTANNSGAAACGGGTGCGAAA
65	TTTTTTCGCACCCGTSNNATCTAAAACCTTTTG	CAAAAGGTTTTAGATNNSACGGGTGCGAAAAAA
66	-	-
67	-	-
68	AATATCCACTTTTTTSNNACCCGTTTCATCTAAAAC	GTTTTAGATGAAACGGGTNNSAAAAAAGTGGATATT
69	GACAATATCCACTTTSNNCGCACCCGTTTCATC	GATGAAACGGGTGCGNNSAAAGTGGATATTGTC
70	-	-
71	-	-
72	GCTGTGAGCGACAATSNNCACTTTTTTCGCACC	GGTGCGAAAAAAGTGNNSATTGTCGCTCACAGC
73	-	-
74	CCCCATGCTGTGAGCSNNAATATCCACTTTTTTC	GAAAAAAGTGGATATTNNSGCTCACAGCATGGGG
75	-	-
76	-	-
77	-	-
78	TGTGTTCGCGCCCCCSNNGCTGTGAGCGACAATATC	GATATTGTCGCTCACAGCNNSGGGGGCGCGAACACA
79	-	-
80	GTAAAGTGTGTTCGCSNNCCCCATGCTGTGAGC	GCTCACAGCATGGGGNNSGCGAACACACTTTAC
81	GTAGTAAAGTGTGTTSNNGCCCCCCATGCTGTG	CACAGCATGGGGGGCNNSAACACACTTTACTAC
82	TATGTAGTAAAGTGTSNNCGCGCCCCCCATGCTG	CAGCATGGGGGGCGCGNNSACACTTTACTACATA
83	TTTTATGTAGTAAAGSNNGTTCGCGCCCCCCATG	CATGGGGGGCGCGAACNNSCTTTACTACATAAAA
84	ATTTTTTATGTAGTASNNTATGTTCGCGCCCCC	GGGGGCGCGAACATANNSTACTACATAAAAAAT
85	CAGATTTTTTATGTASNNAAGTGTGTTCGCGCC	GGCGCGAACACACTTNNSTACATAAAAAATCTG
86	-	-
87	GCCGTCCAGATTTTTSNNGTAGTAAAGTGTGTTC	GAACACACTTTACTACNNSAAAAATCTGGACGGC
88	TCCGCCGTCCAGATTSNNTATGTAGTAAAGTGTG	CACACTTTACTACATANNSAATCTGGACGGCGGA
89	ATTTCCGCCGTCCAGSNNTTTTATGTAGTAAAG	CTTTACTACATAAAANNSCTGGACGGCGGAAAT
90	TTTATTTCCGCCGTCSNNATTTTTTATGTAGTAAAG	CTTTACTACATAAAAAATNNSGACGGCGGAAATAAA
91	AACTTTATTTCCGCCSNNCAGATTTTTTATGTAG	CTACATAAAAAATCTGNNSGGCGGAAATAAAGTT
92	TGCAACTTTATTTCCSNNGTCCAGATTTTTTATG	CATAAAAAATCTGGACNNSGGAAATAAAGTTGCA
93	GTTTGCAACTTTATTSNNGCCGTCCAGATTTTTTATG	CATAAAAAATCTGGACGGCNNSAATAAAGTTGCAAAC
94	GACGTTTGCAACTTTSNNTCCGCCGTCCAGATTTTTTATG	CATAAAAAATCTGGACGGCGGANNSAAAGTTGCAAACGTC
95	CACGACGTTTGCAACSNNATTTCCGCCGTCCAG	CTGGACGGCGGAAATNNSGTTGCAAACGTCGTG
96	CGTCACGACGTTTGCSNNTTTATTTCCGCCGTC	GACGGCGGAAATAAANNSGCAAACGTCGTGACG
97	AAGCGTCACGACGTTSNNAACTTTATTCCTGCC	GGCAGGAATAAAGTTNNSAACGTCGTGACGCTT

Additional Tables, page 2

98	GCCAAGCGTCACGACSNNTGCAACTTTATTTCC	GGAAATAAAGTTGCANNSGTCGTGACGCTTGGC
99	GCCGCCAAGCGTCACSNNGTTTGCAACTTTATTTC	GAAATAAAGTTGCAAACNNSGTGACGCTTGGCGGC
100	CGCGCCGCCAAGCGTSNNGACGTTTGCAAC	GTTGCAAACGTCNNSACGCTTGGCGGCGCG
101	GTTCGCGCCGCCAAGSNNCACGACGTTTGCAAC	GTTGCAAACGTCGTGNNSCTTGGCGGCGCGAAC
102	ACGGTTCGCGCCGCCSNNCGTCACGACGTTTGC	GCAAACGTCGTGACGNNSGGCGGCGCGAACCGT
103	-	-
104	CGTTAAACGGTTCGCSNNGCCAAGCGTCACGAC	GTCGTGACGCTTGGCNNSGCGAACCGTTTAACG
105	TGTCGTTAAACGGTTSNNGCCGCCAAGCGTCAC	GTGACGCTTGGCGGCNNSAACCGTTTAACGACA
106	-	-
107	CTTGCCTGTCGTCAASNNGTTCGCGCCGCCAGG	CCTGGCGGCGCGAACNNSTTGACGACAGGCAAG
108	CGCCTTGCCTGTCGTSNNACGGTTCGCGCCGCC	GGCGGCGCGAACCGTNNSACGACAGGCAAGGCG
109	AAGCGCCTTGCCTGTSNNCAAACGGTTCGCGCC	GGCGCGAACCGTTTGNNSACAGGCAAGGCGCTT
110	CCGAAGCGCCTTGCCSNNCGTCAAACGGTTCGC	GCGAACCGTTTGACGNNSGGCAAGGCGCTTCGG
111	CCCGGAAGCGCCTTSNNTGTCGTCAAACGGTTC	GAACCGTTTGACGACANNSAAGGCGCTTCCGGG
112	TGTTCCCCGAAGCGCSNNGCCTGTCGTCAAACG	CGTTTGACGACAGGCNNSGCGCTTCGGGGAACA
113	ATCTGTTCCCGGAAGSNNCTTGCCTGTCGTTAAAC	GTTTAACGACAGGCAAGNNSCTTCCGGGAACAGAT
114	TGGATCTGTTCCCGGSNNCGCCTTGCCTGTCG	CGACAGGCAAGGCGNNSCCGGGAACAGATCCA
115	ATTTGGATCTGTTCCSNNAAGCGCCTTGCCTGTC	GACAGGCAAGGCGCTTNNSGGAACAGATCCAAAT
116	TTGATTTGGATCTGTSNNCGGAAGCGCCTTGCC	GGCAAGGCGCTTCCGNNSACAGATCCAAATCAA
117	CTTTTGATTTGGATCSNNTCCCGGAAGCGCCTTG	CAAGGCGCTTCCGGGANNSGATCCAAATCAAAAG
118	AATCTTTTGATTTGGSNNTGTTCCCGGAAGCGC	GCGCTTCCGGGAACANNSCCAAATCAAAAGATT
119	TAAAATCTTTTGATTSNNATCTGTTCCCGGAAG	CTTCCGGGAACAGATNNSAATCAAAAGATTTTA
120	GTATAAAATCTTTTGSNNTGGATCTGTTCCCGG	CCGGGAACAGATCCANNSCAAAAGATTTTATAC
121	TGTGTATAAAATCTTSNNATTTGGATCTGTTCC	GGAACAGATCCAAATNNSAAGATTTTATACACA
122	GGATGTGTATAAAATSNNTTGATTTGGATCTG	CAGATCCAAATCAANNSATTTTATACACATCC
123	AATGGATGTGTATAASNNCTTTTGATTTGGATC	GATCCAAATCAAAAGNNSTTATACACATCCATT
124	GTAAATGGATGTGTASNNATTCTTTTGATTTGG	CCAAATCAAAAGAATNNSTACACATCCATTTAC
125	GCTGTAAATGGATGTSNNTAAAATCTTTTGATTTG	CAAATCAAAAGATTTTANNSACATCCATTTACAGC
126	ACTGCTGTAAATGGASNNGTATAAAATCTTTTG	CAAAAGATTTTATACNNSTCCATTTACAGCAGT
127	GGCACTGCTGTAAATSNNTGTGTATAAAATCTT	AAGATTTTATACACANNSATTTACAGCAGTGCC
128	ATCGGCACTGCTGTASNNGGATGTGTATAAAATC	GATTTTATACACATCCNNSTACAGCAGTGCCGAT
129	-	-
130	AATCATATCGGCACTSNNGTAAATGGATGTG	CACATCCATTTACNNSAGTGCCGATATGATT
131	GACAATCATATCGGCSNNGCTGTAAATGGATG	CATCCATTTACAGCNNSGCCGATATGATTGTC
132	CATGACAATCATATCSNNACTGCTGTAAATGG	CCATTTACAGCAGTNNSGATATGATTGTCATG
133	ATTCATGACAATCATSNNGGCACTGCTGTAAATG	CATTTACAGCAGTGCCNNSATGATTGTCATGAAT
134	GTAATTCATGACAATSNNATCGGCACTGCTGTAAATG	CATTTACAGCAGTGCCGATNNSATTGTCATGAATTAC
135	TAAGTAATTCATGACSNNCATATCGGCACTGCTG	CAGCAGTGCCGATATGNNSGTCATGAATTACTTA
136	TGATAAGTAATTCATSNNAATCATATCGGCACTG	CAGTGCCGATATGATTNNSATGAATTACTTATCA
137	TCTTGATAAGTAATTSNNGACAATCATATCGGC	GCCGATATGATTGTCNNSAATTACTTATCAAGA
138	TAATCTTGATAAGTASNNCATGACAATCATATC	GATATGATTGTCATGNNSTACTTATCAAGATTA
139	ATCTAATCTTGATAASNNATTCATGACAATCATATCGGC	GCCGATATGATTGTCATGAATNNSTTATCAAGATTAGAT
140	ACCATCTAATCTTGASNNGTAATTCATGACAATC	GATTGTCATGAATTACNNSTCAAGATTAGATGGT
141	-	-
142	TCTAGCACCATCTAASNNTGATAAGTAATTCATG	CATGAATTACTTATCANNSTTAGATGGTGCTAGA
143	GTTTCTAGCACCATCSNNTCTTGATAAGTAATTC	GAATTACTTATCAAGANNSGATGGTGCTAGAAAC
144	AACGTTTCTAGCACCSNNTAATCTTGATAAGTAATT	AATTACTTATCAAGATTANNSGGTGCTAGAAACGTT
145	-	-
146	GATTTGAACGTTTCTSNNACCATCTAATCTTG	CAAGATTAGATGGTNNSAGAAACGTTCAAATC
147	ATGGATTTGAACGTTSNNAGCACCATCTAATCTTG	CAAGATTAGATGGTGCTNNSAACGTTCAAATCCAT
148	GCCATGGATTTGAACSNNTCTAGCACCATCTAATC	GATTAGATGGTGCTAGANNSGTTCAAATCCATGGC
149	AACGCCATCGATTTGSNNGTTTCTAGCACCATC	GATGGTGCTAGAAACNNSCAAATCGATGGCGTT

Additional Tables, page 3

150	TCCAACGCCATGGATSNNAACGTTTCTAGCACC	GGTGCTAGAAACGTTNNSATCCATGGCGTTGGA
151	GTGTCCAACGCCATGSNNTTGAACGTTTCTAGC	GCTAGAAACGTTCAANNSCATGGCGTTGGACAC
152	GATGTGTCCAACGCCSNNGATTTGAACGTTTCTAG	CTAGAAACGTTCAAATCNNSGGCGTTGGACACATC
153	-	-
154	AAGGCCGATGTGTCCSNNGCCATGGATTTGAAC	GTTCAAATCCATGGCNNSGGACACATCGGCCTT
155	CAGAAGGCCGATGTGSNNAACGCCATGGATTTG	CAAATCCATGGCGTTNNSCACATCGGCCTTCTG
156	GTACAGAAGGCCGATSNNTCCAACGCCATGGATTTG	CAAATCCATGGCGTTGGANNSATCGGCCTTCTGTAC
157	GCTGTACAGAAGGCCSNNGTGTCCAACGCCATG	CATGGCGTTGGACACNNSGGCCTTCTGTACAGC
158	GCTGCTGTACAGAAGSNNGATGTGTCCAACGCC	GGCGTTGGACACATCNNSCTTCTGTACAGCAGC
159	TTGGCTGCTGTACAGSNNGCCGATGTGTCCAAC	GTTGGACACATCGGCNNSCTGTACAGCAGCCAA
160	GACTTGGCTGCTGTASNNAAGGCCGATGTGTCC	GGACACATCGGCCTTNNSTACAGCAGCCAAGTC
161	GTTGACTTGGCTGCTSNNCAGAAGGCCGATGTG	CACATCGGCCTTCTGNNSAGCAGCCAAGTCAAC
162	GCTGTTGACTTGGCTSNNGTACAGAAGGCCGATG	CATCGGCCTTCTGTACNNSAGCCAAGTCAACAGC
163	CAGGCTGTTGACTTGSNNGCTGTACAGAAGGCC	GGCCTTCTGTACAGCNNSCAAGTCAACAGCCTG
164	AATCAGGCTGTTGACSNNGCTGCTGTACAGAAG	CTTCTGTACAGCAGCNNSGTCAACAGCCTGATT
165	-	-
166	TTCTTTAATCAGGCTSNNGACTTGGCTGCTGTAC	GTACAGCAGCCAAGTCNNSAGCCTGATTAAAGAA
167	CCCTCCTTTAATCAGSNNGTTGACTTGGCTGCTG	CAGCAGCCAAGTCAACNNSCTGATTAAAGGAGGG
168	CAGCCCTTCTTTAATSNNGCTGTTGACTTGGCTGCTG	CAGCAGCCAAGTCAACAGCNNSATTAAAGAAGGGCTG
169	GTTCAGCCCTTCTTTSNNCAGGCTGTTGACTTG	CAAGTCAACAGCCTGNNSAAAGAAGGGCTGAAC
170	GCCGTTCAGCCCTTCSNNAATCAGGCTGTTGAC	GTCAACAGCCTGATTNNSGAAGGGCTGAACGGC
171	CCCGCCGTTCAGCCCSNNTTTAATCAGGCTGTT	AACAGCCTGATTAAANNSGGGCTGAACGGCGGG
172	GCCCCCGCCGTTCAGSNNTTCTTTAATCAGGCT	AGC CTG ATT AAA GAA NNS CTG AAC GGC GGG GGC
173	CTGGCCCCCGCCGTTSNNCCCTTCTTTAATCAG	CTG ATT AAA GAA GGG NNS AAC GGC GGG GGC CAG
174	TTCTGGCCCCCGCCSNNCAGCCCTTCTTTAATC	GATTAAAGAAGGGCTGNNSGGCGGGGGCCAGAA
175	CGTATTCTGGCCCCCSNNGTTCAGCCCTTCTTT	AAA GAA GGG CTG AAC NNS GGG GGC CAG AAT ACG
176	ATTCGTATTCTGGCCSNNGCCGTTCAGCCCTTC	GAA GGG CTG AAC GGC NNS GGC CAG AAT ACG AAT
177	TTAATTCGTATTCTGSNNCCCGCCGTTCAGCCC	GGG CTG AAC GGC GGG NNS CAG AAT ACG AAT TAA
178	GCTTGTCGACGGAGCTCTCATTAATTCGTATTSNNGCC	GGCNNSAATACGAATTAATGAGAGCTCCGTCGACAAGC
179	GCTTGTCGACGGAGCTCTCATTAATTCGTSNNCTC	GAGNNSACGAATTAATGAGAGCTCCGTCGACAAGC
180	GCTTGTCGACGGAGCTCTCATTAATTSNNATT	AATNNSAATTAATGAGAGCTCCGTCGACAAGC
181	GCTTGTCGACGGAGCTCTCATTASNNCGT	ACGNNSTAATGAGAGCTCCGTCGACAAGC

Table S2 Oligonucleotide sequences for generation of *lipA* site directed single and double mutants

The forward and reverse oligonucleotide sequence is shown for each variant. Modification sites are underlined.

variant	forward primer	reverse primer
G13S	ATGGTTCACGGTATTTCGGGGGCATCATTCAAT	ATTGAATGATGCCCCCGAAATACCGTGAACCAT
A105N	GTGACGCTTGGCGGCAACAACCGTTTGACGACA	TGTCGTCAAACGGTTGTTGCCGCCAAGCGTCAC
Y139T	ATGATTGTCATGAATACCTTATCAAGATTAGAT	ATCTAATCTTGATAAGGTATTCATGACAATCAT

Table S3 LipA amino acid sequence conservation

A: UniProtKB accession numbers and original organisms for the 64 lipase sequences out of 41 species from the *Firmicutes* phylum used for the alignment. **B**: The number of identical amino acids in this alignment like in *B. subtilis* LipA was counted and calculated in percentage frequency for each position to determine the conservation of this amino acid within the *Firmicutes* phylum. The amino acid position (position), the amino acid (aa) and the percentaged conservation are shown.

A

Lipase	Species	Lipase	Species	Lipase	Species	Lipase	Species
P94444	*Bacillus* sp. BP-6	**Q8VU78**	*Bacillus* sp. B26	**H0FRJ5**	*Bacillus amyloliquefaciens IT-45*	**E0U0Y0**	*Bacillus subtilis spizizenii* ATCC 23059
Q79F14	*Bacillus subtilis* 168	**H6U4T6**	*Bacillus* sp. enrichment culture clone S6	**H2ABY2**	*Bacillus amyloliquefaciens* subsp. *plantarum* CAU B946	**G4NTQ6**	*Bacillus subtilis spizizenii* TU-B-10
B8YLY0	*Bacillus subtilis*	**B2L2K1**	*Bacillus licheniformis*	**H8XE51**	*Bacillus amyloliquefaciens* subsp. *plantarum* YAU B9601-Y2	**E5W0L6**	*Bacillus sp.* BT1B_CT2
Q8RJP5	*Bacillus megaterium*	**A1E152**	*Bacillus pumilus*	**F4E233**	*Bacillus amyloliquefaciens* TA208	**Q65HR4**	*Bacillus licheniformis* ATCC 14580
E8VEC0	Bacillus subtilis SC-8	**A8FGA4**	*Bacillus pumilus* SAFR-032	**G0IK64**	*Bacillus amyloliquefaciens* XH7	**I0UHQ4**	*Bacillus licheniformis* WX-02
G4EYR4	*Bacillus subtilis* SC-8	**Q9K5F4**	*Bacillus licheniformis*	**E3DTQ6**	*Bacillus atrophaeus* 1942	**B3F2Y4**	*Bacillus sp.* RN2
G4P2C8	*Bacillus subtilis* RO-NN-1	**Q6RSN0**	*Bacillus pumilus*	**A5HLW9**	*Bacillus subtilis*	**H6NI24**	*Paenibacillus mucilaginosus* 3016
D5N1Z7	*Bacillus subtilis* subsp. *spizizenii* ATCC 6633	**B1PN85**	*Bacillus pumilus*	**G4PA03**	*Bacillus subtilis* RO-NN-1	**I0BL71**	*Paenibacillus mucilaginosus* K02
E0TW96	*Bacillus subtilis ATCC 23059*	**B4ANV6**	*Bacillus pumilus* ATCC 7061	**B1PN84**	*Bacillus subtilis*	**F8FBS6**	*Paenibacillus mucilaginosus* KNP414
G4NRF1	*Bacillus subtilis* TU-B-10	**Q2LAN2**	*Bacillus pumilus*	**D4G4R9**	*Bacillus subtilis* subsp. *natto* BEST195	**Q5WDN0**	*Bacillus clausii* KSM-K16
D4G6J8	*Bacillus subtilis* subsp. *natto* BEST195	**B7VF67**	*Bacillus pumilus*	**E8VK85**	*Bacillus subtilis* BSn5	**Q8RC83**	*Caldanaerobacter subterraneus* subsp. *tengcongensis* DSM 15242
B0LW76	*Bacillus* sp. NK13	**B8Y3H3**	*Bacillus pumilus*	**G4F0D0**	*Bacillus subtilis* SC-8	**Q6WUB2**	*Caldanaerobacter subterraneus* subsp. *tengcongensis*
D5E2W8	*Bacillus megaterium* ATCC 12872	**B2CX98**	*Bacillus pumilus*	**B7UDC5**	*Bacillus subtilis*	**F1ZT35**	*Thermoanaerobacter ethanolicus* JW 200
D3WK98	*Bacillus pumilus*	**Q2L991**	*Bacillus pumilus*	**I0F008**	*Bacillus* sp. JS	**G2MS56**	*Thermoanaerobacter wiegelii* Rt8.B1
E2CYQ9	*Bacillus pumilus*	**D7URU5**	*Bacillus* sp. HH-01	**Q83ZY1**	*Bacillus subtilis*	**D3FQU1**	*Bacillus pseudofirmus* OF4
A4GUJ6	*Bacillus pumilus*	**A7Z124**	*Bacillus velezensis* DSM 23117	**D5N2V3**	*Bacillus subtilis* subsp. *spizizenii* ATCC 6633	**P37957**	*Bacillus subtilis 168*

Additional Tables, page 5

Skoczinski *et al.*, 2017: LipA single substitutions

B

position	aa	%	position	aa	%	position	aa	%	position	aa	%
1	A	55	52	G	66	103	G	97	154	V	69
2	E	43	53	P	85	104	G	91	155	G	78
3	H	83	54	V	20	105	A	91	156	H	92
4	N	83	55	L	92	106	N	98	157	I	85
5	P	97	56	S	65	107	R	29	158	G	91
6	V	95	57	R	48	108	L	91	159	L	91
7	V	89	58	F	52	109	T	23	160	L	92
8	M	65	59	V	80	110	T	49	161	Y	11
9	V	98	60	Q	28	111	G	15	162	S	49
10	H	97	61	K	38	112	K	23	163	S	94
11	G	100	62	V	98	113	A	89	164	Q	80
12	I	74	63	L	92	114	L	78	165	V	97
13	G	72	64	D	35	115	P	77	166	N	62
14	G	91	65	E	52	116	G	86	167	S	22
15	A	86	66	T	97	117	T	78	168	L	43
16	S	77	67	G	98	118	D	78	169	I	80
17	F	28	68	A	94	119	P	75	170	K	85
18	N	92	69	K	72	120	N	80	171	E	78
19	F	94	70	K	97	121	Q	78	172	G	85
20	A	62	71	V	100	122	K	85	173	L	92
21	G	38	72	D	92	123	I	85	174	N	75
22	I	97	73	I	100	124	L	78	175	G	75
23	K	89	74	V	92	125	Y	92	176	G	83
24	S	45	75	A	95	126	T	92	177	G	86
25	Y	89	76	H	98	127	S	91	178	Q	55
26	L	94	77	S	98	128	I	63	179	N	78
27	V	48	78	M	91	129	Y	97	180	T	71
28	S	65	79	G	98	130	S	91	181	N	71
29	Q	88	80	G	94	131	S	60			
30	G	92	81	A	91	132	A	57			
31	W	89	82	N	91	133	D	92			
32	S	32	83	T	86	134	M	22			
33	R	52	84	L	91	135	I	91			
34	D	22	85	Y	89	136	V	91			
35	K	43	86	Y	97	137	M	26			
36	L	78	87	I	89	138	N	91			
37	Y	77	88	K	85	139	Y	26			
38	A	97	89	N	75	140	L	91			
39	V	26	90	L	89	141	S	98			
40	D	83	91	D	69	142	R	77			
41	F	92	92	G	91	143	L	89			
42	W	20	93	G	86	144	D	28			
43	D	85	94	N	38	145	G	97			
44	K	88	95	K	89	146	A	83			
45	T	85	96	V	34	147	R	58			
46	G	98	97	A	25	148	N	89			
47	T	20	98	N	78	149	V	65			
48	N	92	99	V	86	150	Q	60			
49	Y	28	100	V	83	151	I	69			
50	N	74	101	T	89	152	H	58			
51	N	91	102	L	80	153	G	97			

Table S4 Summary of relative transcript amount, specific activity and protein amount of 38 characterized LipA variants

The table shows the structural position and location of the variant's amino acid substitution. The wild-type (wt) codon and the introduced variant codon are named together with the codon frequency per 1000 bp. Variant I12F as variant with increased specific activity as well as protein amount is shown twice.

false-positive LipA variants										
variant	position in secondary structure	location	wt codon	frequency per 1000bp	variant codon	frequency per 1000bp	rel. change in transcript level[2]	lower and upper deviation in transcript level	rel. specific activity ± standard deviation	rel. lipase amount ± standard deviation
$I12L_{TTG}$	turn	s	ATT	36.2	TTG	15.8	1.3	0.4 0.7	0.8 ± 0.3	1.0 ± 0.4
$I12V_{GTC}$	turn	s	ATT	36.2	GTC	17.3	1.4	0.4 0.7	0.8 ± 0.6	1.0 ± 0.3
G13N	turn	s	GGA	21.8	AAC	17.8	1	0.4 0.6	0.7 ± 0.4	0.8 ± 0.3
Q29H	turn	s	CAG	18.5	CAC	7.5	0.8	0.2 0.2	0.9 ± 0.4	1.0 ± 0.3
T47H	coil	s	ACA	21.6	CAC	7.5	0.3	0.3 1.0	0.9 ± 0.4	1.2 ± 0.3
T47P	coil	s	ACA	21.6	CCA	7.4	1.5	0.4 0.4	1 ± 0.3	1.0 ± 0.3
T47T	coil	s	ACA	21.6	ACG	14.9	1.4	0.4 0.6	1.0 ± 0.4	0.7 ± 0.3
N48G	αB	s	AAT	22.9	GGC	23.3	1.2	0.2 0.4	0.3 ± 0.3	1.4 ± 0.3
L55F	αB	b	TTA	19.8	TTC	14.3	1	0.2 0.2	0.4 ± 0.4	1.2 ± 0.3
T83M	αC	b	ACA	21.6	ATG	26.3	1.5	0.3 0.5	0.6 ± 0.4	0.9 ± 0.4
Y85W	αC	s	TAC	12.6	TTG	15.8	1.4	0.4 0.8	0.7 ± 0.4	0.6 ± 0.3
I87L	αC	s	ATA	9.8	CTC	10.7	1	0.9 6.6	0.1 ± 0.4	1.1 ± 0.3

Skoczinski *et al.*, 2017: LipA single substitutions

LipA variants with increased extracellular specific lipase activity									
variant	positicin in secondary structure	location	wt codon	frequency per 1000bp	variant codon	frequency per 1000bp	rel. change in transcript level[2]	lower and upper deviation in transcript level	rel. specific activity ± standard deviation	rel. lipase amount ± standard deviation
I12F	turn	s	ATT	36.2	TTC	14.3	1.7	0.4 0.5	2.1* ± 0.4	1.6* ± 0.2
$I12L_{CTG}$	turn	s	ATT	36.2	CTG	23.0	1.0	0.4 0.7	2.4* ± 0.3	0.9 ± 0.3
$I12V_{GTG}$	turn	s	ATT	36.2	GTG	17.3	1.3	0.3 0.6	1.8* ± 0.4	1.0 ± 0.3
G13S	turn	s	GGA	21.8	TCG	6.5	1.4	0.5 0.5	1.4* ± 0.3	0.6 ± 0.4
$G13T_{ACC}$	turn	s	GGA	21.8	ACC	9.0	1.2	0.4 0.5	2.3* ± 0.3	0.5 ± 0.2
I87I	αC	s	ATA	9.8	ATC	27.2	3.6*	0.6 0.9	2.3* ± 0.3	0.7 ± 0.4
LipA variants with higher extracellular LipA amount										
variant	positicin in secondary structure	location	wt codon	frequency per 1000bp	variant codon	frequency per 1000bp	rel. change in transcript level[2]	lower and upper deviation in transcript level	rel. specific activity ± standard deviation	rel. lipase amount ± standard deviation
I12F	turn	s	ATT	36.2	TTC	14.3	1.7	0.4 0.5	2.1* ± 0.4	1.6* ± 0.2
$G13T_{ACG}$	turn	s	GGA	21.8	ACG	14.9	2.7*	0.9 1.2	0.8 ± 0.4	1.6* ± 0.3
F17E	αA	s	TTC	14.3	GAG	22.6	2	0.6 0.9	0.5 ± 0.4	1.3* ± 0.3
N48Q	αB	s	AAT	22.9	CAG	18.5	2.2	0.6 1.2	0.6 ± 0.3	2.2* ± 0.2

Additional Tables, page 8

N50D	αB	s	AAC	17.8	GAC	19.0	0.8	0.3 0.4	0.8 ± 0.5	1.3* ± 0.3
P53D	αB	s	CCG	16.3	GAC	19.0	1.0	0.2 0.3	0.7 ± 0.4	1.9* ± 0.3
P53E	αB	s	CCG	16.3	GAG	22.6	1.0	0.1 0.2	0.9 ± 0.4	1.5* ± 0.3
P53V	αB	s	CCG	16.3	GTG	17.3	1.3	0.3 0.5	0.7 ± 0.3	1.4* ± 0.3
R57T$_{ACC}$	αB	s	CGA	4.3	ACC	9.0	0.8	0.3 0.7	0.8 ± 0.4	1.5* ± 0.3
R57T$_{ACG}$	αB	s	CGA	4.3	ACG	14.9	0.9	0.2 0.3	0.7 ± 0.4	1.4* ± 0.2
I87V	αC	s	ATA	9.8	GTG	17.3	1.3	1.2 8.6	1.1 ± 0.3	1.6* ± 0.3
K88K	αC	s	AAA	48.4	AAG	20.8	1.6	0.4 0.5	0.4 ± 0.3	1.4* ± 0.3
A105N	coil	s	GCG	19.8	AAC	17.8	1.2	0.3 0.3	0.5 ± 0.3	2.0* ± 0.3
M134K	coil	s	ATG	26.3	AAG	20.8	1.1	0.3 0.5	0.9 ± 0.4	2.3* ± 0.3
M134P	coil	s	ATG	26.3	CCG	16.3	0.6	0.1 0.1	0.2 ± 0.3	2.5* ± 0.2
M134Q	coil	s	ATG	26.3	CAG	18.5	0.8	0.2 0.2	0.4 ± 0.3	2.1* ± 0.3
Y139G	αE	s	TAC	12.6	GGG	11.2	1.7	0.7 0.8	0.6 ± 0.4	1.5* ± 0.3
Y139T	αE	s	TAC	12.6	ACG	14.9	2.1	0.5 0.5	0.5 ± 0.4	1.8* ± 0.3
L140A	αE	s	TTA	19.8	GCG	19.8	1.6	0.4 0.5	0.4 ± 0.3	2.2* ± 0.2
L140Y	αE	s	TTA	19.8	TAC	12.6	1.6	0.5 1.3	0.6 ± 0.4	1.8* ± 0.2
V154E	coil	s	GTT	18.6	GAG	22.6	0.9	0.3 0.4	0.6 ± 0.3	1.3* ± 0.2

Combined LipA variants									
variant	position in secondary structure	location[1]	wt codon	frequency per 1000bp	variant codon	frequency per 1000bp	rel. change in transcript level[2]	lower and upper deviation in transcript level	rel. specific activity ± standard deviation	rel. lipase amount ± standard deviation
G13S	turn	s	GGA	21.8	TCG	6.5	1.8	0.6 1.2	2.5* ± 0.5	1.0 ± 0.5
A105N	coil	s	GCG	19.8	AAC	17.8	2.1	0.6 1.3	0.8 ± 0.3	2.4* ± 0.4
Y139T	αE	s	TAC	12.6	ACG	14.9	0.9	0.2 0.2	1.1 ± 0.4	2.2* ± 0.3
G13S, A105N							2.1	0.6 1.2	2.9* ± 0.4	1.4* ± 0.5
G13S, Y139T							1.2	0.2 0.2	0.8 ± 0.5	1.4* ± 0.4
A105N, Y139T							1.4	0.2 0.3	0.4 ± 0.4	3.6* ± 0.4

[1]: s: surface exposed; b: buried;

[2]: significant transcript changes compared to wtLipA above the cutoff of 2.2 and a *p*-value < 0.05;

*: significantly increased relative extracellular specific lipase activity or extracellular lipase amount compared to wtLipA with a *p*-value < 0.05.

Additional Methods

B. subtilis wtLipA production analysis

B. subtilis TEB1030 with the plasmid pBSlipA encoding for wtLipA was cultivated as described for the 48-well FlowerPlate® cultivation in the manuscript´s method section. 1 ml cells were harvested after 2, 4, 6, 8, 10, and 24 h by centrifugation (room temperature, 21,000 g, 5 min). The culture supernatant and the cells, resuspended in 50 mM Tris-HCl pH 8, were used for a lipase activity assay as described in the manuscript´s method section. For online biomass measurement by scattered light ($O.D._{600nm}$), replicates were prepared in 48-well Flowerplates and cultivated in the BioLector® (m2p-labs, Germany) under identical conditions (37 °C, 1,100 rpm) for 24 h.

Protein TCA-NaDoc precipitation

A sample volume of 1 ml was mixed with 100 µl cold 10 % (w/v) NaDoc (sodium desoxycholate) and incubated on ice for 10 min. After addition of 100 µl cold 40 % (v/v) TCA and incubation on ice for 20 min, the sample was centrifuged at 4 °C, 21,000 g for 30 min. The supernatant was discarded and the pellet containing the proteins was washed with 500 µl 80 % (v/v) acetone. After discarding the supernatant, the pellet was dried for 5 min. The pellet was resuspended in 50 mM Tris-HCl pH 8 and 2x SDS sample buffer (50 mM Tris-HCl pH 6.8, 4 % (w/v) SDS, 10 % (v/v) glycerol, 2 % (v/v) β-mercaptoethanol, 0.03 % (w/v) Bromophenol blue) to a concentration corresponding to a cell density of $O.D._{580nm}$ = 15 and boiled for 10 min.

Protein separation by SDS-PAGE

Boiled samples were loaded onto a 5 % stacking gel (2.8 ml *A. dest.*, 0.83 ml 37 % (v/v) acrylamide, 1.3 ml Tris-HCl pH 6.8 (0.5 M), 50 µl 10 % (w/v) SDS, 50 µl 10 % (w/v) APS, 5 µl TEMED) on top of a 16 % separation gel (2.1 ml *A. dest.*, 5.3 ml 37 % (v/v) acrylamide, 2.5 ml Tris-HCl pH 8.8 (0.5 M), 100 µl 10 % (w/v) SDS, 100 µl 10 % (w/v) APS, 10 µl TEMED). Discontinuous SDS-gel electrophoresis was carried out at 100 V for 15 min and at 200 V for 40 min using the gadget „Mini Protean II Dual Slap Cell“ (BioRad Laboratories GmbH, Germany) and SDS running buffer (0.025 M Tris, 0.2 M glycine, 0.003 M SDS).

Immunodetection of proteins *via* Western blotting

Proteins from SDS gels were electrophoretically transferred at 150 mA for 15 min, and at 300 mA for 60 min onto a polyvinylidene difluoride (PVDF) membrane in a Mini-Protean 3 Cell (BioRad Laboratories GmbH, Germany) in 1 x Dunn carbonate buffer (0.003 M Na_2CO_3, 0.01

M $NaHCO_3$) with 20 % (v/v) methanol. The PVDF membranes were washed in methanol and *A. dest.* for 1 min before protein transfer. The membrane was blocked with 3 % (w/v) bovine serum albumin dissolved in TBST (0.025 M Tris, 0.15 M NaCl, 0.0015 M KCl, 0.02 % (v/v) Tween 20) at 4 °C for 16 h. The membranes were incubated with a specific polyclonal LipA antibody (Eurogentec, Germany; produced in rabbits immunized with *B. subtilis* LipA overproduced in *E. coli* BL21(DE3)) in dilution of 1:20,000 in TBST and a second antibody goat-anti-rabbit HRP conjugate (BioRad Laboratories GmbH, Germany) in dilution 1:5,000 in TBST for 1h. After each antibody incubation step, the membranes were washed in TBST at room temperature for 30 min and 3 x 10 min. All incubation steps were accomplished on an orbital mixer. Signals were detected using freshly prepared ECL solution and the Stella 3200 Imaging System (Raytest, Germany). The ECL solution was prepared by mixing 1 ml of 4 °C cold solution A (0.025 % (w/v) luminol, 0.1 M Tris-HCl pH 8.6) with 100 µl solution B (0.1 % (w/v) *p*-hydroxy coumarate in 100 % DMSO) and 0.3 µl solution C (30 % H_2O_2).

Constraint Network Analysis (CNA)

For CNA, a protein is represented as a constraint network, where atoms are nodes and covalent and non-covalent interactions form constraints connecting the nodes [1]. The constraints in the network are modeled with different numbers of bars depending on the type and strength of the interaction. Taking into account that the network nodes are considered bodies with six degrees of freedom, covalent single bonds are modeled as five bars (leaving the rotational degree of freedom unlocked), double and peptide bonds as six bars (freezing any relative motion between two bodies), non-covalent hydrogen bonds (including salt bridges) are modeled as five bars, and hydrophobic interactions as two bars. The hydrogen bond energy (E_{HB}) for all hydrogen bonds is computed according to a potential by Dahiyat *et al.* [2]. For thermal unfolding simulations [3, 4], hydrogen bonds are removed from the network in increasing order of their strength: A hydrogen bond is discarded from the network if $E_{HB} > E_{cut}$. In the present study, E_{cut} was varied from -0.1 kcal mol^{-1} to -0.4 kcal mol^{-1} (according to 302 K to 380 K [4]) with a step size of 0.1 kcal mol^{-1} (2 K), as done previously for investigation of the thermostability of LipA [5]. For each network state generated that way, rigid and flexible regions are determined by the program FIRST [6], and from this the local index $rc_{ij,neighbor}$ [7, 8]. $rc_{ij,neighbor}$, a neighbor stability map, characterizes the local rigidity of a protein. For improving the robustness of the analyses [9], CNA was performed on ensembles of network topologies (ENT) generated by the ENT^{FNC} approach, as done previously [5, 10]. The parameter $\widetilde{rc}_{ij,\ neighbor}$ was then computed as the median of $rc_{ij,neighbor}$ averaged over the respective 5,000 conformations. $\widetilde{rc}_{ij,\ neighbor}$ is related to the thermodynamic thermostability of a protein [8].

1. Pfleger C, Rathi PC, Klein DL, Radestock S, Gohlke H. Constraint Network Analysis (CNA): A python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. J Chem Inf Model. 2013;53:1007–15. http://dx.doi.org/10.1021/ci400044m.

2. Dahiyat BI, Benjamin Gordon D, Mayo SL. Automated design of the surface positions of protein helices. Protein Sci. 1997;6:1333–7. http://dx.doi.org/10.1002/pro.5560060622.

3. Radestock S, Gohlke H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng Life Sci. 2008;8:507–22. http://dx.doi.org/10.1002/elsc.200800043.

4. Radestock S, Gohlke H. Protein rigidity and thermophilic adaptation. Proteins Struct Funct Bioinforma. 2011;79:1089–108. http://dx.doi.org/10.1002/prot.22946.

5. Rathi PC, Fulton A, Jaeger K-E, Gohlke H. Application of rigidity theory to the thermostabilization of lipase A from *Bacillus subtilis*. PLoS Comput Biol. 2016;12:e1004754. http://dx.doi.org/10.1371/journal.pcbi.1004754.

6. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. Proteins Struct Funct Genet. 2001;44:150–65. http://dx.doi.org/10.1002/prot.1081.

7. Pfleger C, Radestock S, Schmidt E, Gohlke H. Global and local indices for characterizing biomolecular flexibility and rigidity. J Comput Chem. 2013;34:220–33. http://dx.doi.org/10.1002/jcc.23122.

8. Rathi PC, Jaeger K, Gohlke H. Structural rigidity and protein thermostability in variants of lipase A from *Bacillus subtilis*. PLoS One. 2015;1–24. http://dx.doi.org/10.1371/journal.pone.0130289.

9. Rathi PC, Radestock S, Gohlke H. Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. J Biotechnol. 2012;159:135–44. http://dx.doi.org/10.1016/j.jbiotec.2012.01.027.

10. Pfleger C, Gohlke H. Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. Structure. 2013;1–10. http://dx.doi.org/10.1016/j.str.2013.07.012.

Additional Figures

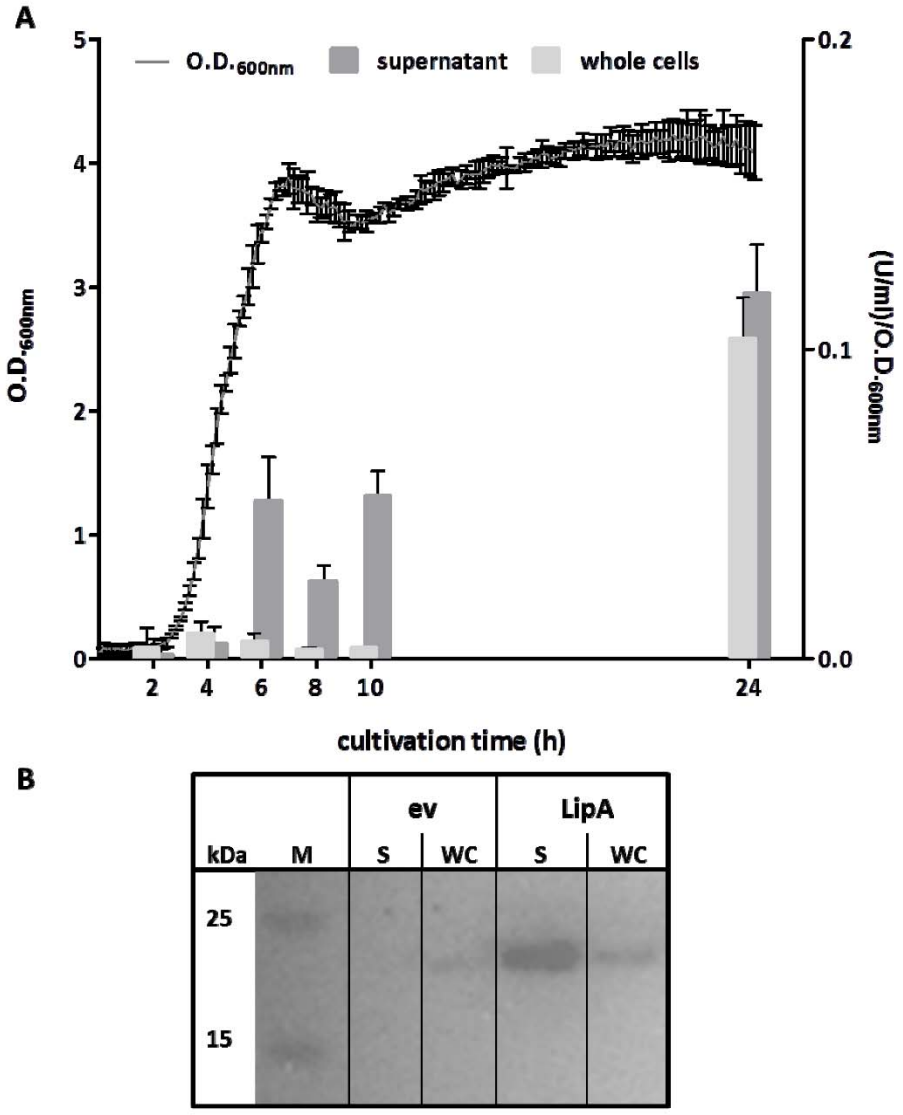


Figure S1 Microfermentation of *B. subtilis* TEB1030 producing wtLipA

A Wild-type LipA production analysis in *B. subtilis* TEB1030. *B. subtilis* TEB1030 producing wtLipA was cultivated for 24 h in a microfermentation system using a 48 well Flowerplate® and online biomass measurement was performed in the BioLector®. The cultivation time (h) is plotted against the optical density at 600 nm on the left handed y-axis and against the volume activity normalized to the optical density ((U/ml)/O.D.$_{600nm}$). The grey line with error bars show *B. subtilis* TEB1030 growth producing wtLipA (O.D.$_{600nm}$). After 2, 4, 6, 8, 10 and 24 h of cultivation samples were taken to determine the lipase activity in the *B. subtilis* culture supernatant (bars in dark grey) and the *B. subtilis* whole cells (bars in light grey) that was normalized to the *B. subtilis* growth at the corresponding sampling time point. **B Western Blot analysis of *B. subtilis* whole cells and culture supernatant after 6 h of wtLipA production.** The culture supernatants of *B. subtilis* TEB1030 harboring the empty vector pBSMul1 (ev) and the *lipA* expression vector pBSlipA (LipA) were precipitated with trichloroacetic acid. The precipitated culture supernatant (S) and the whole cells (WC) were resuspended in 50 mM Tris-HCl pH 8 to an O.D.$_{580nm}$ of 15. 10 µl of each sample were applied on a 16 % discontinuous SDS-PAGE together with a molecular weight standard (M). Immunodetection was performed using a specific polyclonal LipA antibody.

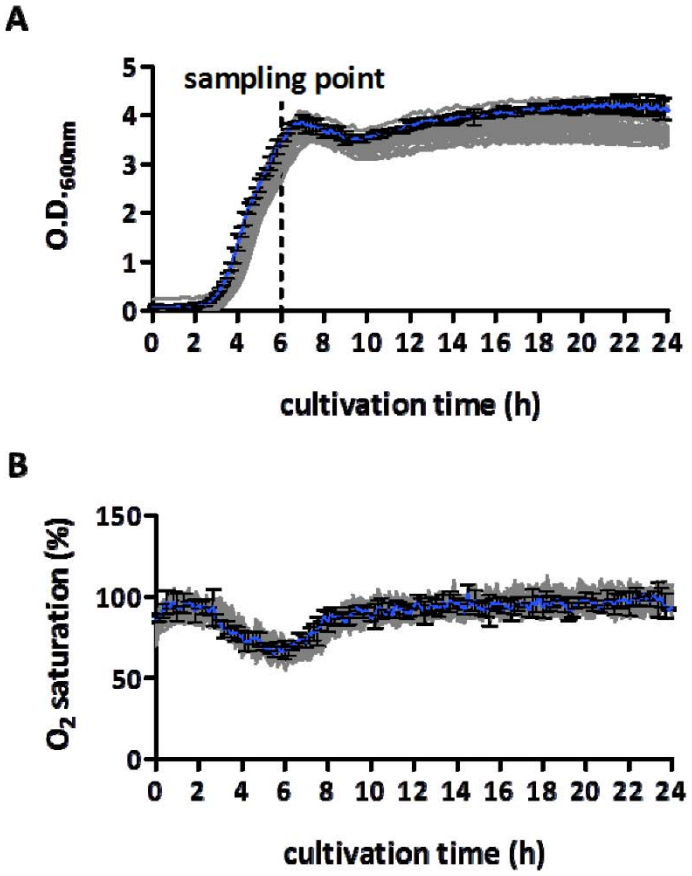


Figure S2 Microfermentation of *B. subtilis* TEB1030 producing the 38 different LipA variants
A 24 h online biomass measurement. *B. subtilis* TEB1030 harboring the 38 different LipA variants were cultivated in a microfermentation system using 48 well FlowerPlates®. Online biomass measurement was performed for 24 h in the BioLector®. The cultivation time (h) is plotted against the optical density at 600 nm (O.D. $_{600nm}$). The blue line indicates wtLipA optical density with corresponding error bars in black. The sampling point after 6 h of LipA production is marked. **B Oxygen saturation during microfermentation.** The cultivation time (h) of *B. subtilis* TEB1030 harboring the 38 different LipA variants is plotted against the percentaged oxygen saturation (%). The blue line indicates wtLipA oxygen saturation with corresponding error bars in black.

CURRICULUM VITAE

Education

Since 11/2016	**PhD student**, Heinrich Heine University Düsseldorf, Germany
10/2011 – 10/2016	**Studies in biochemistry**, Heinrich Heine University Düsseldorf, Germany
04/2016 – 10/2016	**Master thesis**, Heinrich Heine University Düsseldorf, Germany,
06/2014 – 09/2014	**Bachelor thesis**, Heinrich Heine University Düsseldorf, Germany,

Publications

Nutschel, C., Coscolín, C., Mulnaes, D., David, B., Ferrer, M., Jaeger, K.-E., Gohlke, H. *Promiscuous esterases counterintuitively are less flexible than specific ones.* **J Chem Inf Model**. 2020, DOI: 10.1021/acs.jcim.1c00152.

Nutschel, C., Fulton, A., Zimmermann, O., Schwaneberg, U., Jaeger, K.-E., Gohlke, H. *Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for Bacillus subtilis lipase A.* **J Chem Inf Model.** 2020, 60, 1568-1584.

Skoczinski, P., Volkenborn, K., Fulton, A., Bhadauriya, A., Nutschel, C., Gohlke, H., Knapp, A., Jaeger, K.-E. *Contribution of single amino acid and codon substitutions to the production and secretion of a lipase by Bacillus subtilis.* **Microb Cell Fact.** 2017, 16, 160.

Hermans, S.M.A., Pfleger, C., Nutschel, C., Hanke, C.A., Gohlke, H. *Rigidity theory for biomolecules: Concepts, software, and applications.* **WIREs Comput Mol Sci.** 2017, e1311.

Kaschner, M., Schillinger, O., Fettweiss, T., Nutschel, C., Fulton, A., Strodel, B., Stadler, A., Jaeger, K.-E., Krauss, U. *A combination of mutational and computational scanning guides the design of an artificial ligand-binding controlled lipase.* **Sci Rep.** 2017, 7, 42592.

Awards

04/2019	Third lecture award of 33rd Molecular Modeling Workshop (MMW), Erlangen, Germany
11/2017	Master thesis award **„CIC-Förderpreis für Computational Chemistry**“ of Gesellschaft Deutscher Chemiker (GDCh), Mainz, Germany

Oral Conference Contributions

11/2019	**15th German Conference on Chemoinformatics (GCC)**, Mainz, Germany, Topic: *Large-scale analysis of esterase substrate promiscuity- Are predictors of active site flexibility ready for it?*
04/2019	**33rd Molecular Modeling Workshop (MMW)**, Erlangen, Germany, Topic: *Large-scale analysis of protein thermostability and detergent tolerance* (Third lecture award)
01/2019	**International Conference on Advances in Materials Science & Applied Biology (AMSAB)**, Mumbai, India, Topic: *Large-scale analysis of protein thermostability and detergent tolerance*
11/2017	**13th German Conference on Chemoinformatics (GCC)**, Mainz, Germany, Topic: *Large-scale analysis of protein stability: Bacillus subtilis lipase A as test case* (Master thesis award)

REFERENCES

1. Aehle W. *Enzymes in industry: production and applications.* 3rd ed: John Wiley & Sons; 2007.

2. Buchholz K, Kasche V, Bornscheuer UT. *Biocatalysts and enzyme technology.* 2nd ed: John Wiley & Sons; 2012.

3. Robinson PK. Enzymes: principles and biotechnological applications. *Essays Biochem.* 2015;59:1-41.

4. Jaeger K-E, Liese A, Syldatk C. *Einführung in die Enzymtechnologie*: Springer Spektrum, Berlin, Heidelberg; 2018.

5. Kendrew JC, Dickerson RE, Strandberg BE, *et al.* Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution. *Nature.* 1960;185:422-427.

6. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. *Nature.* 1977;267:585-590.

7. Payen A, Persoz J-F. Mémoire sur la diastase, les principaux produits de ses réactions, et leurs applications aux arts industriels. *Ann Chim Phys.* 1833;53:73-92.

8. Payen A, Persoz J-F. Memoir on diastase, the principal products of its reactions, and their applications to the industrial arts. *Ann Chim Phys.* 1833;53:73-92.

9. Berzelius JJ. Sur un Force Jusqu'ici Peu Remarquée qui est Probablement Active Dans la Formation des Composés Organiques, Section on Vegetable Chemistry. *Jahres-Bericht.* 1835;14:237.

10. Wisniak J. The history of catalysis. From the beginning to Nobel Prizes. *Educ Quimica.* 2010;21:60-69.

11. Pandey A, Webb C, Soccol CR, Larroche C. *Enzyme technology*: Springer Science & Business Media; 2006.

12. Priyadarshini A, Pandey P. *Biocatalysis and Agricultural Biotechnology: Fundamentals, Advances, and Practices for a Greener Future*: Apple Academic Press; 2018.

13. Buchner E. Alkoholische Gährung ohne Hefezellen. *Berichte der deutschen chemischen Gesellschaft.* 1897;30:117-124.

14. Fischer E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft.* 1894;27:2985-2993.

15. Koshland DE. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A.* 1958;44:98-104.

16. Foote J, Milstein C. Conformational isomerism and the diversity of antibodies. *Proc Natl Acad Sci U S A.* 1994;91:10370-10374.

17. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol.* 2009;5:789-796.

18. Vogt AD, Pozzi N, Chen Z, Di Cera E. Essential role of conformational selection in ligand binding. *Biophys Chem.* 2014;186:13-21.

19. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. *Protein Eng.* 1999;12:713-720.

20. Savir Y, Tlusty T. Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition. *PloS One.* 2007;2:e468.

21. Radestock S, Gohlke H. Protein rigidity and thermophilic adaptation. *Proteins.* 2011;79:1089-1108.

22. Radestock S, Gohlke H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng Life Sci.* 2008;8:507-522.

23. Ekroos M, Sjögren T. Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc Natl Acad Sci U S A.* 2006;103:13682-13687.

24. Skopalík J, Anzenbacher P, Otyepka M. Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *J Phys Chem B.* 2008;112:8165-8173.

25. Hendrychová T, Anzenbacherová E, Hudeček J, *et al.* Flexibility of human cytochrome P450 enzymes: molecular dynamics and spectroscopy reveal important function-related variations. *Biochim Biophys Acta.* 2011;1814:58-68.

26. https://www.bccresearch.com/market-research/biotechnology/global-markets-for-enzymes-in-industrial-applications.html.

27. Garzón-Posse F, Becerra-Figueroa L, Hernández-Arias J, Gamba-Sánchez D. Whole cells as biocatalysts in organic transformations. *Molecules.* 2018;23:1265.

28. Sheldon RA, Arends I, Hanefeld U. *Green chemistry and catalysis*: John Wiley & Sons; 2007.

29. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K. Engineering the third wave of biocatalysis. *Nature.* 2012;485:185-194.

30. Jaeger K-E, Eggert T. Enantioselective biocatalysis optimized by directed evolution. *Curr Opin Biotechnol.* 2004;15:305-313.

31. Lin B, Tao Y. Whole-cell biocatalysts by design. *Microb Cell Fact.* 2017;16:106.

32. Rigoldi F, Donini S, Redaelli A, Parisini E, Gautieri A. Engineering of thermostable enzymes for industrial applications. *APL Bioeng.* 2018;2:011501.

33. Littlechild JA. Enzymes from extreme environments and their industrial applications. *Front Bioeng Biotechnol.* 2015;3:161.

34. Martínez-Martínez M, Coscolín C, Santiago G, *et al.* Determinants and prediction of esterase substrate promiscuity patterns. *ACS Chem Biol.* 2018;13:225-234.

35. Skoczinski P, Volkenborn K, Fulton A, *et al.* Contribution of single amino acid and codon substitutions to the production and secretion of a lipase by *Bacillus subtilis*. *Microb Cell Fact.* 2017;16:160.

36. Chaparro-Riggers JF, Polizzi KM, Bommarius AS. Better library design: data-driven protein engineering. *Biotechnol J.* 2007;2:180-191.

37. Reetz MT, Carballeira JD, Vogel A. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew Chem Int Ed.* 2006;45:7745-7751.

38. Huang X, Gao D, Zhan CG. Computational design of a thermostable mutant of cocaine esterase via molecular dynamics simulations. *Org Biomol Chem.* 2011;9:4138-4143.

39. Singh J, Ator MA, Jaeger EP, *et al.* Application of genetic algorithms to combinatorial synthesis: A computational approach to lead identification and lead optimization. *J Am Chem Soc.* 1996;118:1669-1676.

40. Thiltgen G, Goldstein RA. Assessing predictors of changes in protein stability upon mutation using self-consistency. *PloS One.* 2012;7:e46084.

41. Modarres HP, Mofrad M, Sanati-Nezhad A. Protein thermostability engineering. *RSC Adv.* 2016;6:115252-115270.

42. Zhang Z, Wang L, Gao Y, Zhang J, Zhenirovskyy M, Alexov E. Predicting folding free energy changes upon single point mutations. *Bioinformatics.* 2012;28:664-671.

43. Kang S, Chen G, Xiao G. Robust prediction of mutation-induced protein stability change by property encoding of amino acids. *Protein Eng Des Sel.* 2009;22:75-83.

44. Nutschel C, Fulton A, Zimmermann O, Schwaneberg U, Jaeger K-E, Gohlke H. Systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance for *Bacillus subtilis* lipase A. *J Chem Inf Model.* 2020;60:1568-1584.

45. Nutschel C, Coscolín C, Mulnaes D, David B, Ferrer M, Jaeger K-E, Gohlke H. Promiscuous esterases counterintuitively are less flexible than specific ones. *J Chem Inf Model*. 2020; DOI: 10.1021/acs.jcim.1c00152.

46. Pfleger C, Rathi PC, Klein DL, Radestock S, Gohlke H. Constraint Network Analysis (CNA): a Python software package for efficiently linking biomacromolecular structure, flexibility,(thermo-) stability, and function. *J Chem Inf Model.* 2013;53:1007-1015.

47. Steiner K, Schwab H. Recent advances in rational approaches for enzyme engineering. *Comput Struct Biotechnol J.* 2012;2:e201209010.

48. Rathi PC, Fulton A, Jaeger K-E, Gohlke H. Application of rigidity theory to the thermostabilization of lipase A from *Bacillus subtilis*. *PLoS Comput Biol.* 2016;12:e1004754.

49. Zeymer C, Hilvert D. Directed evolution of protein catalysts. *Annu Rev Biochem.* 2018;87:131-157.

50. Arnold FH. Directed evolution: bringing new chemistry to life. *Angew Chem Int Ed Engl.* 2018;57:4143-4148.

51. Turner NJ. Directed evolution of enzymes for applied biocatalysis. *Trends Biotechnol.* 2003;21:474-478.

52. Harayama S. Artificial evolution by DNA shuffling. *Trends Biotechnol.* 1998;16:76-82.

53. Esvelt KM, Carlson JC, Liu DR. A system for the continuous directed evolution of biomolecules. *Nature.* 2011;472:499-503.

54. Xiao H, Bao Z, Zhao H. High throughput screening and selection methods for directed enzyme evolution. *Ind Eng Chem Res.* 2015;54:4011-4020.

55. Santoro SW, Schultz PG. Directed evolution of the site specificity of Cre recombinase. *Proc Natl Acad Sci U S A.* 2002;99:4185-4190.

56. Boersma YL, Droge MJ, Quax WJ. Selection strategies for improved biocatalysts. *FEBS J.* 2007;274:2181-2195.

57. Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol.* 2009;10:866-876.

58. Giver L, Gershenson A, Freskgard PO, Arnold FH. Directed evolution of a thermostable esterase. *Proc Natl Acad Sci U S A.* 1998;95:12809-12813.

59. Kuchner O, Arnold FH. Directed evolution of enzyme catalysts. *Trends Biotechnol.* 1997;15:523-530.

60. Korendovych IV. Rational and semirational protein design. *Methods Mol Biol.* 2018;1685:15-23.

61. Damborsky J, Brezovsky J. Computational tools for designing and engineering enzymes. *Curr Opin Chem Biol.* 2014;19:8-16.

62. Bornscheuer UT, Pohl M. Improved biocatalysts by directed evolution and rational protein design. *Curr Opin Chem Biol.* 2001;5:137-143.

63. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel.* 2009;22:553-560.

64. Khan S, Vihinen M. Performance of protein stability predictors. *Hum Mutat.* 2010;31:675-684.

65. Usmanova DR, Bogatyreva NS, Ariño Bernad J, *et al.* Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics.* 2018;34:3653-3658.

66. Pucci F, Bernaerts KV, Kwasigroch JM, Rooman M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics.* 2018;34:3659-3665.

67. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* 2004;32:D120-D121.

68. Kumar MD, Bava KA, Gromiha MM, *et al.* ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic Acids Res.* 2006;34:D204-D206.

69. Chica RA, Doucet N, Pelletier JN. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr Opin Biotechnol.* 2005;16:378-384.

70. Reetz MT, Kahakeaw D, Lohmer R. Addressing the numbers problem in directed evolution. *Chembiochem.* 2008;9:1797-1804.

71. Moore KW, Pechen A, Feng XJ, Dominy J, Beltrani VJ, Rabitz H. Why is chemical synthesis and property optimization easier than expected? *Phys Chem Chem Phys* 2011;13:10048-10070.

72. Hermans SMA, Pfleger C, Nutschel C, Hanke CA, Gohlke H. Rigidity theory for biomolecules: concepts, software, and applications. *WIREs Comput Mol Sci.* 2017;7:e1311.

73. Fulle S, Gohlke H. Statics of the ribosomal exit tunnel: implications for cotranslational peptide folding, elongation regulation, and antibiotics binding. *J Mol Biol.* 2009;387:502-517.

74. Mottonen JM, Jacobs DJ, Livesay DR. Allosteric response is both conserved and variable across three CheY orthologs. *Biophys J.* 2010;99:2245-2254.

75. Pfleger C, Minges A, Boehm M, McClendon CL, Torella R, Gohlke H. Ensemble- and Rigidity Theory-Based Perturbation Approach To Analyze Dynamic Allostery. *J Chem Theory Comput.* 2017;13:6343-6357.

76. Jacobs DJ, Dallakyan S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys J.* 2005;88:903-915.

77. Del Carpio CA, Iulian Florea M, Suzuki A, *et al.* A graph theoretical approach for assessing bio-macromolecular complex structural stability. *J Mol Model.* 2009;15:1349-1370.

78. Gohlke H, Kuhn LA, Case DA. Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins.* 2004;56:322-337.

79. Hespenheide BM, Rader AJ, Thorpe MF, Kuhn LA. Identifying protein folding cores from the evolution of flexible regions during unfolding. *J Mol Graph Model.* 2002;21:195-207.

80. Rader AJ, Bahar I. Folding core predictions from network models of proteins. *Polymer.* 2004;45:659-668.

81. Ahmed A, Gohlke H. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins.* 2006;63:1038-1051.

82. Fulle S, Christ NA, Kestner E, Gohlke H. HIV-1 TAR RNA spontaneously undergoes relevant apo-to-holo conformational transitions in molecular dynamics and constrained geometrical simulations. *J Chem Inf Model.* 2010;50:1489-1501.

83. Wells S, Menor S, Hespenheide B, Thorpe MF. Constrained geometric simulation of diffusive motion in proteins. *Phys Biol.* 2005;2:S127-136.

84. Farrell DW, Speranskiy K, Thorpe MF. Generating stereochemically acceptable protein pathways. *Proteins.* 2010;78:2908-2921.

85. Tan H, Rader AJ. Identification of putative, stable binding regions through flexibility analysis of HIV-1 gp120. *Proteins.* 2009;74:881-894.

86. Livesay DR, Jacobs DJ. Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins.* 2006;62:130-143.

87. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. *Proteins.* 2001;44:150-165.

88. Fox N, Jagodzinski F, Li Y, Streinu I. KINARI-Web: a server for protein rigidity analysis. *Nucleic Acids Res.* 2011;39:W177-W183.

89. Maxwell JC. On the calculation of the equilibrium and stiffness of frames. *Philos Mag.* 1864;27:294-299.

90. Laman G. On graphs and rigidity of plane skeletal structures. *J Eng Math.* 1970;4:331-340.

91. Hendrickson B. Conditions for unique graph realizations. *SIAM J Comput.* 1992;21:65-84.

92. Jacobs DJ. Generic rigidity in three-dimensional bond-bending networks. *J Phys A Math Gen.* 1998;31:6653-6668.

93. Thorpe MF, Jacobs DJ, Chubynsky NV, Rader AJ. Generic rigidity of network glasses. *In Rigidity Theory and Applications*: Kluwer Academic/Plenum Publishers, New York; 1999; 239-277.

94. Whiteley W. Rigidity of molecular structures: generic and geometric analysis. *Rigidity theory and Applications*. New York: Kluwer Academic/Plenum Publishers; 2002:21-46.

95. Tay T-S, Whiteley W. Recent advances in the generic ridigity of structures. *Struct Topol.* 1984;9:31-38.

96. Hespenheide BM, Jacobs DJ, Thorpe MF. Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J Phys Condens Matter.* 2004;16:S5055-S5064.

97. Whiteley W. Counting out to the flexibility of molecules. *Phys Biol.* 2005;2:S116-S126.

98. Fox N, Streinu I. Towards accurate modeling of noncovalent interactions for protein rigidity analysis. *BMC Bioinformatics.* 2013;14:S3.

99. Wells SA, Jimenez-Roldan JE, Römer RA. Comparative analysis of rigidity across protein families. *Phys Biol.* 2009;6:046005.

100. Dahiyat BI, Gordon DB, Mayo SL. Automated design of the surface positions of protein helices. *Protein Sci.* 1997;6:1333-1337.

101. Thorpe MF, Lei M, Rader AJ, Jacobs DJ, Kuhn LA. Protein flexibility and dynamics using constraint theory. *J Mol Graph Model.* 2001;19:60-69.

102. Cheatham 3rd TEI, Cieplak P, Kollman PA. A modified version of the Cornell *et al.* force field with improved sugar pucker phases and helical repeat. *J Biomol Struct Dyn.* 1999;16:845-862.

103. Cornell WD, Cieplak P, Bayly CI, *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc.* 1995;117:5179-5197.

104. Mamonova T, Hespenheide B, Straub R, Thorpe MF, Kurnikova M. Protein flexibility using constraints from molecular dynamics simulations. *Phys Biol.* 2005;2:S137-S147.

105. Pfleger C, Radestock S, Schmidt E, Gohlke H. Global and local indices for characterizing biomolecular flexibility and rigidity. *J Comput Chem.* 2013;34:220-233.

106. Jacobs DJ, Thorpe MF. Generic rigidity percolation: the pebble game. *Phys Rev Lett.* 1995;75:4051-4054.

107. Jacobs DJ, Thorpe MF. *Computer-implemented system for analyzing rigidity of substructures within a macromolecule*: Google Patents; 2000.

108. Lee A, Streinu I. Pebble game algorithms and sparse graphs. *Discrete Math.* 2008;308:1425-1437.

109. Lee A, Streinu I, Theran L. Graded sparse graphs and matroids. *J Univ Comput Sci.* 2007;13:1671-1679.

110. Jacobs DJ, Hendrickson B. An algorithm for two-dimensional rigidity percolation: the pebble game. *J Comput Phys.* 1997;137:346-365.

111. Katoh N, Tanigawa S-i. A proof of the molecular conjecture. *Discrete Comput Geom.* 2011;45:647-700.

112. Rathi PC, Radestock S, Gohlke H. Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J Biotechnol.* 2012;159:135-144.

113. Makhatadze GI, Privalov PL. On the entropy of protein folding. *Protein Sci.* 1996;5:507-510.

114. Pfleger C, Gohlke H. Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. *Structure.* 2013;21:1725-1734.

115. Gohlke H, Case DA. Converging free energy estimates: MM-PB (GB) SA studies on the protein–protein complex Ras–Raf. *J Comput Chem.* 2004;25:238-250.

116. Sljoka A, Wilson D. Probing protein ensemble rigidity and hydrogen–deuterium exchange. *Phys Biol.* 2013;10:056013.

117. Rathi PC, Mulnaes D, Gohlke H. VisualCNA: a GUI for interactive constraint network analysis and protein engineering for improving thermostability. *Bioinformatics.* 2015;31:2394-2396.

118. Krüger DM, Rathi PC, Pfleger C, Gohlke H. CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure,(thermo-) stability, and function. *Nucleic Acids Res.* 2013;41:W340-W348.

119. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF. Protein unfolding: rigidity lost. *Proc Natl Acad Sci U S A.* 2002;99:3540-3545.

120. Rader AJ. Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys Biol.* 2009;7:16002.

121. Privalov PL, Gill SJ. Stability of protein structure and hydrophobic interaction. *Adv Protein Chem.* 1988;39:191-234.

122. Schellman JA. Temperature, stability, and the hydrophobic interaction. *Biophys J.* 1997;73:2960-2964.

123. Stauffer D. Scaling theory of percolation clusters. *Phy Reps.* 1979;54:1-74.

124. Stauffer D, Aharony A. *Introduction to percolation theory.* 2nd ed: Taylor and Francis, London; 1994.

125. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J.* 1948;27:623-656.

126. Andraud C, Beghdadi A, Lafait J. Entropic analysis of random morphologies. *Physica A.* 1994;207:208-212.

127. Rathi PC, Jaeger K-E, Gohlke H. Structural rigidity and protein thermostability in variants of lipase A from *Bacillus subtilis*. *PloS One.* 2015;10:1-24.

128. Polizzi KM, Bommarius AS, Broering JM, Chaparro-Riggers JF. Stability of biocatalysts. *Curr Opin Chem Biol.* 2007;11:220-225.

129. O'Fagain C. Engineering protein stability. *Methods Mol Biol.* 2011;681:103-136.

130. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. *Structure.* 2003;11:1453-1459.

131. Dick M, Weiergräber OH, Classen T, *et al.* Trading off stability against activity in extremophilic aldolases. *Sci Rep.* 2016;6:17908.

132. Kovacic F, Babic N, Krauss U, Jaeger K. Classification of lipolytic enzymes from bacteria. *Aerobic utilization of hydrocarbons, oils and lipids. In Handbook of Hydrocarbon and Lipid Microbiology*: Springer, Berlin, Heidelberg; 2018; 1-35.

133. Ali YB, Verger R, Abousalham A. Lipases or esterases: does it really matter? Toward a new bio-physico-chemical classification. *Methods Mol Biol.* 2012;861:31-51.

134. Verger R. 'Interfacial activation'of lipases: facts and artifacts. *Trends Biotechnol.* 1997;15:32-38.

135. Bornscheuer UT. Microbial carboxyl esterases: classification, properties and application in biocatalysis. *FEMS Microbiol. Rev.* 2002;26:73-81.

136. Jaeger K-E, Ransac S, Dijkstra BW, Colson C, van Heuvel M, Misset O. Bacterial lipases. *FEMS Microbiol Rev.* 1994;15:29-63.

137. Jaeger K-E, Reetz MT. Microbial lipases form versatile tools for biotechnology. *Trends Biotechnol.* 1998;16:396-403.

138. Pandey A, Benjamin S, Soccol CR, Nigam P, Krieger N, Soccol VT. The realm of microbial lipases in biotechnology. *Biotechnol Appl Biochem.* 1999;29:119-131.

139. Sarda L, Desnuelle P. Actions of pancreatic lipase on esters in emulsions. *Biochim Biophys Acta.* 1958;30:513-521.

140. Brockerhoff H. *Lipolytic enzymes*: ACS Publications; 1974.

141. Derewenda ZS, Derewenda U, Dodson GG. The crystal and molecular structure of the Rhizomucor miehei triacylglyceride lipase at 1.9 Å resolution. *J Mol Biol.* 1992;227:818-839.

142. Winkler FK, D'Arcy A, Hunziker W. Structure of human pancreatic lipase. *Nature.* 1990;343:771-774.

143. van Tilbeurgh H, Egloff MP, Martinez C, Rugani N, Verger R, Cambillau C. Interfacial activation of the lipase–procolipase complex by mixed micelles revealed by X-ray crystallography. *Nature.* 1993;362:814-820.

144. Brzozowski AM, Derewenda U, Derewenda ZS, *et al.* A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. *Nature.* 1991;351:491-494.

145. Noble ME, Cleasby A, Johnson LN, Egmond MR, Frenken LG. The crystal structure of triacylglycerol lipase from *Pseudomonas glumae* reveals a partially redundant catalytic aspartate. *FEBS Lett.* 1993;331:123-128.

146. Uppenberg J, Hansen MT, Patkar S, Jones TA. The sequence, crystal structure determination and refinement of two crystal forms of lipase B from *Candida antarctica*. *Structure.* 1994;2:293-308.

147. Jaeger K-E, Ransac S, Koch HB, Ferrato F, Dijkstra BW. Topological characterization and modeling of the 3D structure of lipase from *Pseudomonas aeruginosa*. *FEBS Lett.* 1993;332:143-149.

148. Lesuisse E, Schanck K, Colson C. Purification and preliminary characterization of the extracellular lipase of *Bacillus subtilis* 168, an extremely basic pH-tolerant enzyme. *Eur J Biochem.* 1993;216:155-160.

149. van Pouderoyen G, Eggert T, Jaeger K-E, Dijkstra BW. The crystal structure of *Bacillus subtilis* lipase: a minimal α/β hydrolase fold enzyme. *J Mol Biol.* 2001;309:215-226.

150. Arpigny JL, Jaeger K-E. Bacterial lipolytic enzymes: classification and properties. *Biochem J.* 1999;343:177-183.

151. Hausmann S, Jaeger K-E. Lipolytic enzymes from bacteria. *In Handbook of hydrocarbon and lipid microbiology.:* Springer, Berlin, Heidelberg; 2010; 1099-1126.

152. López-López O, Cerdán ME, González Siso MI. New extremophilic lipases and esterases from metagenomics. *Curr Protein Pept Sci.* 2014;15:445-455.

153. Nthangeni MB, Patterton H, van Tonder A, Vergeer WP, Litthauer D. Over-expression and properties of a purified recombinant *Bacillus* licheniformis lipase: a comparative report on *Bacillus* lipases. *Enzyme Microb Technol.* 2001;28:705-712.

154. Hemilä H, Koivula TT, Palva I. Hormone-sensitive lipase is closely related to several bacterial proteins, and distantly related to acetylcholinesterase and lipoprotein lipase: identification of a superfamily of esterases and lipases. *Biochim Biophys Acta.* 1994;1210:249-253.

155. Pharkya P, Nikolaev EV, Maranas CD. Review of the BRENDA Database. *Metab Eng.* 2003;5:71-73.

156. Schomburg I, Chang A, Ebeling C, *et al.* BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* 2004;32:D431-433.

157. Ollis DL, Cheah E, Cygler M, *et al.* The α/β hydrolase fold. *Protein Eng.* 1992;5:197-211.

158. Wagner UG, Petersen EI, Schwab H, Kratky C. EstB from Burkholderia gladioli: a novel esterase with a β-lactamase fold reveals steric factors to discriminate between esterolytic and β-lactam cleaving activity. *Protein Sci.* 2002;11:467-478.

159. McKay DB, Jennings MP, Godfrey EA, MacRae IC, Rogers PJ, Beacham IR. Molecular analysis of an esterase-encoding gene from a lipolytic psychrotrophic pseudomonad. *J Gen Microbiol.* 1992;138:701-708.

160. Heikinheimo P, Goldman A, Jeffries C, Ollis DL. Of barn owls and bankers: a lush variety of α/β hydrolases. *Structure.* 1999;7:R141-R146.

161. Chow J, Kovacic F, Dall Antonia Y, *et al.* The metagenome-derived enzymes LipS and LipT increase the diversity of known lipases. *PloS One.* 2012;7:e47665.

162. Skjøt M, De Maria L, Chatterjee R, *et al.* Understanding the plasticity of the alpha/beta hydrolase fold: lid swapping on the *Candida antarctica* lipase B results in chimeras with interesting biocatalytic properties. *Chembiochem.* 2009;10:520-527.

163. Brady L, Brzozowski AM, Derewenda ZS, *et al.* A serine protease triad forms the catalytic centre of a triacylglycerol lipase. *Nature.* 1990;343:767-770.

164. Brenner S. The molecular evolution of genes and proteins: a tale of two serines. *Nature.* 1988;334:528-530.

165. Schrag JD, Li YG, Wu S, Cygler M. Ser-His-Glu triad forms the catalytic site of the lipase from Geotrichum candidum. *Nature.* 1991;351:761-764.

166. Jaeger K-E, Dijkstra BW, Reetz MT. Bacterial biocatalysts: molecular biology, three-dimensional structures, and biotechnological applications of lipases. *Annu Rev Microbiol.* 1999;53:315-351.

167. Bajpai P. Application of enzymes in the pulp and paper industry. *Biotechnol Prog.* 1999;15:147-157.

168. Jaeger K-E, Eggert T. Lipases for biotechnology. *Curr Opin Biotechnol.* 2002;13:390-397.

169. Raveendran S, Parameswaran B, Ummalyma SB, *et al.* Applications of microbial enzymes in food industry. *Food Technol. Biotechnol.* 2018;56:16-30.

170. Ferrer M, Bargiela R, Martínez-Martínez M, *et al.* Biodiversity for biocatalysis: a review of the α/β-hydrolase fold superfamily of esterases-lipases discovered in metagenomes. *Biocatal Biotransfor.* 2015;33:235-249.

171. Funke SA, Eipper A, Reetz MT, *et al.* Directed evolution of an enantioselective *Bacillus subtilis* lipase. *Biocatal Biotransfor.* 2003;21:67-73.

172. Gupta R, Gupta N, Rathi P. Bacterial lipases: an overview of production, purification and biochemical properties. *Appl Microbiol Biotechnol.* 2004;64:763-781.

173. Fulton A, Frauenkron-Machedjou VJ, Skoczinski P, *et al.* Exploring the protein stability landscape: *Bacillus subtilis* lipase A as a model for detergent tolerance. *Chembiochem.* 2015;16:930-936.

174. Frauenkron-Machedjou VJ, Fulton A, Zhu L, *et al.* Towards understanding directed evolution: more than half of all amino acid positions contribute to ionic liquid resistance of *Bacillus subtilis* lipase A. *Chembiochem.* 2015;16:937-945.

175. Ortiz C, Ferreira ML, Barbosa O, *et al.* Novozym 435: the “perfect” lipase immobilized biocatalyst? *Catalysis Science & Technology.* 2019;9:2380-2420.

176. Schallmey M, Singh A, Ward OP. Developments in the use of *Bacillus* species for industrial production. *Can J Microbiol.* 2004;50:1-17.

177. van Dijl JM, Hecker M. *Bacillus subtilis*: from soil bacterium to super-secreting cell factory. *Microb Cell Fact.* 2013;12:3.

178. Nijland R, Kuipers OP. Optimization of protein secretion by *Bacillus subtilis*. *Recent Pat Biotechnol.* 2008;2:79-87.

179. Brockmeier U, Caspers M, Freudl R, Jockwer A, Noll T, Eggert T. Systematic screening of all signal peptides from *Bacillus subtilis*: a powerful strategy in optimizing heterologous protein secretion in Gram-positive bacteria. *J Molecular Biol.* 2006;362:393-402.

180. Kunst F, Ogasawara N, Moszer I, *et al.* The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature.* 1997;390:249-256.

181. Kobayashi K, Ehrlich SD, Albertini A, *et al.* Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A.* 2003;100:4678-4683.

182. Buescher JM, Liebermeister W, Jules M, *et al.* Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism. *Science.* 2012;335:1099-1103.

183. Otto A, Bernhardt J, Meyer H, *et al.* Systems-wide temporal proteomic profiling in glucose-starved *Bacillus subtilis*. *Nat Commun.* 2010;1:1-9.

184. Becher D, Büttner K, Moche M, Heßling B, Hecker M. From the genome sequence to the protein inventory of *Bacillus subtilis*. *Proteomics.* 2011;11:2971-2980.

185. Westers L, Westers H, Quax WJ. *Bacillus subtilis* as cell factory for pharmaceutical proteins: a biotechnological approach to optimize the host organism. *Biochim Biophys Acta.* 2004;1694:299-310.

186. Hidaka H, Eida T, Takizawa T, Tokunaga T, Tashiro Y. Effects of fructooligosaccharides on intestinal flora and human health. *Bifidobact Microflora.* 1986;5:37-50.

187. Mitsuoka T. Intestinal flora and human health. *Asia Pacific J Clin Nutr.* 1996;5:2-9.

188. Cooperstock MS, Zedd AJ. Intestinal flora of infants. *Human Intestinal Microflora in Health and Disease*: Academic Press, New York; 1983;79-99..

189. Simonen M, Palva I. Protein secretion in *Bacillus* species. *Microbiol Rev.* 1993;57:109-137.

190. Carrió MM, Cubarsi R, Villaverde A. Fine architecture of bacterial inclusion bodies. *FEBS Lett.* 2000;471:7-11.

191. Tjalsma H, Antelmann H, Jongbloed JD, *et al.* Proteomics of protein secretion by *Bacillus subtilis*: separating the "secrets" of the secretome. *Microbiol Mol Biol Rev.* 2004;68:207-233.

192. Palmer T, Berks BC. The twin-arginine translocation (Tat) protein export pathway. *Nat Rev Microbiol.* 2012;10:483-496.

193. Ling Lin F, Zi Rong X, Wei Fen L, Jiang Bing S, Ping L, Chun Xia H. Protein secretion pathways in *Bacillus subtilis*: implication for optimization of heterologous protein secretion. *Biotechnol Adv.* 2007;25:1-12.

194. Quentin Y, Fichant G, Denizot F. Inventory, assembly and analysis of *Bacillus subtilis* ABC transport systems. *J Mol Biol.* 1999;287:467-484.

195. Nardini M, Lang DA, Liebeton K, Jaeger K-E, Dijkstra BW. Crystal structure of pseudomonas aeruginosa lipase in the open conformation. The prototype for family I. 1 of bacterial lipases. *J Biol Chem.* 2000;275:31219-31225.

196. Salazar O, Cirino PC, Arnold FH. Thermostabilization of a cytochrome P450 peroxygenase. *Chembiochem.* 2003;4:891-893.

197. Pottkämper J, Barthen P, Ilmberger N, *et al.* Applying metagenomics for the identification of bacterial cellulases that are stable in ionic liquids. *Green Chem.* 2009;11:957-965.

198. Liu H, Zhu L, Bocola M, Chen N, Spiess AC, Schwaneberg U. Directed laccase evolution for improved ionic liquid resistance. *Green Chem.* 2013;15:1348-1355.

199. Carter JL, Bekhouche M, Noiriel A, Blum LJ, Doumèche B. Directed evolution of a formate dehydrogenase for increased tolerance to ionic liquids reveals a new site for increasing the stability. *Chembiochem.* 2014;15:2710-2718.

200. Chen Z, Pereira JH, Liu H, *et al.* Improved activity of a thermophilic cellulase, Cel5A, from Thermotoga maritima on ionic liquid pretreated switchgrass. *PloS One.* 2013;8:e79725.

201. Lehmann C, Bocola M, Streit WR, Martinez R, Schwaneberg U. Ionic liquid and deep eutectic solvent-activated CelA2 variants generated by directed evolution. *Appl Microbiol and Biotechnol.* 2014;98:5775-5785.

202. Nordwald EM, Armstrong GS, Kaar JL. NMR-guided rational engineering of an ionic-liquid-tolerant lipase. *ACS Catal.* 2014;4:4057-4064.

203. Zhao J, Frauenkron-Machedjou VJ, Fulton A, *et al.* Unraveling the effects of amino acid substitutions enhancing lipase resistance to an ionic liquid: a molecular dynamics study. *Phys Chem Chem Phys.* 2018;20:9600-9609.

204. Brissos V, Eggert T, Cabral JM, Jaeger K-E. Improving activity and stability of cutinase towards the anionic detergent AOT by complete saturation mutagenesis. *Protein Eng Des Sel.* 2008;21:387-393.

205. Akbulut N, Öztürk MT, Pijning T, Öztürk Sİ, Gümüşel F. Improved activity and thermostability of *Bacillus* pumilus lipase by directed evolution. *J Biotechnol.* 2013;164:123-129.

206. Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc Natl Acad Sci U S A.* 2012;109:16858-16863.

207. Foit L, Morgan GJ, Kern MJ, *et al.* Optimizing protein stability in vivo. *Mol Cell.* 2009;36:861-871.

208. Deng Z, Huang W, Bakkalbasi E, *et al.* Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *J Mol Biol.* 2012;424:150-167.

209. Klesmith JR, Bacik J-P, Wrenbeck EE, Michalczyk R, Whitehead TA. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc Natl Acad Sci U S A.* 2017;114:2265-2270.

210. Vogt G, Woell S, Argos P. Protein thermal stability, hydrogen bonds, and ion pairs. *J Mol Biol.* 1997;269:631-643.

211. Chakravarty S, Varadarajan R. Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Lett.* 2000;470:65-69.

212. Pack SP, Kang TJ, Yoo YJ. Protein thermostabilizing factors: high relative occurrence of amino acids, residual properties, and secondary structure type in different residual state. *Appl Biochem Biotechnol.* 2013;171:1212-1226.

213. Irback A, Mohanty S. PROFASI: A Monte Carlo simulation package for protein folding and aggregation. *J Comput Chem.* 2006;27:1548-1555.

214. Mohanty S, Meinke JH, Zimmermann O. Folding of Top7 in unbiased all-atom Monte Carlo simulations. *Proteins.* 2013;81:1446-1456.

215. Chen R, Gao B, Liu X, *et al.* Molecular insights into the enzyme promiscuity of an aromatic prenyltransferase. *Nat Chem Biol.* 2017;13:226-234.

216. Huang H, Pandya C, Liu C, *et al.* Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc Natl Acad Sci U S A.* 2015;112:E1974-E1983.

217. Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem.* 2010;79:471-505.

218. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. *Nat Biotechnol.* 2009;27:157-167.

219. Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. *Mol Biol Evol.* 2015;32:132-143.

220. Mulnaes D, Porta N, Clemens R, *et al.* TopModel: Template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks. *J Chem Theory Comput.* 2020;16:1953-1967.

221. Mulnaes D, Gohlke H. TopScore: Using Deep Neural Networks and Large Diverse Data Sets for Accurate Protein Model Quality Assessment. *J Chem Theory Comput.* 2018;14:6117-6126.

222. Copley SD. Shining a light on enzyme promiscuity. *Curr Opin Struct Biol.* 2017;47:167-175.

223. Stockwell GR, Thornton JM. Conformational diversity of ligands bound to proteins. *J Mol Biol.* 2006;356:928-944.

224. Perola E, Charifson PS. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem.* 2004;47:2499-2510.

225. Fedyunin I, Lehnhardt L, Böhmer N, Kaufmann P, Zhang G, Ignatova Z. tRNA concentration fine tunes protein solubility. *FEBS Lett.* 2012;586:3336-3340.

226. Kamal MZ, Ahmad S, Yedavalli P, Rao NM. Stability curves of laboratory evolved thermostable mutants of a *Bacillus subtilis* lipase. *Biochim Biophys Acta.* 2010;1804:1850-1856.

227. Wijma HJ, Floor RJ, Janssen DB. Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr Opin Struct Biol.* 2013;23:588-594.

228. Kamal MZ, Ahmad S, Molugu TR, *et al.* In vitro evolved non-aggregating and thermostable lipase: structural and thermodynamic investigation. *J Mol Biol.* 2011;413:726-741.