Can Constraint Network Analysis guide the identification phase of KnowVolution? A case study on improved thermostability of an endo-β-glucanase

Francisca Contreras[a#], Christina Nutschel[b#], Laura Beust[a], Mehdi D. Davari[a], Holger Gohlke[b,c*], Ulrich Schwaneberg[a,d*]

[a] Institute of Biotechnology, RWTH Aachen University, Worringerweg 3, 52074 Aachen, Germany. f.contreras@biotec.rwth-aachen.de (F.C), laura.beust@rwth-aachen.de (L.B.), m.davari@biotec.rwth-aachen.de (M.D.D.)

[b] 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. Christina.Nutschel@uni-duesseldorf.de (C.N.), h.gohlke@fz-juelich.de (H.G.)

[c] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. gohlke@uni-duesseldorf.de (H.G.)

[d] DWI-Leibniz Institute for Interactive Materials, Forckenbeckstraße 50, 52074 Aachen, Germany. u.schwanberg@biotec.rwth-aachen.de (U.S.)

[#]Shared first authors

*Corresponding authors:

Holger Gohlke: +49 211 81 13662, h.gohlke@fz-juelich.de or gohlke@uni-duesseldorf.de

Ulrich Schwaneberg: +49 241 80 24170, e-mail: u.schwanberg@biotec.rwth-aachen.de

Running title: CNA-guided engineering of EGLII thermostability

Abstract

Cellulases are industrially important enzymes, e.g., in the production of bioethanol, in pulp and paper industry, feedstock, and textile. Thermostability is often a prerequisite for high process stability and improving thermostability without affecting specific activities at lower temperatures is challenging and often time-consuming. Protein engineering strategies that combine experimental and computational are emerging in order to reduce experimental screening efforts and speed up enzyme engineering campaigns. Constraint Network Analysis (CNA) is a promising computational method that identifies beneficial positions in enzymes to improve thermostability. In this study, we compare CNA and directed evolution in the identification of beneficial positions in order to evaluate the potential of CNA in protein engineering campaigns (e.g., in the identification phase of KnowVolution). We engineered the industrially relevant endoglucanase EGLII from *Penicillium verruculosum* towards increased thermostability. From the CNA approach, six variants were obtained with an up to 2-fold improvement in thermostability. The overall experimental burden was reduced to 40% utilizing the CNA method in comparison to directed evolution. On a variant level, the success rate was similar for both strategies, with 0.27% and 0.18% improved variants in the epPCR and CNA-guided library, respectively. In essence, CNA is an effective method for identification of positions that improve thermostability.

Keywords

KnowVolution, protein engineering, Constraint Network Analysis, thermostability, cellulase, GH5 endoglucanase.

1. Introduction

Thermal stability is often a prerequisite for process stability of enzymes in industrial chemical production [1]. Cellulases are employed in processes as biomass depolymerization and animal feed, in which high temperature is needed. In biomass depolymerization, the enzyme dosage can be decreased if a biocatalyst that can withstand high temperatures is employed, and, therefore, a diminution of costs can be achieved [2]. Cellulases used for animal feed pellets are required to withstand high temperatures for a short period of time, e.g., a few minutes; in that period, the cellulases should not be inactivated because their catalytic activity is needed in a posterior stage [3].

The current problem is that high activity and high thermal resistance are often contradictory properties, and thermostable enzymes that can be found in nature, e.g., hot-springs and hydrothermal vents, have evolved to be active at high temperatures (>90 °C). For this reason, their activity at low temperatures (<40 °C) is, in most cases, nonexistent or very low [4, 5]. In order to keep enzymes as a competitive option for catalysis, they should possess high thermal stability and maintain a high activity in a broad spectrum of temperatures.

In the last decades, protein engineering has emerged as an important tool to improve the thermostability of enzymes, and directed evolution has become a widely utilized technique in protein engineering [6]. KnowVolution is a strategy that combines directed evolution and computational analyses; this combination allows us to decrease the experimental efforts and maximize improvements [7]. The KnowVolution campaign encompasses four phases: Phase I, identification of beneficial positions is done by random mutagenesis; Phase II, full diversity is generated on the identified beneficial positions, and the best substitution is determined; Phase III, the interaction between

substitutions is determined by computationally assisted structural analyses; and Phase IV, recombination of selected positions. To further decrease screening efforts, Phase IV can be guided by the CompassR rule [8] that discards recombinations that are potentially unstable. At the end of a KnowVolution campaign, a molecular understanding of each substitutions' role in improvement is generated.

A common strategy employed to decrease the screening burden is to combine screening systems like two-step screening systems [9]. The first step consists of a high-throughput screening (HTS) where viable clones are selected, which can be carried, e.g., in agar plates. Later, in a second step, an enriched library can be further screened to determine improved clones expressing the desired property; this enrichment can be performed, e.g., using a liquid assay in a multi-well plate [10]. Another option to decrease the screening burden is to complement the identification phase of KnowVolution (Phase I) with computational techniques, which can reduce the time and effort spent in the laboratory. Several *in silico* methods are available to improve the thermostability of enzymes, including B-FITTER [11], PoPMuSiC [12], FoldX [13-15], FireProt [16], FRESCO [17, 18], PROSS [19], iSAR [20], and Constraint Network Analysis (CNA) [21-24].

CNA functions as front- and back-end to the graph theory-based software Floppy Inclusions and Rigid Substructure Topography (FIRST) [25]. Applying CNA to biomolecules aims at identifying their composition of rigid clusters and flexible regions, which can aid in the understanding of the biomolecular structure, stability, and function [23, 24]. In CNA, biomolecules are modeled as constraint networks, in which covalent and non-covalent interactions compose the edges, and atoms represent the nodes, as described in detail by Hesphenheide *et al* [26]. 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 [27]. In order to

monitor the hierarchy of rigidity and flexibility of biomolecules, CNA performs thermal unfolding simulations by consecutively removing non-covalent constraints from a network in increasing order of their strength [28-30]. CNA has been applied before retro- [21, 30-33] and prospectively [34, 35] in the context of improving protein thermostability, but never compared directly to the performance of random mutagenesis.

Endoglucanases are enzymes commonly used in an ample variety of industries as food and feed, biofuels, detergents, as well as pulp and paper [36]. Nowadays, cellulases from *Penicillium* strains appear as promising biocatalysts for cellulose degradation due to their high activity. Endoglucanase II (PvCel5A) is an endo-β-1,4-glucanase from the fungi *Penicillium verruculosum*, which is highly active and the major endoglucanase of this organism [37]. It pertains to the glycosyl hydrolase family 5 (GH5). Cellulases from GH5 possess a $(\beta/\alpha)_8$ structure consisting of 8 β-sheets in the core of the protein and eight α-helices in the exterior.

In this work, we evaluate the potential of CNA to advance Phase I of KnowVolution by comparing the identification of beneficial positions through random mutagenesis (directed evolution) and CNA-based variant predictions (rational design) on the example of Cel5A to improve its thermostability. The performance of both strategies to identify beneficial positions was evaluated by comparing experimental effort and success in improved thermostability. Consequently, the CNA method is evaluated as an entry point in KnowVolution campaigns towards improved thermostability.

2. Materials and methods

2.1. Plasmids and strains

The strains *Escherichia coli* DH5α (Agilent Technologies, Santa Clara, CA, USA) and *Pichia pastoris* BSYBG11 (Bisy e.U., Hofstaetten/Raab, Austria) were used as a cloning and expression host, respectively. EGLII mutant libraries were cloned into the shuttle vector pBSYA1S1Z (Bisy e.U., Hofstaetten/Raab, Austria) and generated in *P. pastoris* BSYBG11. Endo-β-glucanase gene eglII from *Penicillium verruculosum* (EGLII; UniProtKB/Swiss-Prot: A0A1U7Q1U3) was purchased as codon-optimized synthetic gene fragment from ThermoFisher (Germany) and cloned into pBSYA1S1Z as described previously [38].

2.2. Library generation

Random mutagenesis was generated in the endoglucanase *egl*II gene by error-prone PCR (ep-PCR), as described by Contreras et al. [38]. Briefly, test libraries (consisting of 180 clones) were generated by ep-PCR with varying concentrations of $MnCl_2$ ranging from 0.1 to 0.4 mM. Test libraries were cloned by MEGAWHOP[39] and screened for improved thermostability as described in section "Screening of thermostable EGLII cellulase variants." The library generated from an ep-PCR supplemented with 0.3 mM $MnCl_2$ was selected for screening.

Site-saturation mutagenesis libraries at positions 76, 77, 92, 93, 114, 129, 130, 134, 189, 190, 222, 240, 244, 255, 256, 273, 299, 308, and 312 were generated by site-saturation mutagenesis (SSM) method [40]. SSM libraries were produced as described previously [38], and used NNK primers are detailed in **Table S1 in SI**. The resulting PCR product was digested using DpnI (37 °C, 18 h), purified by using the NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel), and transformed into *P. pastoris* BSYBG11 for expression.

2.3. Cell culture and expression

EGLII was expressed in *P. pastoris* BSYBG11 strain (Bisy e.U., Austria) cultured in 96-well microtiter plates (MTPs, Greiner, Frickenhausen, Germany). For expression, Yeast Extract–Peptone–Dextrose (YPD) medium (1% (w/v) yeast extract, 2% (w/v) peptone and 2% (w/v) D-glucose) supplemented with 100 µg mL^{-1} Zeocin was transferred to a MTP. A volume of 5 µL pre-culture (160 µL, 900 rpm, 30 °C, 48 h, and 70% humidity) was used to inoculate the main culture (160 µL, 900 rpm, 25 °C, 96 h, and 70% humidity) in an MTP supplemented with appropriate antibiotics. The supernatant containing EGLII was separated from the cells by centrifugation (Eppendorf 5810R; 4 °C, 3220 xg, 15 min), and the cell-free supernatant was transferred to a new MTP for further analysis.

For flask expression, a pre-culture of *P. pastoris* BSYBG11 was cultured in YPD-Zeocin medium (3 mL, 200 rpm, 30 °C, 24 h) and used to inoculate the main culture to an initial OD600mn of 0.25 for EGLII expression (50 mL, 200 rpm, 25 °C, 72 h). Cells were centrifuged (4 °C, 10,000 ×g, 20 min; Sorvall, Thermo Fischer Scientific, Darmstadt, Germany), and EGLII containing supernatant was used for further analysis.

2.4. Purification by ion-exchange chromatography

Purification of the endoglucanase EGLII was performed by anion exchange chromatography as described previously [38]. After flask expression, 50 mL of EGLII containing supernatant was concentrated by centrifugal ultrafiltration (10 kDa MWCO PES; VivaSpin turbo 15, Sartorius) to 2 mL, and the buffer was exchanged to Bis-Tris buffer (pH 6.5, 20 mM; buffer A). The endoglucanase EGLII was purified by FPLC (ÄKTAprime plus chromatography system, GE Healthcare, Solingen, Germany). The concentrated supernatant was loaded into an anion exchange chromatography column (GE Healthcare HiTrap Capto Q ImpRes, 5 mL), and equilibrated with buffer A. EGLII was eluted in a step-wise program; first, impurities were eluted with buffer B 26% (Bis-

177 tris buffer, pH 6.2, 20 mM, NaCl 1M) and later, EGLII was eluted with buffer B 33%.
178 Fractions were analyzed by SDS-Page (Figure S4 in SI) [37]. Endoglucanase EGLII
179 protein concentration was measured by A_{280nm} (NanoDrop™ 1000 spectrophotometer
180 by Thermo Scientific™, Bremen, Germany). Amino acid composition was used to
181 determine the theoretical extinction coefficients with ProtParam on the ExPASy
182 server [41].
183

184 **2.5. Screening of thermostable EGLII cellulase variants**

185 **2.5.1. Hydrolytic activity assays**

186 A Two-step screening system was employed, as described by Contreras *et al* [38].
187 Briefly, an agar-based pre-screening step was performed with Azo-carboxymethyl
188 cellulose (Azo-CMC; Megazyme, Bray, Ireland) supplemented YPD agar plates,
189 colonies presenting clear halos were selected as active for hydrolytic activity. In the
190 second step, the hydrolytic activity was screened by using solubilized Azo-CMC as
191 substrate. After cultivation in MTPs, EGLII-containing supernatant was diluted with
192 sodium acetate buffer (0.1 M, pH 4.5) and incubated without the substrate at 78 °C for
193 60 min. The diluted supernatant (40 µL) was transferred into an MTP for activity
194 measurement. The enzyme reaction was initiated by the addition of 40 µL of Azo-CMC
195 in sodium acetate buffer (2.0%, 0.1 M, pH 4.5). The reaction mixture was incubated at
196 50 °C with shaking (ELMI Ltd., SkyLine DTS-4 Digital Thermo Shaker, 900 rpm) for
197 exactly 10 min. The reaction was stopped by precipitating high-molecular-weight dyed
198 azo-CMC fragments with an ethanol-based precipitating solution (80% (v/v) technical
199 grade ethanol, 40 g L^{-1} sodium acetate, 4 g L^{-1} $ZnCl_2$, pH 5.0). The precipitated reaction
200 mix was centrifuged at 1000 xg for 10 min. Afterward, 100 µL of the clear supernatant
201 was transferred into an MTP, and the absorbance was measured at 590 nm (Tecan
202 sunrise, Crailsheim, Germany).

2.5.2. Thermostability of EGLII

For the quantification of the endoglucanase EGLII thermostability, the hydrolytic activity was measured in two conditions: without incubation of the supernatant containing EGLII ($Activity_{t0}$), and after incubation without the substrate at 78 °C for 60 min ($Activity_{t60}$). The residual activity of the EGLII WT and variants was determined as the ratio between the $Activity_{t60}$ and $Activity_{t0}$. A variant improvement was determined as the ratio between the residual activity of the wildtype and variant. As described previously [38], EGLII variants that maintained >80 % of EGLII wild type initial activity and presented increased thermostability were selected.

2.6. Specific activity determination

The hydrolytic activity of the purified endoglucanase EGLII was determined by the dinitrosalicylic acid assay (DNS), which quantifies the amount of reducing sugars released in the reaction [42] as described in[38]. Briefly, 20 µL of EGLII solution was mixed with 80 µL carboxymethyl cellulose (CMC) solved in sodium acetate buffer (50 mM, pH 4.5) to a final concentration of CMC 1 % (w/v) and incubated in a PCR cycler (96-PCR plate). The reaction mix was stopped with 200 µL of the DNS solution after exactly 10 min and incubated for 15 min at 95 °C to allow color revelation, followed by incubation for 10 min at 10 °C. The resulting color change was measured at 540 nm in a microtiter plate reader (Tecan Sunrise, Germany). The released sugar was calculated with cellobiose as standard. One unit of activity was defined as the amount of enzyme releasing 1 µmol of cellobiose equivalents from substrate per minute.

2.7. Generation of structural ensembles

As done previously [43], structural ensembles of wildtype EGLII (PDB ID 5L9C) were generated by all-atom MD simulations of in total 5 µs simulation time. For details on starting structure preparation, parametrization, equilibration, and production runs, see

SI Method M1. All minimization, equilibration, and production simulations were performed with the *pmemd.cuda* module [44] of Amber19 [45]. 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 [46]. 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.

2.8. Constraint Network Analysis

As done previously [43], the thermal unfolding simulations of wildtype EGLII was performed with the Constraint Network Analysis (CNA) software package (version 3.0) [21-24]. For details on thermal unfolding simulations, see SI Method M2. To improve the robustness and investigate the statistical uncertainty, we carried out CNA on ensembles of network topologies (ENT^{MD}) generated from structural ensembles (see section Generation of structural ensembles) [47].

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 $C^{E_{\mathrm{cut}}}$ [22]. 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, based on the coarse-grained, residue-wise network representation of the underlying protein structure [31]. In the present study, we applied the neighbor stability map $rc_{ij,neighbor}$ to investigate short-range rigid contacts. For this, as done previously [31, 33, 43], rc_{ij} was filtered such that only rigid contacts between two residues that are at most 5 Å apart from each other were considered.

2.9. Evolutionary conservation analysis

The degree of conservation of each amino acidic position was determined with the ConSurf server [48]. The amino acid sequence of the endoglucanase EGLII from *P. verroculosum* (PDB ID 5L9C) was used as template for the alignment. The conservation score was obtained after the alignment of 150 sequences with a similarity between 90 and 35%. Results are represented on a scale from 1 (not conserved) to 9 (highly conserved) (Figure S1 in SI).

3. Results and discussion

A main limitation that prevents a broader use of directed evolution in industrial applications is the time requirement of the campaigns. Through combined strategies of directed evolution and (semi-rational) design, like KnowVolution, experimental screening efforts can be minimized. The identification phase (Phase I, Figure 1) of a KnowVolution campaign can be changed by the CNA method, which could further reduce the time-requirement and screening burden in a protein engineering campaign towards improved thermostability. The potential of CNA in the identification of beneficial position for protein engineering campaigns is evaluated by the analysis of two types of libraries generated for improving EGLII thermostability (see workflow in Figure 1). First, we describe the directed evolution library generated by random mutagenesis; second, we explain the CNA approach for the prediction of positions ("structural weak spots") towards increased thermostability, and, finally, the generation and screening of a semi-rationally designed library guided by CNA is described. We propose the CNA approach as the first step in a protein engineering campaign. The computational screening will envisage the generation of a reduced library for thermostability improvement.

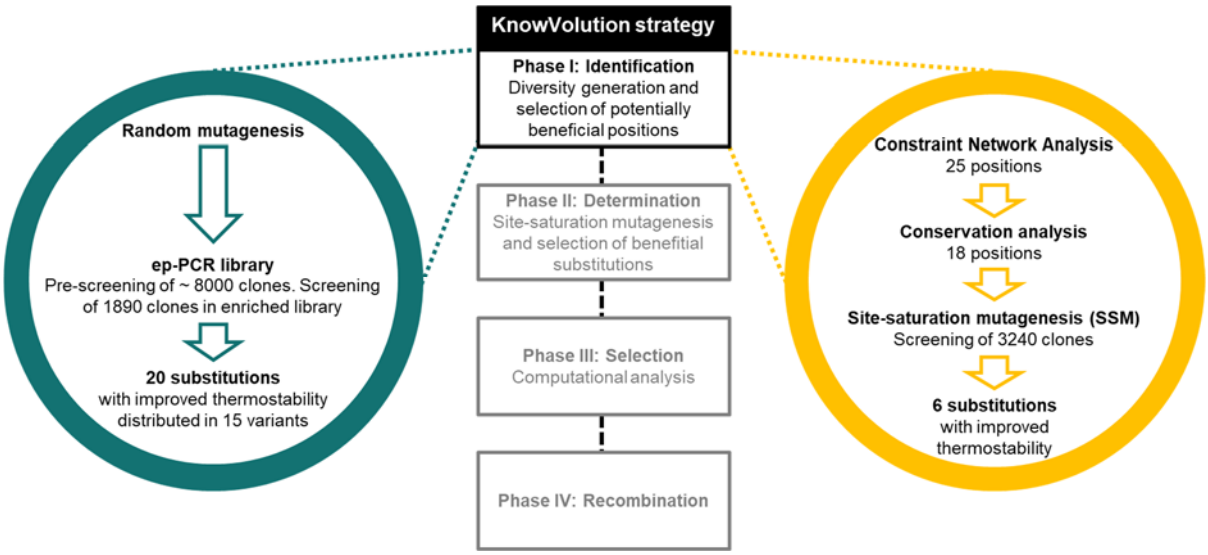

Figure 1: Protein engineering strategies for thermal stabilization of EGLII. Center: KnowVolution strategy with its four Phases (I, II, III, and IV); **Left:** A directed evolution campaign was performed in the complete endoglucanase EGLII gene by ep-PCR. **Right:** A semi-rational library design was performed, starting with a computational screening by the CNA approach, followed by an evolutionary conservation analysis, and finally, an SSM library of the 18 predicted "structural weak spots".

3.1. Random mutagenesis of EGLII towards increased thermostability

Randomly mutagenized libraries are commonly the first step in protein engineering campaigns for the identification of beneficial positions. Although it is a great advantage that no information is needed about the protein structure, the knowledge generated from random mutagenesis is not extensive. Frequently, improved variants can present several substitutions; therefore, it is difficult to acknowledge the role of each position in the improvement of the desired property (Figure 1).

In this work, a randomly mutagenized EGLII library was generated by ep-PCR. The ep-PCR library was optimized as described in "Library generation" section, and EGLII variants were screened towards increased thermostability as described previously [38]. The generated library was pre-screened in an agar plate-based assay, and later the library was enriched by transferring the active clones to a liquid culture and screened for improved thermostability utilizing an optimized Azo-CMC assay. In the agar plate-based assay, a library of ~8 000 clones was pre-screened and presented a 0.23

active/inactive ratio. An enriched library was produced with 1 890 active clones and was screened for improved thermostability. From the enriched library, 22 clones presented up to a 3.1-fold increase in thermostability compared to EGLII wild type, and after sequencing, 15 variants and 20 different substitutions were identified (Figure 2). The identified variants carried single, double, and triple substitutions. In total, 18 different positions were identified, but as they constitute double or triple variants, it is difficult to distinguish with certainty between the substitutions that are neutral, improving, or reducing EGLII thermostability. It is also uncertain if the found substitutions represent the best amino acids to improve the thermostability. These 18 positions represent 5.7% of the total EGLII amino acids (18 positions out of 314 amino acids), and it cannot be determined if all these positions influence the thermostability of the enzyme.

At the variant level, the yield of clones with increased thermostability from the ep-PCR library was 0.27% (22 out of ~8 000 clones). The yield increased to 1.16% (22 out of 1 890 clones) when an enriched library of only active clones was selected for screening. The latter is in agreement with previous directed evolution campaigns towards improved thermostability, in which the improved clones represented <1% of the total and enriched library [49-52].

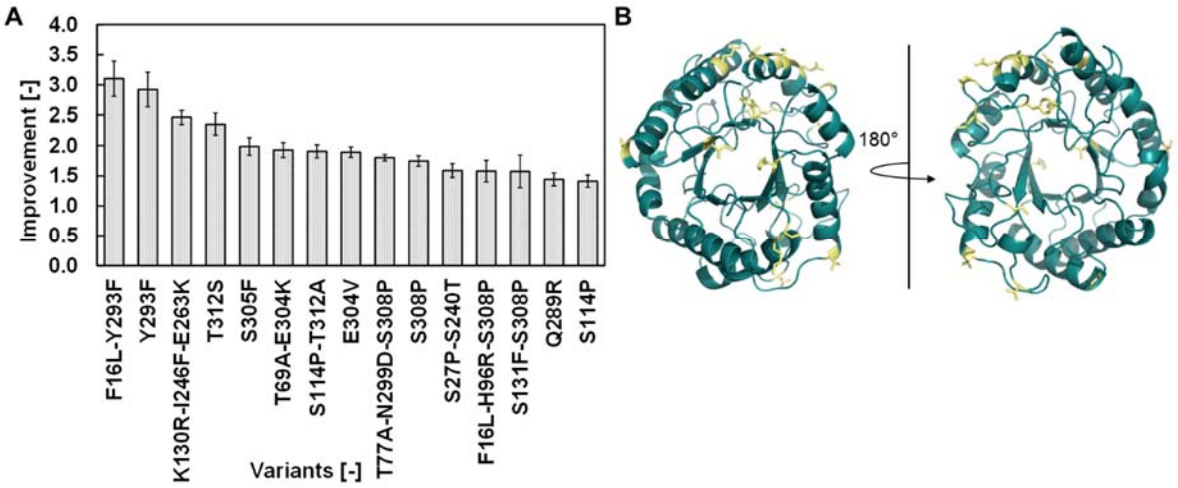

Figure 2: EGLII variants obtained from ep-PCR library. (A) Fifteen variants of EGLII with a significant thermostability improvement compared to the EGLII wild type. The improvement is defined as the ratio between the residual activity of the EGLII variants and the EGLII wild type, in AU. Given is the mean over experiments performed in biological replicates (*n* = 3). Error bars denote the standard error of the mean. **(B)** Representation of substituted positions (yellow sticks) in EGLII wildtype obtained from the ep-PCR library.

3.2. CNA approach towards increased thermostability

3.2.1. Prediction of structural weak spots

For identifying structural weak spots on EGLII, thermal unfolding simulations were carried out by CNA (section *Constraint Network Analysis*) on structural ensembles of wildtype EGLII generated by MD simulations (section *Generation of structural ensembles*), as done previously [21, 31, 34, 43]. By visual inspection of the unfolding trajectory, four major phase transitions, T1 – T4, were identified (Figure 3A). During unfolding, first, helix αA, second, helices αD and αG-I, then αC and αF, and, finally, αE segregated from the largest rigid cluster at 326, 338, 342, and 344 K. The hierarchy of rigid and flexible regions of EGLII showed that most helices segregated from the largest rigid cluster at T2, followed by T3. As most helices that are located at the C-terminus segregate from the largest rigid cluster at T2 and T3, this region is particularly promising for increasing thermostability, considering that substitutions there can improve the interaction strength with the largest rigid cluster and, hence, delay the disintegration of that cluster with increasing temperature.

Next, weak spots were identified as the fringe residues of such helices. In doing so, we followed the hypothesis that the more structurally stable the fringes of the helices are, the more structurally stable those regions will become. Therefore, if the fringe residues are targeted by substitutions, the likelihood to stabilize the protein should be high. Additionally, rc_{ij} and $rc_{ij,neighbor}$ were calculated to monitor the local rigidity of EGLII wild type at a per-residue level (Figure 3B). Both indices supported the above-mentioned observations. In total, 25 weak spots were predicted, i.e., 8% of all 314 EGLII residues (Figure 3C, Table 1).

CNA has been applied before retro- [21, 30-33] and prospectively [34, 35] in the context of improving protein thermostability on pairs [21, 30] and series [31, 32, 35] of proteins from psychro-, meso-, thermo-, and hyperthermophilic organisms and to proteins of different folds. Furthermore, the CNA approach was benchmarked against a complete site saturation mutagenesis library of *Bacillus subtilis* lipase A [53] for systematically scrutinizing the impact of substitution sites on thermostability and detergent tolerance [33]. Additionally, CNA was used to understand the impact of dimer interface stability on thermostability and the role of active site flexibility for turnover numbers in aldolases [35]. The breadth of these applications is rooted in the identification of structural weak spots [21] based on rigorous analysis of structural rigidity [25] applied to structural ensembles to improve robustness [47]. Hence, the use of CNA is not limited to a specific class or fold of proteins such that the strategy applied herein to identify structural weak spots can be transferred to non-TIM barrel proteins. This also includes structural water molecules, metal ions, and cofactors, which can be considered in the analysis. Finally, CNA has been used to understand signal propagation [54] and allosteric effects [55, 56] within biomolecules, primarily within membrane proteins, expanding the application scope of CNA.

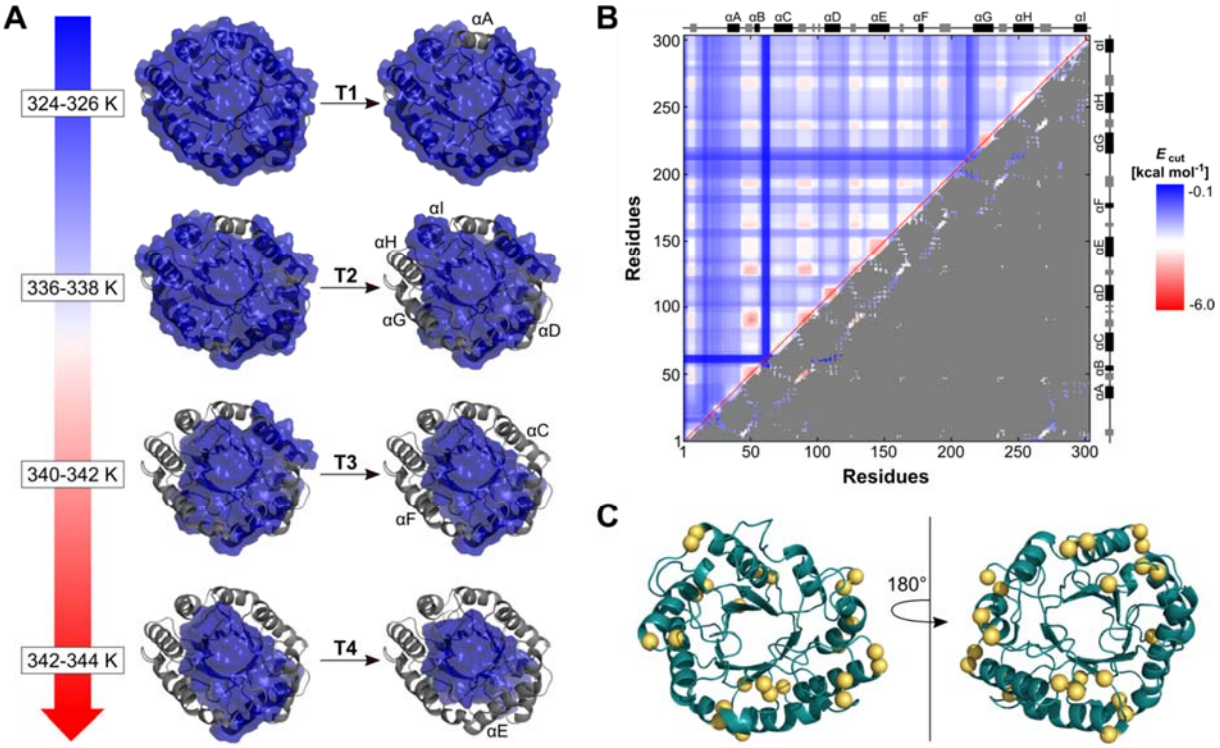

Figure 3: Prediction of the thermal unfolding pathway, local rigidity, and weak spots of wildtype EGLII. (A) Thermal unfolding pathway of wildtype EGLII (PDB ID: 5L9C [57]) showing four major phase transitions, T1-T4. The largest rigid cluster at each phase transition is represented as uniformly colored blue body. Helices that segregate from the largest rigid cluster at a phase transition are labeled. **(B)** Stability map rc_{ij} for wildtype EGLII including E_{cut} values at which a rigid contact between two residues (i, j) is lost during the thermal unfolding simulation **(upper triangle)**; the neighbor stability map $rc_{ij,neighbor}$ for wildtype EGLII 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)**. A red (blue) color indicates that contacts between residue pairs are more (less) rigid. α-helices and β-strands are depicted at the top. **(C)** Localization of predicted weak spots of wildtype EGLII (yellow spheres).

After the CNA identification, the evolutionary conservation of the selected weak spots was analyzed to further reduce experimental efforts [48]. A high conservation score is an indicator of the functional or structural importance of a position in the protein. Thus, weak spots with a conservation score ≥ 8 (Table 1, Figure S1 in SI) were excluded from the experimental analysis. This resulted in the selection of 18 weak spots (D76, T77, G92, K93, S114, F129, K130, L134, N189, T190, V222, S240, L244, S256, S273, N299, S308, T312), which were targeted for site-saturation mutagenesis (SSM),

allowing to focus substitution efforts on only ~6% of the protein residues. A similar percentage was found for weak spot predictions on *Bs*LipA [33]. The selected weak spots are mainly located in the outer α-helices and loops (including turns, bends, and beta bridges) of the $(\beta/\alpha)_8$ barrel structure (Table 1, Figure S1 in SI).

Table 1: Phase transitions during the thermal unfolding simulation of wildtype EGLII, predicted weak spots, and their evolutionary conservation scores.

Phase transition	Weak spot	Secondary structure element[a]	Conservation score[b]
T3	Asp76	Loop	6
T3	Thr77	αC	1
T3	Gly92	αC	1
T3	Lys93	Turn	3
T2	Ile112	Loop	**8**
T2	Ser114	Loop	6
T2	Phe129	Turn	7
T2	Lys130	Turn	2
T2	Leu134	Turn	2
T3	Gly181	Beta Bridge	**9**
T3	Ala182	Bend	**9**
T2	Asn189	Loop	5
T2	Thr190	3/10 helix	1
T2	Cys221	Loop	**9**
T2	Val222	Bend	7
T2	Ser240	αG	1
T2	Leu244	Loop	4
T2	Asn255	Bend	**8**
T2	Ser256	αH	1
T2	Ser273	Turn	3
T2	Asp274	Turn	**8**
T2	Asn299	Turn	3
T2	Gly300	Loop	**8**
T2	Ser308	αI	4
T2	Thr312	Turn	1

[a] Localization of the respective weak spot.

[b] Values $\geq$ 8 are marked in bold and led to the exclusion of the weak spot from experimental analysis.

3.2.2. Effects of substitutions at weak spots on EGLII thermostability

The 18 selected weak spots were saturated by SSM to systematically probe the effects of the weak spots on the thermostability of EGLII. For each selected weak spot, primers were designed (Table S1 in SI), a site-saturation reaction was performed and individually transformed in *P. pastoris*. Each SSM library consisted of 180 clones, and 3 240 clones were screened in total (18 positions with 180 clones each), covering ~98% [58] of possible substitutions in each position. This results in a reduction of experimental efforts to ~40% compared with the pre-screening of ~8 000 clones of the ep-PCR library. In our opinion, experimental efforts derived from library production (PCR reaction, cloning, transformation, and library optimization) are comparable in ep-PCR and SSM approaches.

On the basis of the SSM libraries, the selected weak spots showed different influences on the activity of EGLII, with an inactivation percentage ranging from 1 to 87% in positions 308 and 76, respectively (Table S2 in SI). At five weak spots (T77, V222, L244, S308, and T312), at least one substitution improved EGLII thermostability. The variant T312R showed the highest improvement (1.99-fold) compared to EGLII wildtype (Table 2). Of note, when T312 was substituted by S in the random mutagenesis approach (Figure 2), a ~2.4-fold improvement was found, confirming that position 312 is a weak spot. Likewise, for S308P, between ~1.7 and ~1.4-fold improvement was found in the random mutagenesis approach (Figure 2), whereas the improvement is 1.16-fold when screening the SSM libraries (Table 2). The improvement difference may result from different expression levels of the variants due to silent mutations present in the genes. These five positions represent 1.6% of the total EGLII amino acids (5 positions out of 314 amino acids). As for substitutions at predicted weak spots, six substitutions (T312R, T77V, T77E, L244R, V222P, and S308P) were found to yield increased thermostability.

Saturation of 18 selected weak spots enables an accurate assessment of each positions' influence over EGLII thermostability. For example, in random mutagenesis, position T77 was found as a triple variant (Figure 2). Therefore, it is unclear if thermostability improvement is driven from position T77, N299, or S308; or if they possess an additive effect. Substitutions must be single tested either by SDM or SSM to determine their influence on EGLII thermostability. Considering all identified variants by random mutagenesis (section "Random mutagenesis of EGLII towards increased thermostability"), 18 SSM libraries (or 18 SDM) should be produced additionally for each ep-PCR identified position. As a result, screening efforts would be incremented.

Table 2: Variants identified in SSM libraries at predicted weak spots with increased thermostability

Phase transition	Secondary structure element	Weak spot	Substitution	Improvement[a]
T2	Turn	312	T312R	1.99 ± 0.30
T3	αC	77	T77V	1.25 ± 0.04
T3	αC	77	T77E	1.24 ± 0.07
T2	Loop	244	L244R	1.23 ± 0.09
T2	Bend	222	V222P	1.16 ± 0.09
T2	αI	308	S308P	1.16 ± 0.16

[a] Improvement is defined as the ratio between the residual activity of the EGLII variants and the EGLII wild type in AU. Given is the mean ± SEM over n = 3 experiments performed in biological replicates.

High thermostability and high activity at lower temperatures are properties not found together in enzymes in nature [59, 60]. Protein engineering has advanced as an important toolbox for improving more than one feature within a protein. Through a screening system that select enzymes with increased thermostability and retained activity at lower temperatures, variants that meet both characteristics can be identified. The specific activity of the identified variants with improved thermostability (T312R, T77V, T77E, L244R, V222P, and S308P) was determined at 75 °C, EGLII wild type

optimum temperature [38], and 30 °C. Endoglucanase EGLII wild type presents a specific activity of 249 ± 38 U mg^{-1} at 75 °C and 58 ± 4 U mg^{-1} at 30 °C (Table S3 in SI). It is noteworthy that all six variants obtained from the CNA strategy retained >90% of EGLII wild type specific activity at 75 °C, and 100% at 30 °C. The CNA strategy enables the identification of variants with improved thermostability and retained activity at lower temperatures.

As to the structural basis of increased thermostability in EGLII variants, in both T312R and L244R a salt bridge with residues on neighboring secondary structure elements can form, which was absent in EGLII wildtype. Salt bridges are considered to enhance thermostability in the majority of cases [61] (Figure S2 in SI). Concerning the other substitutions, favorable enthalpic contributions appear less likely to be determinants of thermostabilization. Position 77 is exposed to the solvent, such that direct interactions to neighboring residues are not possible (Figure S2 in SI). In thermophilic proteins, glutamate residues in the protein surface can stabilize exposed structures by increasing the polar surface area, which could explain the stabilization of substitution T77E [62-64]. Substitution T77V may yield a stabilization of the N-terminal part of the helix because of the more favorable helix propensity of valine compared to threonine [65]. Positions 222 and 308 comprise substitutions to proline, which have a unique role in determining local conformation [66, 67] that may lead to an entropy-driven thermostabilization [68].

At the variant level, the screened 3 240 clones yielded a success rate of 0.18% (6 out of 3 240) of variants with increased thermostability. Depending on the influence of each selected weak spot in the activity of EGLII, the screening effort could be further reduced by pre-screening in an agar plate >180 clones in each position and doing an enrichment with just the necessary clones needed to fulfill a >98% of coverage [58]. If, analogously to ep-PCR, an enriched SSM library of only 1 660 active clones is

476 produced (Table S2 in SI), the success rate can rise to 0.36% (6 out of 1 660) and
477 further reduce the screening in MTP by ~40%. Within a KnowVolution campaign, the
478 major screening load comes from Phases I and II and comprises libraries of thousands
479 of clones. Therefore, the CNA approach could be used beneficially in Phase I
480 (Identification) of a KnowVolution campaign for improved thermostability [38], and,
481 consequently, reduce the screening effort further.
482 Due to the lack of a complete site saturation library for EGLII, as available, e.g., for
483 *Bs*LipA [33] and the domain protein G (Gβ1) [69], the true number of weak spots in
484 EGLII at which at least one substitution leads to increased thermostability remains
485 unknown. Hence, the *precision in random classification* cannot be calculated rigorously
486 and, thus, neither the *gain in precision over random classification* (*gip*) due to our CNA
487 and conservation score analyses. Nevertheless, a lower estimate of *gip* is possible
488 when one assumes that the 18 positions identified in the random mutagenesis
489 approach constitute the true weak spots of the protein. Then, *gip* = (# of confirmed
490 predicted weak spots / # of predicted weak spots) / (# of true weak spots / # of amino
491 acids) = (5 / 18) / (18 / 314) = 4.8 [33], demonstrating a ~5-fold higher likelihood to
492 identify weak spots by CNA and conservation score analyses over random
493 identification.
494 Note that, by the design of the CNA approach, the identification of structural weak
495 spots aims at improving thermodynamic thermostability [3, 4, 31], whereas kinetic
496 parameters leading to irreversible denaturation elude the CNA analysis. In this context,
497 CNA has not yet been applied to scrutinize the effect of insertions or deletions on
498 thermostability. Likely, the most direct application would be to identify structurally less
499 stable loop regions that may give rise to local unfolding events, from which irreversible
500 inactivation may occur, and suggest those for deletion. Alternatively, the impact of
501 insertions or deletions suggested from sequence analyses on structural stability could

be analyzed with CNA because such effects may not be restricted locally but lead to changes across the protein, presumably through packing changes [70].

Finally, several advantages arise from the SSM libraries generated from the CNA analysis compared with the random mutagenesis library. First, positions that improve EGLII thermostability can be identified with certainty. Second, it can be established which amino acid represents the best substitution for a position. Third, the influence of each position can be quantitatively determined for each improved variant.

4. Conclusion

Constraint Network Analysis (CNA) is a promising method for the identification of beneficial positions in Phase I of a KnowVolution campaign for thermostability improvement of the endo-β-glucanase Cel5A and can likely be applied to other enzymes. Screening efforts can be reduced to ~40% compared to a randomly mutagenized library based on an estimated ~5-fold higher likelihood to identify weak spots by CNA and conservation score analyses over random identification. The focused work performed in CNA-predicted weak spots yields a success rate (0.18%) in identifying variants with increased thermostability similar to random mutagenesis (0.27%). These results reduce time-requirements in directed evolution campaigns. The CNA-based identification of beneficial positions becomes particularly interesting if high-throughput screening systems are not available.

Authors information

ORCID

Francisca Contreras: 0000-0001-8134-1445

Christina Nutschel: 0000-0001-5498-0911

Mehdi D. Davari: 0000-0003-0089-7156

Ulrich Schwaneberg: 0000-0003-4026-701X

Holger Gohlke: 0000-0001-8613-1447

Notes

Declarations of interest: none

Author contributions

M.D.D. and H.G. conceived the study. F.C. conceived, planned, and performed the experiments, analyzed the results, and wrote the manuscript. C.N. conceived, planned the computer experiments, performed the computational analyses, analyzed the results, and wrote the manuscript. L.B. planned and performed the experiments, analyzed the results, and revised the manuscript. U.S., M.D.D., and H.G. discussed the results and revised the manuscript.

Acknowledgment

The work was supported by the German Federal Ministry of Education and Research (BMBF, FKZ: 031B0506, EnzyBioDeg project, Bioökonomie International 2016). CN is funded through a grant ("Vernetzungsdoktorand") provided by the Forschungszentrum Jülich. 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)) [71]. We acknowledge Prof. Arkady P. Sinitsyn, Dr. Aleksandra M. Rozhkova,

and Dr. Ivan N. Zorov (Federal Research Centre "Fundamentals of Biotechnology", Russia) for providing the EGLII sequence.

Abbreviations

CNA, Constraint Network Analysis; PCR, polymerase chain reaction; HTS, high-throughput screening; EGLII, endoglucanase II; SSM, site-saturation mutagenesis; MTP, 96-well microtiter plates; CMC, carboxymethyl cellulose; MD, molecular dynamics; AU, absorbance units.

References

1. Atalah J, Cáceres-Moreno P, Espina G, Blamey JM. Thermophiles and the applications of their enzymes as new biocatalysts. Bioresour Technol. 2019;280:478-88, doi: 10.1016/j.biortech.2019.02.008.
2. Xu Z, Cen Y-K, Zou S-P, Xue Y-P, Zheng Y-G. Recent advances in the improvement of enzyme thermostability by structure modification. Crit Rev Biotechnol. 2020;40(1):83-98, doi: 10.1080/07388551.2019.1682963.
3. Amerah A, Gilbert C, Simmins P, Ravindran V. Influence of feed processing on the efficacy of exogenous enzymes in broiler diets. Worlds Poult Sci J. 2011;67(1):29-46, doi: 10.1017/S0043933911000031.
4. Stepnov AA, Fredriksen L, Steen IH, Stokke R, Eijsink VG. Identification and characterization of a hyperthermophilic GH9 cellulase from the Arctic Mid-Ocean Ridge vent field. PloS One. 2019;14(9):e0222216, doi: 10.1371/journal.pone.0222216.
5. Graham JE, Clark ME, Nadler DC, Huffer S, Chokhawala HA, Rowland SE, et al. Identification and characterization of a multidomain hyperthermophilic cellulase from an archaeal enrichment. Nat Commun. 2011;2(1):1-9, doi: 10.1038/ncomms1373.
6. Bommarius AS, Paye MF. Stabilizing biocatalysts. Chem Soc Rev. 2013;42(15):6534-65, doi: 10.1039/C3CS60137D.
7. Cheng F, Zhu L, Schwaneberg U. Directed evolution 2.0: improving and deciphering enzyme properties. Chem Commun. 2015;51(48):9760-72, doi: 10.1039/C5CC01594D.
8. Cui H, Cao H, Cai H, Jaeger KE, Davari MD, Schwaneberg U. Computer-assisted Recombination (CompassR) teaches us how to recombine beneficial substitutions from directed evolution campaigns. Chem Eur J. 2019;26(3):643-9, doi: 10.1002/chem.201903994.
9. Nannemann DP, Birmingham WR, Scism RA, Bachmann BO. Assessing directed evolution methods for the generation of biosynthetic enzymes with potential in drug biosynthesis. Future Med Chem. 2011;3(7):809-19, doi: 10.4155/fmc.11.48.
10. Leemhuis H, Kelly RM, Dijkhuizen L. Directed evolution of enzymes: library screening strategies. IUBMB life. 2009;61(3):222-8, doi: 10.1002/iub.165.
11. Reetz MT, Carballeira JD. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. Nat Protoc. 2007;2(4):891-903, doi: 10.1038/nprot.2007.72
12. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinform. 2011;12(1):1-12, doi: 10.1186/1471-2105-12-151.
13. Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: Working with RNA, small molecules and a new graphical interface. Bioinformatics. 2019;35(20):4168-9, doi: 10.1093/bioinformatics/btz184.
14. Buss O, Rudat J, Ochsenreither K. FoldX as protein engineering tool: Better than random based approaches? Comput Struct Biotechnol J. 2018;16:25-33, doi: 10.1016/j.csbj.2018.01.002.
15. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33(Web Server issue):W382-W8, doi: 10.1093/nar/gki387.
16. Musil M, Stourac J, Bendl J, Brezovsky J, Prokop Z, Zendulka J, et al. FireProt: web server for automated design of thermostable proteins. Nucleic Acids Res. 2017;45(W1):W393-W9, doi: 10.1093/nar/gkx285.

17. Wijma HJ, Floor RJ, Jekel PA, Baker D, Marrink SJ, Janssen DB. Computationally designed libraries for rapid enzyme stabilization. Protein Eng Des Sel. 2014;27(2):49-58, doi: 10.1093/protein/gzt061.
18. Wijma HJ, Furst M, Janssen DB. A Computational Library Design Protocol for Rapid Improvement of Protein Stability: FRESCO. In: Bornscheuer U, M. H, editors. Methods in molecular biology 1685. 2017/11/01 ed. New York, NY: Humana Press; 2018. p. 69-85.
19. Goldenzweig A, Fleishman SJ. Principles of Protein Stability and Their Application in Computational Design. Annu Rev Biochem. 2018;87:105-29, doi: 10.1146/annurev-biochem-062917-012102.
20. Cadet F, Fontaine N, Vetrivel I, Chong MNF, Savriama O, Cadet X, et al. Application of fourier transform and proteochemometrics principles to protein engineering. BMC Bioinform. 2018;19(1):1-11, doi: 10.1186/s12859-018-2407-8.
21. Radestock S, Gohlke H. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng Life Sci. 2008;8(5):507-22, doi: 10.1002/elsc.200800043.
22. Pfleger C, Radestock S, Schmidt E, Gohlke H. Global and local indices for characterizing biomolecular flexibility and rigidity. J Comput Chem. 2013;34(3):220-33, doi: 10.1002/jcc.23122.
23. 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(W1):W340-W8, doi: 10.1093/nar/gkt292.
24. Hermans SM, Pfleger C, Nutschel C, Hanke CA, Gohlke H. Rigidity theory for biomolecules: concepts, software, and applications. Wiley Interdiscip Rev Comput Mol Sci. 2017;7(4):e1311, doi: 10.1002/wcms.1311.
25. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. Proteins: Struct Funct Bioinform. 2001;44(2):150-65, doi: 10.1002/prot.1081.
26. Hespenheide B, Jacobs D, Thorpe M. Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. J Phys: Condens Matter. 2004;16(44):S5055, doi: 10.1088/0953-8984/16/44/003.
27. Jacobs DJ, Thorpe MF. Generic rigidity percolation: the pebble game. Phys Rev Lett. 1995;75(22):4051-4, doi: 10.1103/PhysRevLett.75.4051.
28. Rader A, Hespenheide BM, Kuhn LA, Thorpe MF. Protein unfolding: rigidity lost. Proc Natl Acad Sci USA. 2002;99(6):3540-5, doi: 10.1073/pnas.062492699.
29. Livesay DR, Jacobs DJ. Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. Proteins: Struct Funct Bioinform. 2006;62(1):130-43, doi: 10.1002/prot.20745.
30. Radestock S, Gohlke H. Protein rigidity and thermophilic adaptation. Proteins: Struct Funct Bioinform. 2011;79(4):1089-108, doi: 10.1002/prot.22946.
31. Rathi PC, Jaeger K-E, Gohlke H. Structural rigidity and protein thermostability in variants of lipase A from Bacillus subtilis. PloS One. 2015;10(7):e0130289, doi: 10.1371/journal.pone.0130289.
32. Rathi PC, Radestock S, Gohlke H. Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. J Biotechnol. 2012;159(3):135-44, doi: 10.1016/j.jbiotec.2012.01.027.
33. 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(3):1568–84, doi: 10.1021/acs.jcim.9b00954.

34. Rathi PC, Fulton A, Jaeger K-E, Gohlke H. Application of rigidity theory to the thermostabilization of lipase A from Bacillus subtilis. PLoS Comp Biol. 2016;12(3):e1004754, doi: 10.1371/journal.pcbi.1004754.
35. Dick M, Weiergräber OH, Classen T, Bisterfeld C, Bramski J, Gohlke H, et al. Trading off stability against activity in extremophilic aldolases. Sci Rep. 2016;6:17908, doi: 10.1038/srep17908.
36. Sharma A, Tewari R, Rana SS, Soni R, Soni SK. Cellulases: classification, methods of determination and industrial applications. Appl Biochem Biotechnol. 2016;179(8):1346-80, doi: 10.1007/s12010-016-2070-3
37. Morozova VV, Gusakov AV, Andrianov RM, Pravilnikov AG, Osipov DO, Sinitsyn AP. Cellulases of Penicillium verruculosum. Biotechnol J. 2010;5(8):871-80, doi: 10.1002/biot.201000050
38. Contreras F, Thiele MJ, Pramanik S, Rozhkova AM, Dotsenko AS, Zorov IN, et al. KnowVolution of a GH5 cellulase from Penicillium verruculosum to improve thermal stability for biomass degradation. ACS Sustain Chem Eng. 2020;8(33):12388–99, doi: 10.1021/acssuschemeng.0c02465.
39. Miyazaki K. Creating Random Mutagenesis Libraries by Megaprimer PCR of Whole Plasmid (MEGAWHOP). In: Arnold FH, Georgiou G, editors. Directed Evolution Library Creation. Totowa, NJ: Humana Press; 2003. p. 23-8.
40. Wang W, Malcolm BA. Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange Site-Directed Mutagenesis. BioTechniques. 1999;26(4):680-2, doi: 10.2144/99264st03.
41. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editor. The proteomics protocols handbook. Totowa, NJ: Humana Press; 2005. p. 571-607.
42. Miller GL. Use of dinitrosalicylic acid reagent for determination of reducing sugar. Anal Chem. 1959;31(3):426-8, doi: 10.1021/ac60147a030.
43. Nutschel C, Mulnaes D, Coscolin C, Ferrer M, Jaeger K-E, Gohlke H. Promiscuous esterases counterintuitively are less flexible than specific ones. bioRxiv. 2020, doi: 10.1101/2020.06.02.129015.
44. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. J Chem Theory Comput. 2013;9(9):3878-88, doi: 10.1021/ct400314y.
45. D.A. Case KB, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York and P.A. Kollman AMBER 2019. University of California, San Francisco2019.
46. Hopkins CW, Le Grand S, Walker RC, Roitberg AE. Long-time-step molecular dynamics through hydrogen mass repartitioning. J Chem Theory Comput. 2015;11(4):1864-74, doi: 10.1021/ct5010406.
47. Pfleger C, Gohlke H. Efficient and robust analysis of biomacromolecular flexibility using ensembles of network topologies based on fuzzy noncovalent constraints. Structure. 2013;21(10):1725-34, doi: 10.1016/j.str.2013.07.012.
48. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation

in macromolecules. Nucleic Acids Res. 2016;44(W1):W344-W50, doi: 10.1093/nar/gkw408.
49. Anbar M, Lamed R, Bayer EA. Thermostability enhancement of Clostridium thermocellum cellulosomal endoglucanase Cel8A by a single glycine substitution. ChemCatChem. 2010;2(8):997-1003, doi: 10.1002/cctc.201000112.
50. Chen C, Su L, Xu F, Xia Y, Wu J. Improved Thermostability of Maltooligosyltrehalose Synthase from Arthrobacter ramosus by Directed Evolution and Site-Directed Mutagenesis. J Agric Food Chem. 2019;67(19):5587-95, doi: 10.1021/acs.jafc.9b01123.
51. Schmidt A, Shvetsov A, Soboleva E, Kil Y, Sergeev V, Surzhik M. Thermostability improvement of Aspergillus awamori glucoamylase via directed evolution of its gene located on episomal expression vector in Pichia pastoris cells. Protein Eng Des Sel. 2019;32(6):251-9, doi: 10.1093/protein/gzz048.
52. Shivange AV, Serwe A, Dennig A, Roccatano D, Haefner S, Schwaneberg U. Directed evolution of a highly active Yersinia mollaretii phytase. Appl Microbiol Biotechnol. 2012;95(2):405-18, doi: 10.1007/s00253-011-3756-7
53. Fulton A, Frauenkron-Machedjou VJ, Skoczinski P, Wilhelm S, Zhu L, Schwaneberg U, et al. Exploring the protein stability landscape: Bacillus subtilis lipase A as a model for detergent tolerance. ChemBioChem. 2015;16(6):930-6, doi: 10.1002/cbic.201402664.
54. Milić D, Dick M, Mulnaes D, Pfleger C, Kinnen A, Gohlke H, et al. Recognition motif and mechanism of ripening inhibitory peptides in plant hormone receptor ETR1. Sci Rep. 2018;8(1):3890, doi: 10.1038/s41598-018-21952-3.
55. Wifling D, Pfleger C, Kaindl J, Ibrahim P, Kling RC, Buschauer A, et al. Basal Histamine H4 Receptor Activation: Agonist Mimicry by the Diphenylalanine Motif. Chem Eur J. 2019;25(64):14613-24, doi: 10.1002/chem.201902801.
56. Preising MN, Görg B, Friedburg C, Qvartskhava N, Budde BS, Bonus M, et al. Biallelic mutation of human SLC6A6 encoding the taurine transporter TAUT is linked to early retinal degeneration. FASEB J. 2019;33(10):11507-27, doi: 10.1096/fj.201900914RR.
57. Nemashkalov V, Vakhrusheva A, Tishchenko S, Gabdulkhakov A, Kravchenko O, Gusakov A, et al. Crystal structure of an endoglucanase from Penicillium verruculosum in complex with cellobiose. http://www.rcsb.org/structure/5L9C; 2016.
58. Reetz MT, Kahakeaw D, Lohmer R. Addressing the numbers problem in directed evolution. ChemBioChem. 2008;9(11):1797-804, doi: 10.1002/cbic.200800298.
59. Schreiber G, Buckle AM, Fersht AR. Stability and function: two constraints in the evolution of barstar and other proteins. Structure. 1994;2(10):945-51, doi: 10.1016/s0969-2126(94)00096-4.
60. Harms MJ, Thornton JW. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat Rev Genet. 2013;14(8):559-71, doi: 10.1038/nrg3540.
61. Kumar S, Tsai C-J, Nussinov R. Factors enhancing protein thermostability. Protein Eng. 2000;13(3):179-91, doi: 10.1093/protein/13.3.179.
62. Vogt G, Argos P. Protein thermal stability: hydrogen bonds or internal packing? Fold Des. 1997;2:S40-S6, doi: 10.1016/s1359-0278(97)00062-x.
63. Pack SP, Yoo YJ. Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. J Biotechnol. 2004;111(3):269-77, doi: 10.1016/j.jbiotec.2004.01.018.
64. Li W, Zhou X, Lu P. Structural features of thermozymes. Biotechnol Adv. 2005;23(4):271-81, doi: 10.1016/j.biotechadv.2005.01.002.

761 65. Pace CN, Scholtz JM. A helix propensity scale based on experimental studies
762 of peptides and proteins. Biophys J. 1998;75(1):422-7, doi: 10.1016/s0006-
763 3495(98)77529-0.
764 66. MacArthur MW, Thornton JM. Influence of proline residues on protein
765 conformation. J Mol Biol. 1991;218(2):397-412, doi: 10.1016/0022-2836(91)90721-h.
766 67. Betts MJ, Russell RB. Amino acid properties and consequences of
767 substitutions. In: Barnes MR, Gray IC, editors. Bioinformatics for geneticists. 317:
768 Wiley; 2003. p. 289.
769 68. Liu Z, Lemmonds S, Huang J, Tyagi M, Hong L, Jain N. Entropic contribution to
770 enhanced thermal stability in the thermostable P450 CYP119. Proc Natl Acad Sci USA.
771 2018;115(43):E10049-E58, doi: 10.1073/pnas.1807473115.
772 69. Nisthal A, Wang CY, Ary ML, Mayo SL. Protein stability engineering insights
773 revealed by domain-wide comprehensive mutagenesis. Proc Natl Acad Sci USA.
774 2019;116(33):16367-77, doi: 10.1073/pnas.1903888116.
775 70. Zeiske T, Stafford KA, Palmer III AG. Thermostability of enzymes from
776 molecular dynamics simulations. J Chem Theory Comput. 2016;12(6):2489-92, doi:
777 10.1021/acs.jctc.6b00120.
778 71. Krause D. JUWELS: Modular Tier-0/1 Supercomputer at the Jülich
779 Supercomputing Centre. JLSRF. 2019;5:135, doi.
780