ACS Catalysis

Perspective

Cite This: ACS Catal. 2019, 9, 1033−1054

pubs.acs.org/acscatalysis

Computational Design of Stable and Soluble Biocatalysts

Milos Musil,[†,‡,§,∥] Hannes Konegger,[†,§,∥] Jiri Hon,[†,‡,§,∥] David Bednar,[†,§] and Jiri Damborsky*[,†,§]

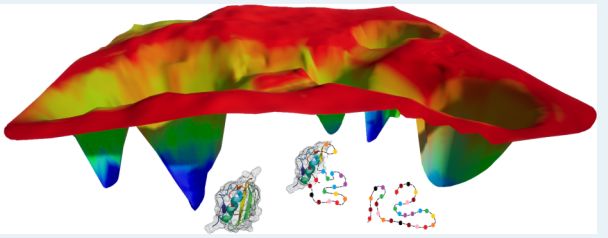
[†]Loschmidt Laboratories, Centre for Toxic Compounds in the Environment (RECETOX), and Department of Experimental Biology, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic

[‡]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic

[§]International Clinical Research Center, St. Anne's University Hospital, Pekarska 53, 656 91 Brno, Czech Republic

S *Supporting Information*

ABSTRACT: Natural enzymes are delicate biomolecules possessing only marginal thermodynamic stability. Poorly stable, misfolded, and aggregated proteins lead to huge economic losses in the biotechnology and biopharmaceutical industries. Consequently, there is a need to design optimized protein sequences that maximize stability, solubility, and activity over a wide range of temperatures and pH values in buffers of different composition and in the presence of organic cosolvents. This has created great interest in using computational methods to enhance biocatalysts' robustness and solubility. Suitable methods include (i) energy calculations, (ii) machine learning, (iii) phylogenetic analyses, and (iv) combinations of these approaches. We have witnessed impressive progress in the design of stable enzymes over the last two decades, but predictions of protein solubility and expressibility are scarce. Stabilizing mutations can be predicted accurately using available force fields, and the number of sequences available for phylogenetic analyses is growing. In addition, complex computational workflows are being implemented in intuitive web tools, enhancing the quality of protein stability predictions. Conversely, solubility predictors are limited by the lack of robust and balanced experimental data, an inadequate understanding of fundamental principles of protein aggregation, and a dearth of structural information on folding intermediates. Here we summarize recent progress in the development of computational tools for predicting protein stability and solubility, critically assess their strengths and weaknesses, and identify apparent gaps in data and knowledge. We also present perspectives on the computational design of stable and soluble biocatalysts.



KEYWORDS: *aggregation, computational design, force field, expressibility, machine learning, phylogenetic analysis, enzyme stability, enzyme solubility*

1. INTRODUCTION

Nature has developed a remarkable diversity of biochemical reactions that are vital to the continuing evolution of living organisms and the preservation of life. Enzymes are the most prominent catalytic entities in living cells and are collectively capable of catalyzing a vast range of biochemical reactions. The advent of next-generation sequencing together with recent advances in bioinformatics and molecular and structural biology have granted ready access to these rich genetic resources, facilitating the identification of efficient biocatalysts for diverse applications.[1−4] Moreover, the field of protein engineering has matured to a level that allows tailoring of native enzymes for specific practical applications.[5] However, the redesign of an enzyme sequence often imposes unintended secondary effects, frequently reducing the solubility and stability of the target enzyme.[6−9] Strategies for mitigating or eliminating these negative effects include chaperone buffering,[10] chemical modification of the protein structure,[11,12] protein immobilization,[13] medium engineering,[13] the addition of fusion proteins,[14,15] and the introduction of stabilizing or solubilizing mutations by protein engineering.[16−18]

Of particular interest for a mutational strategy is "directed evolution", which refers to experimental methods that emulate natural evolution by coupling molecular diversity generation to a selection or screening process. However, the immensity of an enzyme's sequence space prohibits global evaluation of all possible mutational combinations,[19] frequently causing optimization trajectories to become stuck in evolutionary dead ends.[20,21] This restricts the scope for creating stable and soluble biocatalysts by directed evolution alone and calls for knowledge-guided approaches to navigate the mutational space.[22] Rational protein design strategies can dramatically reduce the experimental effort required for successful directed evolution by consolidating pre-existing information.[23] Semirational strategies that combine directed evolution with structural and sequence data to help identify mutational hotspots amenable to focused screening efforts have been particularly popular recently.[24−26]

Received: September 7, 2018
Revised: December 15, 2018
Published: December 18, 2018

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

Table 1. Selected Experimentally Validated Cases of Successful Computational Redesigns of Stable and Soluble Biocatalysts

Stable Biocatalysts

enzyme | UniProt ID	substrate	method[j]	mutant code	mutations[a]	wild-type T_m [°C]	ΔT_m [°C][b]	$t_{1/2}$[c]	specific activity[c]	k_{cat}/K_m[c]	ref
cutinase | P52956	4-nitrophenyl butyrate	force field	variant 10	7 of 197	62.3	5.7	12.9× (60 °C)	0.64× (25 °C)	n.d.[d]	41
keratinase | Q1EM64	keratin	machine learning	quadruple mutant	4 of 379	n.d.	n.d.	8.6× (60 °C)	n.d.	4.11× (40 °C)	42
adenylate kinase | P16304	Mg/ATP, AMP	phylogeny (ASR)	ANC1	66 of 218	53.6	35.4	n.d.	n.d.	1.79× (25 °C)	43
β-lactamase | P62593	benzylpenicillin	phylogeny (CD)	ALL-CON	122 of 262	55.0	23.6	n.d.	n.d.	0.03× (25 °C)	44
kemp eliminase | Q06121	5-nitrobenzisoxazole	phylogeny (CD)[e]	R2−4/3D	9 of 247	72.0	10.0	n.d.	n.d.	11.46× (25 °C)	31
haloalkane dehalogenase | P59336	1-iodohexane	hybrid[f]	DhaA115	11 of 294	49.0	24.6	200× (60 °C)	0.31× (37 °C)	2.77× (37 °C)	45
halohydrin dehalogenase | Q93D82	*rac-p*-nitro-2-bromo-1-phenylethanol	hybrid[g]	HheC-H12	13 of 253	57.0	25.5	n.d.	n.d.	0.88× (30 °C)	9

Soluble Biocatalysts

enzyme | UniProt ID	substrate	method[j]	mutant code	mutations[a]	wild-type T_m [°C]	ΔT_m [°C][b]	expr. yield[c]	specific activity[c]	expr. host	ref
haloalkane dehalogenase | P59337	1,2-dibromoethane	phylogeny (ASR)	AncHLD2	69 of 317	53.6	21.9	4.8× (20 °C)	1.86× (37 °C)	*E. coli*	46
α-galactosidase | P06280	α-D-galactose	hybrid[h]	A348R/A368P/S405L	3 of 397	n.d.	n.d.	1.4× (37 °C)	2.00× (37 °C)	*H. gartleri*	18
acetylcholinesterase | P22303	acetylcholine	hybrid[i]	dAChE4	51 of 542	44.0	18.3	2000× (20 °C)	0.89× (25 °C)	*E. coli*	47

[a]Number of introduced mutations and total number of residues. [b]ΔT_m value of the mutant with respect to the wild-type enzyme. [c]Fold change in the specified property of the mutant relative to the wild-type enzyme. The temperature at which the given property was measured is given in parentheses. [d]n.d.: not determined. [e]Spiked Consensus Design, Directed Evolution. [f]FireProt: Rosetta, FoldX, Consensus Design. [g]FRESCO: Rosetta, FoldX, Disufide Bonds, MD. [h]SOLUBIS: TANGO, FoldX. [i]PROSS: Consensus Design, Rosetta. [j]CD - Consensus Design, ASR - Ancestral Sequence Reconstruction.

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

Table 2. Advantages and Disadvantages of Methods for the Computational Design of Stable and Soluble Biocatalysts

method	advantages	disadvantages
energy calculations	• granularity of predictions can be adjusted via different force fields	• high computational cost of accurate methods
	• web servers make predictions accessible to inexperienced users	• dependence on high-resolution structures
	• ever-growing structural databases together with advances in homology modeling and molecular threading	• trade-offs between stability and activity
	• high accuracy for the prediction of single-point mutations	• predicted stable mutants may not be expressible
		• epistatic effects are not well resolved
machine learning	• very rapid predictions	• lack of balanced high-quality experimental data
	• easy to implement and use	• limited accuracy of current models
	• wide applicability of features	• risk of overtraining
	• no need to understand all dependencies	
	• previously unknown patterns can be discovered	
phylogenetics[a]	• rich abundance of sequence data	• selection of relevant sequences is nontrivial
	• structures not needed for predictions	• profound understanding of the gene family is required
	• web servers available for certain tasks	• CD: epistatic effects are not considered
	• CD: simple and fast	• ASR: small data set size due to computational costs
	• CD: several filters are available to enhance prediction accuracies	• ASR: requires technical skills and experience
	• ASR: prediction of highly thermostable variants is achievable	
	• ASR: sequences of extremophilic proteins are not required	
	• ASR: sequence context and epistasis are maintained	

[a]CD, consensus design; ASR, ancestral sequence reconstruction.

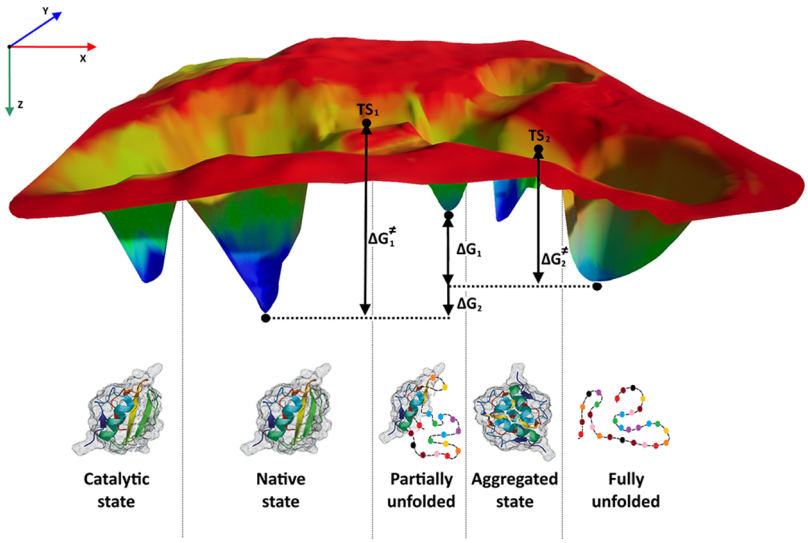


Figure 1. Simplified energy landscape with characteristic conformational states accessible from the native-state ensemble of a folded enzyme. Each point on the plane defined by the *X* axis and *Y* axis resembles a different conformation of the enzyme. The corresponding value on the *Z* axis is the free energy of folding, which has been color-coded to depict the spectrum from less probable high-energy states (red) to more probable low-energy states (blue). The catalytic state is readily accessible from the native-state ensemble but clearly separated by a free energy barrier. Catalysis based on a conformational selection model is assumed, which requires a distinct set of conformations prior to substrate binding and catalysis.[48] A reversible transition from the native state to a partially unfolded state via TS_1 is characterized by the free energy difference of folding ΔG_1 and its free energy barrier $\Delta G_1^{\ddagger}$. The partially unfolded state can also constitute the starting point for an irreversible unfolding transition via TS_2, leading to the fully unfolded state. Another irreversible pathway emanating from the partially unfolded state leads to an aggregated state, which is often characterized by the interactions of several biomolecules. ΔG_1 and ΔG_2 relate to thermodynamic stability, while $\Delta G_1^{\ddagger}$ and $\Delta G_2^{\ddagger}$ relate to kinetic stability.

This Perspective provides a thorough overview of contemporary data sets and computational protein redesign tools for enhancing enzyme stability or solubility. Preservation of enzymatic activity is of paramount importance in all protein engineering projects.[21,27] However, highly active and stable catalysts are evolutionarily disfavored because they could disrupt the host organism's homeostatic balance[28] or interfere with the cell's complicated metabolic regulatory networks.[29,30] Accordingly, several studies have indicated that most natural enzymes operate in a suboptimal regime,[21,28] leaving

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

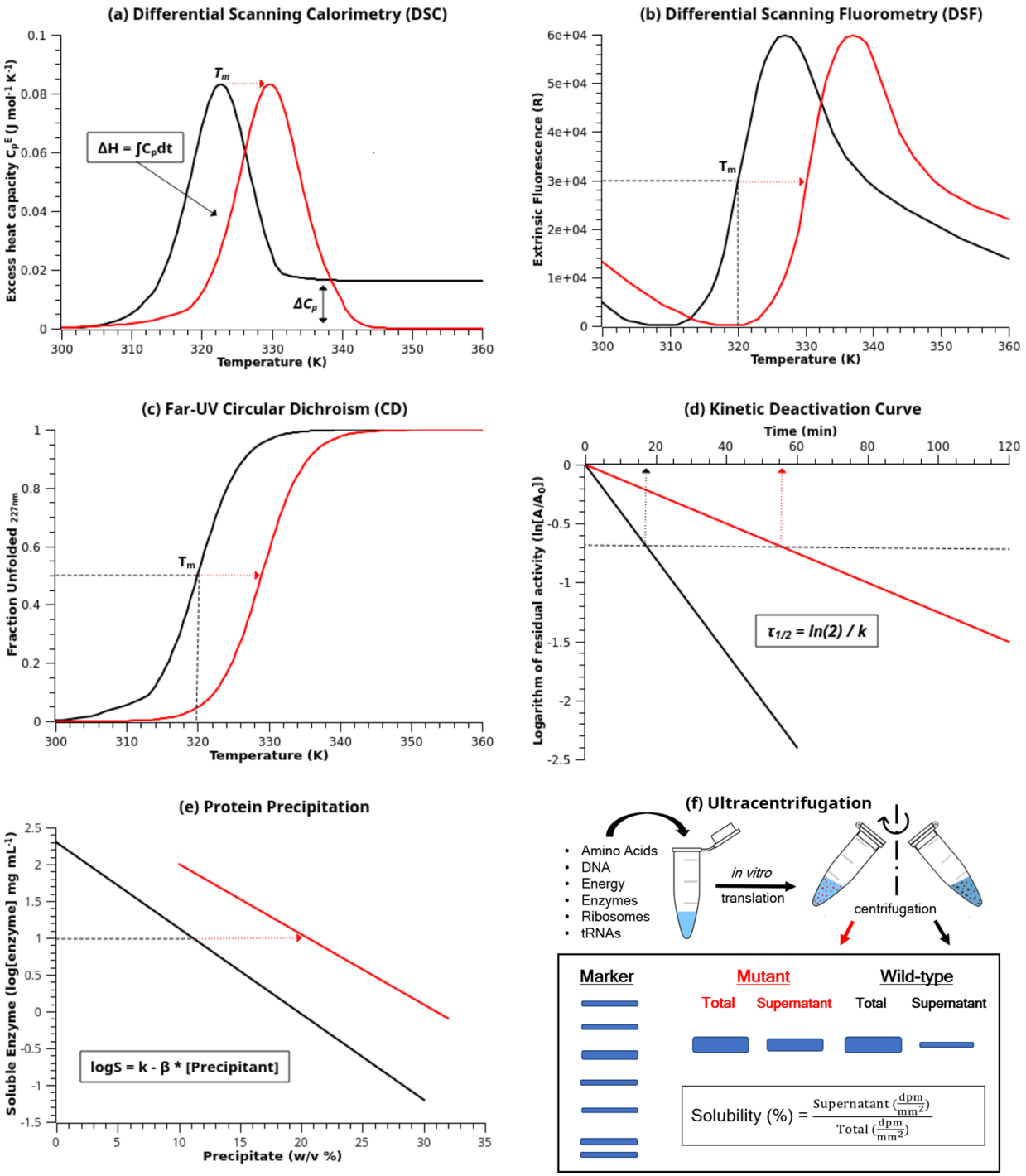


Figure 2. Representative experimental methods to quantify (a–d) protein stability and (e, f) solubility. Curves for a hypothetical wild-type enzyme (black) and an improved variant exhibiting higher stability or solubility (red) are shown. (a) Differential scanning calorimetry (DSC) curve. T_m is the midpoint of the transition, ΔC_p is the difference between the pre- and post-transition baselines, and ΔH is the area under the curve between the pre- and post-transition baselines. (b) Differential scanning fluorimetry (DSF) curve. Fluorescent dyes progressively bind to exposed hydrophobic regions of unfolding proteins, and the fluorescence signal is detected at different temperatures. T_m corresponds to the midpoint value of the stability curve. (c) Far-UV circular dichroism (CD) curve. Following the change of molar ellipticity at a specific wavelength over a wider temperature range monitors the change in secondary structure of an unfolding protein. The midpoint of the sigmoid curve is related to T_m of the protein. (d) Kinetic deactivation curve. For first-order deactivations, a plot of ln(activity) vs time yields a straight line with a slope of $-k$. The half-life can be calculated using the equation $\tau_{1/2} = \ln(2)/k$ and hence corresponds to the point $(\tau_{1/2}, -0.69)$ on the fitted line. (e) Protein precipitation experiment. The addition of a precipitant is negatively correlated with the solubility of the folded protein. The parameter β is protein-specific and characterizes the dependence of the solubility on the precipitant concentration. (f) Record from ultracentrifugation. In vitro translation followed by ultracentrifugation allows quantification of protein solubility independent of the proteostatic network of a living cell (the PURE system). The solubility percentage is calculated as the ratio of protein in the supernatant to the total protein measured by autoradiography.[60] Adapted with permission from ref 37. Copyright 2007 Elsevier.

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033–1054

considerable room for further optimization (Table 1). Unfortunately, activity enhancements often come at the cost of reduced enzyme stability. The protein redesign tools presented here offer ways to avoid this trade-off and also to solubilize the polypeptides, facilitating the purposeful adaptation of natural enzymes.[31] Here we outline the theoretical frameworks of methods commonly used to analyze protein stability and solubility. We also critically review the data sets and software tools available for predictive purposes. This Perspective strives to evaluate the tools from the perspective of users, who are typically interested in accuracy, reliability, user-friendliness, and the strengths and weaknesses of the underlying methods (Table 2). We also present a personal perspective on existing gaps in knowledge and propose possible directions for future development.

2. EXPERIMENTAL FRAMEWORK TO DETERMINE PROTEIN STABILITY AND SOLUBILITY

2.1. Experimental Determination of Protein Stability. Globular proteins are known to be marginally stable, with free energy differences between the folded and unfolded states (Figure 1) being as low as 5 kcal/mol.[32] Two key concepts in the analysis of protein stability are thermodynamic and kinetic stability.[30,33−35] Thermodynamic stability can be defined on the basis of equilibrium thermodynamics as the Gibbs free energy difference of folding (ΔG). Exact quantification of absolute ΔG values is difficult,[36] so most stability predictors and experimental procedures determine the relative change in free energy ($\Delta\Delta G$) upon mutation. A commonly used experimental quantity related to $\Delta\Delta G$ is the change in melting temperature (ΔT_m). The melting temperature, T_m, is defined as the temperature at which half of the sample is in the unfolded state, and it can be determined using biophysical techniques (Figure 2) such as circular dichroism spectroscopy (CD), fluorescence spectroscopy (FS), dynamic light scattering (DLS), differential scanning microcalorimetry (DSC), or differential scanning fluorimetry (DSF).[37] The chemical equivalent of T_m is the half-concentration ($C_{1/2}$), i.e., the concentration of denaturant at which half the sample exists in the unfolded state. Kinetic stability, on the other hand, is a time-dependent property that is quantified by the height of the free energy barrier of unfolding ($\Delta G^{\ddagger}$) separating distinct folding states (Figure 1). Predicting kinetic stability is challenging,[38] and experimentally determined biological half-lives ($t_{1/2}$) are preferred to theoretical estimates (Figure 2). The kinetic stability is a key determinant of an enzyme's functional competence[30] because it is related to the rate at which the protein's structure is irreversibly altered by proteolysis or aggregation.[29,39,40]

2.2. Experimental Determination of Protein Solubility. Protein solubility is a thermodynamic parameter defined as the concentration of folded protein in a saturated solution that is in equilibrium with a crystalline or amorphous solid phase under given conditions.[49] Two methods can be used to estimate protein solubility in aqueous solutions in vitro: (i) adding lyophilized protein to the solvent and (ii) concentrating a protein solution by ultrafiltration and then estimating the protein fractions in the supernatant and the pellet. Both methods require that the concentration of protein in solution is increased until saturation is reached, which can be difficult to achieve.[49] The difficulties of measuring protein solubility can be alleviated by adding an agent—a precipitant—to reduce the protein's solubility. Precipitants may be salts, organic solvents, or long-chain polymers.

The term solubility can also be applied to the in vivo observable that describes protein expression quantitatively (expression yield) or qualitatively (soluble/insoluble). Besides the previously given definition of solubility, these two observables critically depend on the expressibility of a given enzyme inside the cell.[50,51] As a polypeptide is synthesized in the ribosome, the emerging chain enters the cell's highly regulated proteostasis network,[29,35,52] which assists the enzyme to attain its native-state structure. Protein folding does not rely on the random scanning of all accessible conformational states but follows a deterministic folding pathway[53,54] or multiple folding pathways.[55,56] Changes in the protein sequence can perturb such folding pathways, frequently diminishing the expressibility and solubility of an enzyme with a negative impact on its aggregation propensity or the formation of inclusion bodies.[8,9,57,58] One high-throughput in vivo experimental screening assay to test for properly folded enzyme variants is the Split-GFP system.[59] Besides the calculation of the expression yields via the Bradford method and the quantification of mRNA levels of the cells, the PURE system[60] might be a valuable experimental platform to investigate determinants of protein solubility and folding under in vitro conditions (Figure 2).

3. THEORETICAL FRAMEWORK FOR THE DESIGN OF ROBUST PROTEINS

3.1. Principles of Methods Based on Energy Calculations. In silico design of protein stability based on energy calculations has taken a long way from fairly simple[61,62] to more accurate and versatile methods, facilitating reliable high-throughput predictions of thermodynamically and kinetically stable enzymes.[41,63] A force field is a collection of bonded and nonbonded interaction terms[64,65] that are related by a set of equations that can be used to estimate the potential energy of a molecular system.[66] For stability predictions, such potential energy functions can be applied to a protein's structure to assess the energetic changes caused by the mutations. The most accurate but also the most computationally expensive methods are free energy methods, which rely on molecular dynamics (MD) or Metropolis Monte Carlo simulations. Free energy perturbation has proven to be a potent and rigorous alchemical approach that generates the most meaningful stability predictions, but only for a limited number of mutations.[67] Less accurate but considerably more performant are end-point methods such as molecular mechanics generalized Born[68] or linear interaction energy.[69] These free energy methods require a high level of technical expertise and access to supercomputing facilities, which can be challenging for experimental groups. Over the last 20 years, simpler and simulation-independent stability predictors have been developed. A subdivision into three categories has been proposed, namely, (i) statistical effective energy functions (SEEFs), (ii) empirical effective energy functions (EEEFs), and (iii) physical effective energy functions (PEEFs).[70,71]

SEEFs are fast and can predict changes in stability over the entire sequence space of an average-sized enzyme in a matter of seconds.[72,73] They are derived from curated data sets of folded protein structures, which are projected into a number of stability descriptors. An effective potential can be extracted for every descriptor distribution, and these can be combined to create an overall energy function.[72,74] SEEFs do not explicitly

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

model physical molecular interactions, and the exact physical nature of statistical potentials remains obscure.[71] Consequently, overlapping and double counting of terms relating to the same causative interactions should be avoided.[70] EEEFs include both physical and statistical terms, which are carefully weighted and parametrized to match experimental data.[70,71] The thermodynamic data used in their derivation typically originate from mutational experiments conducted under standard conditions, which can be obtained from databases such as ProTherm.[75−77] EEEFs provide a reasonable compromise between computational cost and accuracy of the free energy function.[78] A major drawback of EEEFs and SEEFs is that their applicability is restricted to the environmental conditions under which the experimental data used for parametrization were acquired.[79,80] PEEFs are closely related to classical molecular mechanics force fields[81,82] and allow a fundamental analysis of molecular interactions.[66] PEEFs have more complex mathematical formalisms[71] and higher computational costs than EEEFs.[70] However, they are versatile, accurate, and capable of predicting behavior of the enzymes under nonstandard conditions, for instance at elevated temperature, nonphysiological pH, or nonstandard salinity.[83]

The accuracies of stability predictors based on such energy functions are still suboptimal[77,79,84−86] because of (i) imbalances in the force fields,[87,88] (ii) insufficient conformational sampling,[85,88] (iii) the occurrence of insoluble species,[8,9] and (iv) intrinsic problems with existing data sets (Table 2). The concept of free energy change upon mutation ($\Delta\Delta G$) was introduced for a fundamental analysis of the causative factors leading to these deficits. The computation of $\Delta\Delta G$ is based on a thermodynamic cycle (Figure 3), which requires modeling of

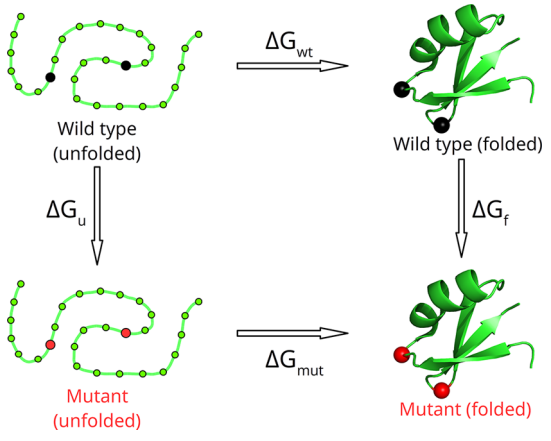


Figure 3. Thermodynamic cycle used to compute the free energy change upon mutation ($\Delta\Delta G$). $\Delta\Delta G$ is calculated according to the formula $\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt} = \Delta G_f - \Delta G_u$. For better illustration, the hypothetical folded and unfolded states of the wild type and a two-point mutant are shown. The respective substitution sites have been color-coded in black (wild type) and red (mutant). Adapted with permission from ref 69. Copyright 2012 Wiley.

the folded states of both the wild type and the mutant as well as their unfolded states.[36,67] Contemporary force fields describe enthalpic interactions reasonably well, although they are known to overestimate hydrophobicity and tend to favor nonpolar substitutions.[6,9,89] EEEFs and PEEFs generally underestimate the stability of buried polar residues because they overestimate the energetic cost of unsatisfied salt bridges and hydrogen bonds in the protein core.[58,90,91] The estimation of both conformational and solvent-related entropy is imprecise[9,92] because of the necessity of using computationally less expensive terms.[83] The inability of force field methods to account for entropy-driven contributions can be mitigated by using hybrid methods that incorporate complementary evolution-based approaches.[45,47,92,93] Moreover, most stability predictors have been parametrized using single-point-mutation data sets, resulting in higher prediction errors upon application to multiple-point mutants.[69,94] Whenever epistatic effects[20] are present between two or more individual mutations, force field predictions deviate from experimental results.

This shortcoming can be attributed to insufficient conformational sampling of the mutant's folded state, particularly when the introduced mutations induce large-scale backbone movements.[95] Tools based on EEEFs or PEEFs often apply rotamer libraries to fixed protein backbones, thereby reducing computational costs while providing comparable accuracies for the prediction of single-point mutations.[88] Multistate design[80,96] and flexible backbone sampling techniques[84,97−99] have partly alleviated the sampling problem for multiple-point substitutions by generating conformational ensembles and utilizing energetically more favorable conformations. Enzymes are intrinsically dynamic molecules and populate a high number of heterogeneous conformational substates[100] (Figure 1). Consequently, an adequate treatment of an enzyme's conformational plasticity[96,97] in the folded states of the wild type and mutant may be crucial for further advances of these methods. Structures obtained by X-ray crystallography do not essentially reflect the global energy minimum of the native state of an enzyme in its natural environment[101] and may therefore be nonideal starting points for stability predictions.[80,102] Besides the folded states, $\Delta\Delta G$ computations rely on sampling of the unfolded states of the wild type and the mutant. Simplifying and less realistic models (random coil or tetrapeptide) are frequently employed for explicit computations of the unfolded-state energies.[68,69] Generally, it is assumed that the free energy of the unfolded state does not change much upon mutation.[68,84]

The aforementioned explanations primarily relate to the prediction of thermodynamic stability. Not much work has been anticipated to predict kinetic stability, which can mostly be explained by the time-dependent nature[30] of this property and the time scales[103] assessable by energy-based methods. However, it is recognized that enhanced thermodynamic stability frequently goes hand in hand with enhanced kinetic stability.[41,45] One energy-based strategy to enhance the kinetic stability of an enzyme is to optimize solvent−solute interactions by introducing surface charges,[104] which can affect its expressibility.[105] The latter property may also be enhanced by computational linker design,[106] providing fusion enzymes with solubilizing protein tags.

3.2. Principles of Methods Based on Machine Learning. Machine learning is a field of computer science that allows computational systems to be constructed without being explicitly programmed. Statistical techniques are used to analyze training data sets and recognize patterns that might be difficult to detect given the limitations of human knowledge and cognitive abilities. Machine learning systems can be trained with or without supervision. In supervised approaches, the system is given a set of example inputs and the corresponding desired outputs in the form of labels indicating the correct classification of each input. Supervised approaches are suitable for training predictive systems, while unsupervised

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

approaches are more suitable for tasks involving data clustering. In recent years, machine learning has become one of the most common approaches for predicting the effects of mutations on protein stability[107−109] and solubility.[57,110] Machine learning does not require full understanding of the mechanistic principles underpinning the target function because they are modeled during the learning process. An important advantage of machine learning methods is that they are very flexible because any characteristic extracted from the data can be used as a feature if it improves the prediction accuracy, i.e., minimizes the prediction error (Table 2). Consequently, machine learning methods can reveal previously unrecognized patterns, relationships, and dependencies that are not considered in knowledge-based models. Moreover, machine learning is much less time-intensive than other methods because once a model has been constructed using the available data, predictions can be obtained almost instantaneously.

The reliability of machine learning approaches depends on the size and quality of the training data set. The weights representing the relative importance of the individual features and the relationships between them are based on experimental observations. Consequently, it is essential to use high-quality experimental data with high consistency when training and testing machine learning methods. The size and balance of the training data set must also be considered carefully. A modest data set with only a few hundred or a few thousand cases might be too small to identify useful descriptors during the learning process. Additionally, lower diversity of the training data set leads to a greater risk that the prediction tool will lose its ability to generalize. In such cases, the weights assigned to individual descriptors might be influenced by over-representation of some descriptors in the training data, while other descriptors that might be very important for general predictive ability could be omitted. Unbalanced training data sets with large differences in the numbers of cases representing individual categories could also lead to erroneous overestimations. For example, a training data set in which 80% of the mutations are destabilizing would allow the predictor to classify most mutations as destabilizing because of the prevalence of such mutations during the learning process. Methods like support vector machines and random forests are known to be more resistant to overfitting caused by unbalanced data sets,[111−113] while standard neural networks and decision trees are particularly sensitive to them. If the data set is too small to be balanced, the problem can be partially addressed by using cost-sensitive matrices,[114] which penalize the predictor more strictly for misclassifying mutations that are sparsely represented in the training data.

In parallel to the issue of the quality and availability of training data, one must address the problem of model validation. Ideally, the validation data set should be balanced and completely independent of the training set. In bioinformatics, it has become common to use *k*-fold cross-validation as a standard method for testing the performance of newly developed tools. This method entails randomly partitioning the original data set into *k* subsets. During the learning process, one of the *k* subsets is used for validation, while the remaining subsets are used as a training data set. This process is performed for each of the *k* subsets. The main reason for using cross-validation instead of splitting the data set into independent training and validation subsets is that the data set may be too small to support such splitting without harming the model's ability to learn the important predictive patterns. However, the combination of unbalanced data sets with the random aspect of *k*-fold cross-validation increases the risk of serious overestimation. Therefore, cross-validation is not a reliable method for measuring model accuracy when lower-quality data sets are used.[115] In conclusion, machine learning is a powerful approach that can reveal unknown interactions that are poorly defined in current force fields (Table 2). However, great care must be taken when constructing the training data set and during validation to avoid overfitting and overestimation of the results.

3.3. Principles of Methods Based on Phylogenetic Analysis. The two most widely used phylogeny-based approaches for stability engineering are consensus design (CD) and ancestral sequence reconstruction (ASR). Continuous cycles of variation and selection have created an enormous diversity of modern-day enzyme sequences that can be processed using phylogenetic techniques (Table 2). Over the last two decades, the advent of next-generation sequencing methods has revolutionized life science but has also introduced new challenges arising from the vast amounts of sequence data that are now available.[116] When phylogenetic analyses are performed, this results in a selection problem: one must carefully decide which sequences to include in any analysis. Identifying suitable homologous sequences to a given target can be particularly challenging. Local alignment algorithms such as the Basic Local Alignment Search Tool (BLAST)[117] offer reasonable accuracy at minimal computational cost. More complex and computationally demanding signature-based and profile-based search algorithms[118−120] have further extended the boundaries of homology detection[121] beyond the twilight zone.[122] The twilight zone is an alignment-length-dependent pairwise sequence identity range above which homologous sequences can reliably be distinguished. When pairwise sequence identities fall within or below this specific range, a large number of false negative sequences will get incorporated into multiple sequence alignments (MSAs). Great care is needed in the construction of biologically relevant MSAs from distantly related homologues. The treatment of nontrivial evolutionary artifacts such as indels, translocations, and inversions within the coding sequence can profoundly affect the quality of an MSA.[123,124] Progressive, iterative, and consistency-based alignment algorithms[125] exclusively consider sequence data and often introduce topological inconsistencies that require manual correction.[126] These deficiencies have been alleviated by incorporating complementary structural or evolutionary information, but such approaches can be computationally demanding.[25,126,127]

CD starts from a set of homologous protein sequences. A genuine MSA is generated using a small number (between a dozen and a few hundred) of homologous sequences, which permits the computation of the frequency distribution of every amino acid position in the alignment.[128] A user-specified conservation threshold is then used to distinguish between ambiguous and conserved "consensus" positions. The core assumption of this method is that the most frequent amino acid at a given position is more likely to be stabilizing.[128−133] It has been noted that high levels of sequence diversity in the MSA can interfere with the preservation of catalytic activity in consensus enzymes; this problem can be particularly acute when the MSA incorporates both prokaryotic and eukaryotic sequences.[129,134] However, the assumption of statistical independence is central to CD. Excessively homogeneous

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

MSAs may violate this assumption, introducing phylogenetic bias that hinders the discovery of more thermostable proteins.[133] The proportions of neutral and destabilizing consensus mutations have been estimated to be 10 and 40%, respectively, among all characterized variants produced using consensus design to date, suggesting a need for a more focused selection of substitution sites.[128,132] To this end, Sullivan et al.[129] discarded mutations of residues with high statistical correlations to other positions in the MSA, thereby increasing the proportion of identified stabilizing mutations to 90%. Vazquez-Figueroa et al.[135] adopted a different approach, successfully using structural information (e.g., the distance between a possible mutation and the active site, secondary structure data, and the total number of intramolecular contacts) to complement traditional CD predictions. Another example of an effective structure-based CD approach involved the analysis of molecular fluctuations based on crystallographic B-factors.[136] Important drawbacks of CD are its inability to account for epistatic interactions[137,138] and an apparent phylogenetic bias in cases where the MSA is dominated by a few subfamilies.[130,139]

ASR is a probabilistic method for inferring primordial enzymes and ancestral mutations, which have proven to be very effective for thermostability engineering.[43,44,46,140] ASR explores the deep evolutionary history of homologous sequences to reassemble a gene's evolutionary trajectory.[138,141] As a starting point, a phylogenetic gene tree can be inferred from a manually curated MSA and a suitable evolutionary model using either the maximum-likelihood method[142,143] or Bayesian inference.[144] In the simplest case, such statistical inference methods derive parameters from the given MSA for the selected empirical evolutionary model, which defines the underlying amino acid substitution process. Once the gene phylogeny has been established, ancestral sequences corresponding to specific nodes of the tree can be computed, synthesized, overexpressed, and characterized in vitro. In addition to the difficulty of identifying and aligning legitimate sequences,[124] a major challenge encountered in ASR is the computation of a plausible phylogenetic tree that adequately explains the evolutionary relationships of the given sequences. Homogenous evolutionary models assume that amino acid substitutions are homogeneously distributed over time and among sites and are therefore heavily oversimplified models of evolution.[145] Maximum-likelihood methods have been shown to systematically overestimate the thermodynamic stability of deeper ancestors,[140,146] so Bayesian inference methods have been recommended as alternatives to account for this bias. However, Bayesian inference computes ancestral sequences with considerably lower posterior probabilities, sometimes leading to the loss of the biological function.[147] It is not entirely clear why ASR is successful at identifying sequences with improved thermostability.[141] One hypothesis states that its success is an artifact of the ancestral inference methods and resembles a possible bias toward stabilizing consensus sequences.[140,146] Another plausible explanation is based on the thermophilic origin of primordial life.[148,149] Regardless of the reasons for its effectiveness, ASR is clearly a very robust and efficient method for identifying enzyme sequences with high thermodynamic stability and elevated expression yields (Table 2). Furthermore, increases in kinetic stability resulting in higher $\tau_{1/2}$ have frequently been reported for ancestral enzymes in comparison with their extant forms.[140,150] The sequence context is maintained in the resurrected ancestral enzymes, enabling the conservation of historic mutations causing functionally important epistatic effects.[20,137,138] The fundamental drawbacks of ASR are that users must have considerable methodological skill and a good level of knowledge about the targeted gene family.

4. DATA SETS AND SOFTWARE TOOLS FOR DESIGNING STABLE PROTEINS

4.1. Data Sets for Protein Stability. The accuracy and reliability of computational methods depends strongly on the size, structure, and quality of the chosen training and validation data sets. The primary source of validation data for protein stability is the ProTherm database.[75] ProTherm is the most extensive freely available database of thermodynamic parameters such as $\Delta\Delta G$, ΔT_m, and ΔC_p. It currently contains almost 26 000 entries representing both single- and multiple-point mutants of 740 unique proteins. Although ProTherm is the most common source of stability data, it suffers from high redundancy and serious inconsistencies. Particularly troubling are differences in the pH values at which the thermodynamic parameters were determined, missing values, redundancies, and strikingly even disagreements about the signs of $\Delta\Delta G$ values. ProTherm also neglects the existence of intermediate states.[57,107] To overcome the problems of the ProTherm database, the data must be filtered and manually repaired to construct a reliable data set.

Several subsets of the ProTherm database have been developed (Table S1) and used widely to train and validate new prediction tools. The most popular is the freely available PopMuSiC data set,[151] which contains 2648 mutations extracted from the ProTherm database. The data set is unbalanced because only 568 of its mutations are classified as stabilizing or neutral, while 2080 are classified as destabilizing. Furthermore, 755 of its 2648 mutations have reported $\Delta\Delta G$ values in the interval $\langle -0.5, 0.5\rangle$. Mutations with such $\Delta\Delta G$ values cannot be considered either stabilizing or destabilizing because the average experimental error in $\Delta\Delta G$ measurements is 0.48 kcal/mol.[152] Additionally, the data extracted from ProTherm are insufficiently diverse: around 20% of the PopMuSiC data set comes from a single protein, and 10 proteins (of 131 represented in the data set) account for half of the available data. Inspection of the data reveals that mutations to more hydrophobic residues located on the surface of the protein tend to be stabilizing, whereas mutations that increase the hydrophilicity in the protein core are usually destabilizing. Consequently, most computational tools are likely to identify mutations that increase surface hydrophobicity as stabilizing even though such designs often fail because of poor protein solubility.[58]

Some predictive tools use alternative data sets derived from ProTherm or PopMuSiC for training and validation. The most common benchmarking data set utilized for independent tests is S350,[151] which contains 90 stabilizing and 260 destabilizing mutations in 67 unique proteins. However, this data set is still small for comprehensive evaluation and unbalanced. The recently published PoPMuSiC^{sym} data set[153] tries to address these issues, containing 342 mutations inserted into 15 wild-type proteins and their inverse mutations inserted into the mutant proteins. A comparative study conducted using this data set showed a bias of the existing tools (Table S2) toward destabilizing mutations, as they performed significantly worse on the set of inverse mutations. Because of the overlaps of the mutations in training and validation data sets, the results of the

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

individual tools can be overestimated. Even the new derivatives of the ProTherm database do not solve the problems arising from the size and structure of the available data. Therefore, there is an urgent need for new experimental data, particularly on the side of stabilizing mutations. Moreover, it would be of immense help for the future development of predictive tools to proceed with the standardization of the stability data, e.g., a unified definition of $\Delta\Delta G$ as a subtraction of the ΔG values for the mutant and the wild type. FireProt DB, a new publicly available database collecting carefully curated protein stability data, is being established at https://loschmidt.chemi.muni.cz/fireprotdb/.

Until the new unbiased data sets arise, a regular accuracy measure considering only the number of correctly predicted mutations from the testing set is not suitable for validation of the predictive tools. For binary classification, the Matthews correlation coefficient (MCC) can be utilized, as it was designed as a balanced measure that is usable even for data sets with a significant difference in the sizes of individual classes.[113] Similarly, when binary predictions are utilized as a filtration step in the hybrid approaches, metrics like sensitivity, specificity, and precision might be useful. When numerical measures are considered, the linear correlation between the predicted and experimental values can be estimated with the use of the Pearson correlation coefficient (PCC) and the average error established as the root-mean-square error (RMSE). Finally, the bias of the computational tools can be estimated as the sum of $\Delta\Delta G$ for the direct and inverse mutations according to Thiltgen and Goldstein.[94] Critical evaluation of the existing tools using the S350 data set revealed that the PCC ranges from 0.29 to 0.81 with an average RMSE of about 1.3 kcal/mol (Table S5).

4.2. Software Tools for Predicting Protein Stability Based on Energy Calculations. Software tools relying on force field calculations are based on either modeling the physical bonds between atoms (PEEFs) or utilizing methods of mathematical statistics (SEEFs). Rosetta[88] is one of the most versatile software suites for macromolecular modeling and consists of several modules. Rosetta Design is a generally applicable module for protein design experiments that evaluates mutations and assigns them scores (in physically detached Rosetta energy units) reflecting their predicted stability. In its newest version, the Rosetta force field converts Rosetta energy units into well-interpretable $\Delta\Delta G$ values.[83] Furthermore, the stand-alone ddg_monomer module was built on top of Rosetta Design and is parametrized specifically for predicting $\Delta\Delta G$ values and protein stability. The Rosetta suite is also supplemented by a wide variety of usable force fields and protocols. The Eris software[154] is based on the Medusa force field and incorporates a side-chain packing algorithm and backbone relaxation method. A similar physical approach is adopted in the Concoord/Poisson−Boltzmann surface area (CC/PBSA) method,[155] which uses the GROMACS force field[156] to evaluate an ensemble of structures initially generated by the Concoord program.[157]

Unlike the previously mentioned methods, in which the values of the individual terms in the force field equation are evaluated by performing calculations based on Newtonian physics, some tools simply fit equations using values derived from the available data. One of the main representatives of this approach is PopMuSiC,[73] whose force field equation includes 13 physical and biochemical terms with values derived from databases of known protein structures. Similar approaches are used by other statistical and empirical tools, including FoldX[78] and Dmutant.[158] Another tool in this class is HotMuSiC,[159] which is based on PopMuSiC and was parametrized specifically for estimating ΔT_m, since the correlation coefficient between $\Delta\Delta G$ and ΔT_m is −0.7.[159] HotMuSiC makes predictions using five temperature-dependent potentials based exclusively on data extracted from mesostable and thermostable proteins.

While PEEFs provide generally more accurate predictions of the effect of mutations on protein stability, there is an apparent trade-off between predictive power and computational demands. In the majority of cases, SEEFs still perform fairly well compared with most machine learning methods and are orders of magnitude faster than PEEFs. Therefore, SEEFs seem to be an acceptable compromise between accuracy and time demands, especially when utilized as filters for prioritization of the mutations in hybrid workflows.

4.3. Software Tools for Predicting Protein Stability Based on Machine Learning. Machine learning methods do not require comprehensive knowledge of the physical forces governing protein structure; their predictions are based exclusively on the available data. The most popular machine learning tools are based on the support vector machines (e.g., EASE-MM,[107] MuStab,[108] I-Mutant,[160] and MuPro[161]) and random forest (e.g., ProMaya[162] and PROTS-RF[163]) methods, which are known to be comparatively resistant to overtraining even when used with unbalanced training data sets (Table S2). Neural networks are rarely used for protein stability engineering because of their high sensitivity to the quality and size of the training data set.

In recent years, several new machine learning approaches have been applied to diverse problems in the field of bioinformatics. Deep learning is used to predict the effects of mutations on human health in DANN[164] and to predict protein secondary structure in SSREDNs.[165] Unfortunately, like regular neural networks, deep learning methods are prone to overfitting because adding extra layers of abstraction increases their ability to model rare dependencies, resulting in a loss of generality. This shortcoming can be addressed by using regularization methods such as Ivakhnenko's unit pruning.[166,167] However, this does not eliminate problems arising from inadequate training data sets because deep learning has very stringent data requirements. Consequently, deep-learning-based tools such as TopologyNet[168] still have very limited applicability in predicting protein stability.

The robustness and accuracy of computational tools can be increased by combining several machine learning approaches into a single multiagent system, as in the case of MAESTRO.[169] In MAESTRO, neural networks are combined with support vector machines, multiple linear regression, and statistical potentials. The outputs of the individual methods are then averaged to provide users with a single consensus prediction. In such tools, machine learning can be used to train the arbiter that decides how to combine the outputs of the individual methods and their weights, balancing the relative strengths of each method when applied to the type of mutation under consideration. This approach is widely used in metapredictors.[58]

It is difficult to compare individual tools on the basis of the results presented in the publications where they were first reported because most of them were validated using different data sets. This can bias a tool's performance toward particular proteins or mutation types, causing its general prediction accuracy to be overestimated. Therefore, independent

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

comparative studies are needed. The critical evaluations reported by Kellogg et al.,[88] Potapov et al.,[77] and Khan and Vihinen[170] revealed that methods based on PEEF calculations systematically outperform tools relying only on machine learning techniques or statistical potentials in independent tests. Furthermore, machine learning methods tend to be more biased,[153,171] and their reported accuracies are overestimated as a result of overtraining. The PCC upper bound for the most commonly used stabilization data sets is about 0.8, and the lower bound of the RMSE is 1 kcal/mol.[172] The applicability of machine learning methods will increase with the size and diversity of the available data in the future.

4.4. Software Tools for Predicting Protein Stability Based on Phylogenetics. Phylogeny-based methods do not require knowledge of high-resolution protein structures; they can be applied to any protein with a known amino acid sequence and a sufficiently high number of sequence homologues. However, although phylogeny-based methods often improve some protein characteristics, the influence of individual mutations manifested during evolution is uncertain. About 50% of all mutations identified by CD are stabilizing, but some may affect protein solubility rather than stability.[131] CD-based methods are therefore frequently utilized as filters during core calculations of hybrid workflows or as components of predictive tools for hotspot identification.

CD is available in several bioinformatics suits (e.g., EMBOSS,[173] 3DM,[25] VectorNTI,[174] and HotSpot Wizard[175]). Although there are no stand-alone tools for CD, there are several for ASR, some using maximum-likelihood methods (e.g., RAxML,[176] FastML,[177] and Ancestors[178]) and others using Bayesian inference (e.g., HandAlign[179] and MrBayes[180]). A major limitation of these methods is that most of the tools require users to upload their own MSA and phylogenetic tree. Constructing these input data is the most important and demanding step of the entire process. To obtain reliable predictions, the initial set of homologue sequences must be filtered to identify a reasonably sized subset of biologically relevant sequences. At present, sets of homologous sequences obtained using BLAST,[117] profile-based methods such as position-specific iterated BLAST,[118] or hidden Markov models[120,181] must be manually curated to ensure reliable ancestral reconstructions.

4.5. Software Tools for Predicting Protein Stability Based on Hybrid Approaches. Hybrid methods make predictions by combining information from several fundamentally different approaches. They offer greater robustness and reliability than individual tools, allowing multiple-point mutants to be designed while reducing the risk of combining mutations with antagonistic effects. Consequently, several research groups are focusing on hybrid methods in their efforts to improve the rational design of thermostable proteins.

The Framework for Rapid Enzyme Stabilization by Computational Libraries (FRESCO)[93] is available as a set of individual tools and scripts, and its use requires a good knowledge of bioinformatics. FRESCO initially selects a pool of potentially stabilizing mutations (FoldX or Rosetta energy cutoff of −5 kJ/mol) and also filters out all residues in close proximity (<10 Å) to active sites. Disulfide bridges are designed by dynamic disulfide discovery using snapshots from MD simulations and subsequently evaluated using the set of geometric criteria. An energy criterion for the maximal molecular mechanics energy of the disulfide bond was also adopted. Furthermore, very short MD simulations predict changes in backbone flexibility upon mutation to remove designs with unreasonable features that are expected to destabilize the protein. About a hundred of the single-point mutants are then subjected to experimental validation to select mutations to be included in the combined multiple-point mutant. Experimental validation of individual mutations greatly reduces the risk of false positives and maximizes the stabilization effect but requires a substantial investment of time and effort.

FireProt[45,89] combines energy- and evolution-based approaches in a fully automated process for designing thermostable multiple-point mutants (Figure 4). FireProt integrates

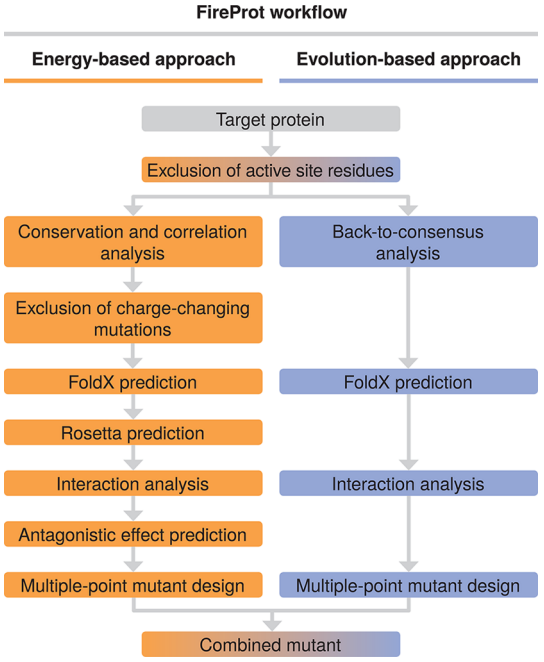


Figure 4. Workflow of the protein thermostabilization platform FireProt. The hybrid method combines evolutionary- and energy-based approaches and designs stable multiple-point mutants by fundamentally different methods.[45] The user is offered three different designs, two based solely on the energy- and evolution-based approaches and a third combining all of the identified mutations. FireProt has been made available as a fully automated and user-friendly web application[89] and is free of charge for academic users at http://loschmidt.chemi.muni.cz/fireprot.

16 computational tools, utilizing both sequence and structural information in the prediction process. When the energy-based approach is applied, information extracted from the protein sequences (e.g., lists of conserved and correlated residues) is used to exclude potentially deleterious mutations, while structural information is used to obtain estimated $\Delta\Delta G$ values with both FoldX and Rosetta. The second approach is based on back-to-consensus analysis followed by energy filtration using FoldX. Finally, a distance-based graph algorithm is used to create a multiple-point mutant by selecting the most favorable mutually nonconflicting mutations from the pool of all potentially stabilizing mutations. A stand-alone version of FireProt[45] has been implemented as an intuitive web-based application,[89] making this complex modeling workflow accessible to a wide user community. The automation of the whole procedure eliminates the need to select, install, and

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

evaluate tools, optimize their parameters, and interpret intermediate results.

Protein Repair One-Stop Shop (PROSS)[47] is another automated web-based protein stabilization platform. The PROSS workflow begins with a Rosetta design calculation in which the amino acids constituting the protein's active and binding sites are not eligible for mutation. A position-specific substitution matrix is analyzed to steer the design process away from amino acids that are rarely observed in the sequence homologues,[182] and Rosetta's computational mutation scanning tool[183] is used to scan the remaining pool of potential amino acid mutations. Finally, Rosetta's combinatorial sequence design tool is used to find an optimal combination of potentially stabilizing mutations, and an energy function is applied that favors amino acid identities on the basis of their frequency in the multiple-sequence alignment. This phylogeny-based biasing potential allows the designed variants to incorporate mutations found to be neutral or even slightly destabilizing in the Rosetta calculations,[35] which is desirable because some of these mutations might positively influence properties such as catalytic activity or protein solubility.

Hybrid methods represent a step forward in the prediction of protein stability because of their higher reliability at a decreased computational cost. These methods utilize evolution-based approaches as filters for removing potentially deleterious mutations in the conserved or correlated regions of the target protein. Furthermore, hybrid methods identify stabilizing mutations that would be missed by using only force field or phylogeny methods, as these two approaches are often complementary.[92] The increased robustness of the hybrid methods allows for a safer combination of single-point mutations into a multiple-point mutant. Hybrid methods can be further expanded by predictions of protein solubility or catalytic activity.

5. DATA SETS AND SOFTWARE TOOLS FOR THE DESIGN OF SOLUBLE PROTEINS

5.1. Protein Solubility Data Sets. Protein solubility, aggregation propensity, and expressibility are complex properties governed by several distinct biophysical and biological mechanisms. Progress in understanding these mechanisms depends on the availability of large, high-quality, diverse experimental data sets. In addition, the performance of prediction methods must be assessed with respect to the data used during their training. It is therefore important to recognize the strengths and limitations of the available experimental data sets on protein solubility and expressibility. To this end, this section presents a comprehensive review of the data sets available at the time of writing (Table S3).

5.1.1. Protein Solubility Data Sets Based on Full-Length Proteins. Data sources of this type contain information on the solubility of entire proteins produced in a specific expression system, either in vitro using a cell-free expression system or in vivo. Solubility can be determined by separating the liquid component of a sample by centrifugation or filtration and measuring the protein content in a solution, which is normalized by the protein content in the unseparated sample. The normalization removes the relationship between the solubility value and varying protein expression level. Alternatively, proteins may be simply classified as soluble or insoluble.

The Solubility Database of *E. coli* Proteins (eSOL)[60] contains experimentally measured solubilities for over 4000 *E. coli* proteins. The solubilities were determined by expressing the proteins using the PURE cell-free expression system[184] and using ultracentrifugation to measure their solubility as the ratio of the protein content in the supernatant to the total protein content of the sample. The limitations of eSOL are that only a moderate number of proteins are represented and that all of them originate from *E. coli*. In addition, in vitro cell-free expression systems cannot reproduce the post-transcriptional molecular processes that occur during protein expression in vivo. Interestingly, adding the three main cytosolic *E. coli* chaperones (TF, DnaKJE, and GroEL/GroES) to the in vitro cell-free expression system reduced the number of insoluble proteins from 788 to 24.[185]

TargetTrack,[186] formerly PepcDB or TargetDB, integrates vast amounts of information from the Protein Structure Initiative, a large-scale structure determination project. It contains data from over 900 000 protein crystallization trials using almost 300 000 unique protein sequences, which are termed targets. The database is not focused on solubility, but target proteins can be considered soluble if they reached a particular state in the experimental trial. We note that strictly speaking, this parameter reflects both the expressibility and the solubility of the target proteins. The major drawback of this database is the low quality of the annotations. No reason for failure is recorded for most of the unsuccessful crystallization attempts. Moreover, the experimental protocols are described in free text with no common structure. Therefore, it is difficult to automatically extract information about the expression systems. As a result, the application of strict rules to the target annotations dramatically reduces the number of usable records.

The Northeast Structural Consortium (NESG)[187] database is a subset of TargetTrack containing data on 9644 targets analyzed between 2001 and 2008. The NESG database contains explicit data on protein expression and solubility levels based on uniform protein production in *E. coli*. Two integer scores are recorded for each target, indicating the protein's level of expression and the recovery of the soluble fraction. The major drawback of this data set is that it was generated using outdated experimental methods; some of the targets could probably be solubilized using current techniques. Additionally, the database is too small to be used to train new machine learning algorithms. However, it can be used as a high-quality benchmark data set because its explicit experimental observations are more trustworthy than any other data in TargetTrack.

The Human Gene and Protein Database (HGPD)[188] contains expression and solubility measurements on over 9000 human proteins expressed in *E. coli*, a wheat-germ cell-free expression system, or *Brevibacillus*. The expression data were obtained using the Gateway system coupled with SDS-PAGE of C-terminal V5- or His-tagged proteins. Like the NESG data, these results originate from a uniform high-throughput protein production pipeline and thus constitute a consistent data set. Moreover, the HGPD provides information at the DNA level, so it includes codon composition data. Its major drawback is that it is focused exclusively on human proteins, so predictors constructed on the basis of its data will have an implicit bias toward human proteins.

AMYPdb[189] contains data on over 12 000 proteins belonging to amyloid precursor families as well as over 6000 generalized sequence patterns useful for assigning new sequences to poorly soluble amyloid precursor families. These data are derived from the literature and by UniProt

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

and PROSITE mining, so they are useful only as training data and for concept verification; they are not suitable for performance validation. This database has not been updated since its release in 2008.

5.1.2. Protein Solubility Data Sets Based on Protein Fragments. Fragment databases often describe properties of short peptides and their tendency to aggregate when exposed to solvent. This tendency does not necessarily correlate with the peptide's behavior when it is incorporated into a larger globular protein, in which case it may be protected by the formation of a hydrophobic core. Therefore, great care is necessary when using these databases as a basis for solubility prediction.

AmylHex and AmylFrag[190] are literature-based collections of nearly 200 short peptide sequences known to form amyloid fibrils. The major flaws of this database are its strong over-representation (51%) of point variants of a single amyloidogenic hexapeptide (STVIIE) and its low content of data on longer protein fragments.

WALTZ-DB[191] integrates data obtained from the literature and by in-house experimental verification on over 1000 hexapeptides tested for amyloidogenicity. As such, it is a unique resource containing primary experimental data. Of the peptides represented in the data set, 22% were found to be amyloidogenic and 78% were found to be non-amyloidogenic.

AmyLoad[192] combines data collected from WALTZ-DB, AmylHex, AmylFrag, the AGGRESCAN and TANGO validation data sets, and manual reviews of over 90 publications. The data set contains information on almost 1500 amyloidogenic and non-amyloidogenic protein fragments that have been characterized experimentally or computationally. About 30% of the fragments are considered amyloidogenic.

The Human Protein Atlas (HPA)[193] contains data on over 16 000 protein epitope signature tags (PrESTs) that were produced using a uniform *E. coli* production pipeline. PrESTs are substantial fragments of human proteins ranging from 20 to 150 amino acids. Their expression and solubility were measured and are quantified using integer scores ranging from 0 to 5.

The Curated Protein Aggregation Database (CPAD)[194] is an integrated database that includes data on almost 1700 amyloidogenic protein fragments and aggregation changes upon mutation. The fragments represented in the database include peptides with known and unknown structures, almost 100 verified aggregation-prone regions, and over 2300 aggregation rate changes upon mutation. The database represents a unique resource for validating the effect of mutations on protein aggregation. Unfortunately, it is poorly structured, and the data are not easily downloadable in a machine-friendly format.

5.1.3. Protein Solubility Data Sets Based on Mutants. The existing data sets containing information on protein variants with measured effects on protein solubility are very small and were constructed ad hoc by the authors of prediction software on the basis of literature data. Three representatives of this small group of solubility data sources are OptSolMut,[195] CamSol,[17] and PON-Sol.[57] OptSolMut contains binary solubility data on 137 protein variants, and the amounts of positive and negative samples are nearly balanced. CamSol contains data on 56 protein variants, of which only three are classified as reducing solubility. The PON-Sol data set contains 443 protein variants, of which 222 reportedly have no effect on protein solubility.

5.2. Software Tools for Predicting Protein Solubility. Unlike stability prediction tools, solubility prediction tools differ in their outputs rather than their fundamental operating principles. Almost all solubility prediction tools use some form of machine learning, ranging from simple statistical approaches to modern nonlinear methods such as support vector machines, random forests, or deep neural networks. The tools also use similar sets of features based on amino acid composition and physicochemical properties. Their outputs typically fall into one of three categories: (i) a single solubility score for the entire input sequence, (ii) a solubility profile with a unique score for each amino acid, or (iii) a score reflecting the effect of a specific mutation on solubility. All three outputs are expressed using arbitrary solubility scales with no physical meaning. The following section discusses the available predictive tools and their theoretical underpinnings and critically assesses their reliability (Table S4). Tools that predict single solubility scores for entire protein sequences are most useful for genomic projects because they can help prioritize protein sequences for laboratory production. Conversely, algorithms that provide quantitative scores over fixed-size sequence windows generate solubility profiles that can be used in the rational design of soluble proteins.

5.2.1. Software Tools for Protein Solubility Based on Primary Sequences. One of the first single-score solubility methods was the linear prediction model proposed by Wilkinson and Harrison,[196] which was later simplified by Davis and co-workers.[197] The revised model is surprisingly simple, using only two features (the approximate-charge average and turn-forming residue content) that both measure the relative abundance of specific amino acid types in the sequence. Despite its simplicity, the model can be useful for analyzing certain protein families. For example, it achieved a Spearman correlation coefficient of 0.54 and outperformed several newer tools in the same category (Table S4) when its predictions were compared to experimental data for 20 sequences closely related to a recently characterized haloalkane dehalogenase family.[4]

SOLpro,[198] PROSO II,[199] ccSOL omics,[200] and DeepSol[201] use the TargetTrack database as the source of training data. Consequently, although they use different features and machine learning models, they are quite similar to one another and have many shared strengths and weaknesses. Their most significant drawback is that they do not focus on any one expression system because it is hard to automatically extract expression system data from TargetTrack. Therefore, when validating these tools on a set of proteins expressed in a single expression system (e.g., *E. coli*), the observed prediction performance might differ significantly from that reported by the tools' creators. Published results suggest that DeepSol should have the highest prediction accuracy in general. However, this algorithm was created by using deep learning with a moderately sized training set and was validated against a data set representing protein families similar to those included in the training set. Moreover, although good performance is commonly claimed for tools based on TargetTrack, these claims have been strongly questioned.[199,201] In conclusion, the validation of these tools should be evaluated carefully, and further external validation using test sets independent of TargetTrack is needed. Unfortunately, the limitations of the TargetTrack database, from which solubility data can be

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

extracted only via automated parsing, impose a strong performance limit on any tool that relies heavily on its data.

Periscope[202] attempts to predict soluble protein expression in the periplasm of *E. coli* rather than the cytosol. Although it was trained on a small data set, it was validated against an independent set of proteins and thus might be useful for predicting periplasmatic expression in *E. coli*.

ESPRESSO[203] estimates protein expression and solubility in both cell-free (wheat germ) and in vivo (*E. coli*) expression systems. The system has three unique aspects. First, it is based on measured expression and solubility levels of human proteins from the HGPD and thus may be useful for production of human proteins in either of the two relevant expression systems. Second, it offers two types of prediction: property-based and motif-based. The former type resembles the predictions offered by the other machine learning tools in this category. In contrast, motif-based predictions identify positive and negative solubility motifs extracted from the training data. For each negative motif, ESPRESSO suggests point mutations that should turn the negative motif into a positive one, so the tool can be used for the rational design of soluble proteins. Third, ESPRESSO also uses DNA-level features in its property-based method. However, direct verification of its reported performance is currently complicated because the original training and testing data are unavailable.

SoluProt[204] is one of the latest additions to the family of solubility predictors. Its training set is based on the TargetTrack database,[186] which was carefully filtered to keep only targets expressed in *E. coli*. The negative and positive samples were balanced and equalized with respect to protein length. The independent validation set was derived from the NESG data set.[187] The current version of the tool uses a predictor based on a random forest regression model that employs 36 sequence-based features, including amino acid content, predicted disorder, α-helix and β-sheet content, sequence identity to the Protein Data Bank (PDB), and several aggregated physicochemical properties. SoluProt currently achieves a prediction accuracy of 58.2%, which exceeds that of other currently available tools, and is under active development. An intuitive web interface to the tool will soon be made available to the community at https://loschmidt.chemi.muni.cz/soluprot/.

5.2.2. Software Tools for Predicting Protein Solubility Based on Sequence Profiles. A solubility profile is an abstract construct in which each residue of a given protein sequence is assigned a solubility score that contextually describes its relative contribution to the solubility of the protein as a whole. The solubility scores within a profile may represent aggregation rates or values on an arbitrary scale with no corresponding physical units. In either case, the highest scores represent solubility hotspots. Predictions based on such profiles must be interpreted with care because they rest on a hidden assumption: most profile-predicting methods are trained with data on short linear and unstructured peptides (Table S4), so there is an inherent assumption that the protein of interest is also at least partially unstructured. Therefore, these tools lack specificity when applied to natively folded globular proteins, in which many predicted low-solubility (or aggregation-prone) segments are stabilized by the interactions that maintain the protein's secondary and tertiary structure. If the protein's structure or a reasonable homology model is available, it is possible to compensate for these problems by applying structural corrections.

There are several profile-based tools, most of which share at least some concepts and/or training data sets. Zyggregator[205,206] uses a model fitted to the measured aggregation rates of nearly 100 variants of 15 proteins mined from the scientific literature. AGGRESCAN[207] is based on data from a single-codon saturation mutagenesis study of amyloid β 42 protein, in which aggregation rates were measured for 20 protein variants. Because both methods are based on very small data sets, the authors took care to bolster their credibility by applying the models in several case studies.

TANGO,[208] WALTZ,[209] and PASTA[210] predict amyloid plaque formation propensity on the basis of data for short experimentally characterized peptides (mostly hexapeptides). TANGO is the most famous of these tools and has been cited hundreds of times. However, the models used by the newer tools WALTZ and PASTA were inferred from larger experimental data sets, so they are claimed to outperform TANGO. A common concern is that the data sets of amyloidogenic peptides are unbalanced, containing too few non-amyloidogenic fragments (Table S3), which limits the generalizability of predictions obtained with these tools.

BETASCAN,[211] FoldAmyloid,[212] ZipperDB,[213] and ArchCandy[214] learn from experimentally determined structures of amyloidogenic proteins and apply the discovered general concepts at the sequence level. BETASCAN calculates likelihood scores for potential β-strands and strand pairs in sequences based on correlations observed in parallel β-sheets of experimental structures. FoldAmyloid uses the number of contacts per residue and statistics on hydrogen bonds in nearly 4000 PDB structures. In ZipperDB, the input protein is threaded onto a template cross-β spine structure, and the relative threading energy is used to predict amyloidogenicity. ArchCandy evaluates whether a protein segment can fold into β-arcade structures, which are often disease-related, and uses an empirical scoring function to evaluate interactions that disrupt β-arcade formation. These structure-based tools are expected to be inherently more specific than sensitive because structure-derived criteria tend to be relatively strict. When a high sensitivity is required and a structure is available, methods based on short peptides are expected to be more sensitive than structure-based alternatives. It is possible to compensate for false positives by checking the tool's output against known structures.

Because individual solubility prediction tools have different strengths and weaknesses, efforts have been made to create consensus-based methods that combine multiple tools to mitigate against the weaknesses of individual tools while preserving their strengths. The advantages of consensus methods have been proven both theoretically[215] and empirically.[216] Both AmylPred2[217] and MetAmyl[218] implement 11 individual methods, including AGGRESCAN, TANGO, and WALTZ. Although the primary publication on AmylPred2 claims superior performance to all of the individual methods, these results should be treated with care because the consensus threshold was validated using the entire data set chosen by the developers. Consequently, there was no independent validation set, and the claimed performance is very likely to be overestimated. MetAmyl uses a specially developed peptide set derived from the WALTZ data set to establish a logistic regression model that integrates the outputs of the individual tools. An evaluation using the AmylPred2 data

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

set indicated that MetAmyl outperformed AmylPred2 despite having been optimized with a different data set.[218] This strongly suggests that MetAmyl performs better than AmylPred2 in general.

5.2.3. Software Tools for Protein Solubility Based on Mutations. While the profile-based tools discussed above can be used to design solubilizing mutations, the methods described in this section are tailored for this purpose and therefore are easier to use. Importantly, most of the methods discussed here require a protein structure as an additional input (Table S4).

OptSolMut[195] uses the concepts from computational geometry to define a scoring function reflecting the changes in solubility due to mutations. The scoring function was optimized using linear programming on the basis of a set of protein variants extracted from the literature. The reported 81% overall accuracy should be taken with care, as the training set was small and the model might not generalize well. In contrast to other tools in this section, OptSolMut is able to predict the effect of multiple-point mutations.

Several tools for predicting the effect of mutations on solubility have been developed from tools for predicting solubility profiles. For example, CamSol,[17] AGGRESCAN3D,[219] SolubiS,[220,221] and SODA[110] are based on the previously published profile-based methods Zyggregator, AGGRESCAN, TANGO, and PASTA, respectively. The workflows of these tools are all very similar: first a solubility profile is predicted, then a correction based on knowledge of the protein's structure is applied, and finally solubility hotspots are identified and specific mutations targeting low-solubility regions are suggested. CamSol, AGGRESCAN3D, and SODA use structural corrections to refine the predicted solubility profiles by averaging physicochemical properties over residues proximal in three-dimensional space or on the basis of solvent exposure of individual residues. SolubiS uses free energy calculations based on the FoldX force field to avoid potentially destabilizing mutations in aggregation-prone regions and can thus be classified as a hybrid method (Figure 5). CamSol and SODA can make predictions even without structural data.

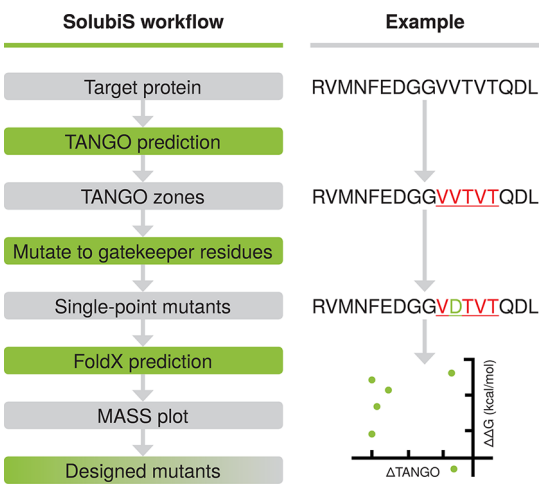


Figure 5. Workflow of the protein solubilization platform SolubiS. The platform uses free energy calculations performed with FoldX to avoid potentially destabilizing mutations in aggregation-prone regions identified by TANGO. The results are presented in form of a mutant aggregation and stability spectrum plot.[220] The web server is free of charge for academic users at http://solubis.switchlab.org/.

However, this necessarily eliminates the potential to exploit structure-based corrections and thus tends to reduce the prediction accuracy. The main issue with all of these tools is in the difficulty of validating their output. The data sets available for both training and testing are small, and they have only been validated using data for a small number of experimentally characterized protein variants.

PON-Sol[57] uses a machine learning algorithm designed from scratch for solubility prediction of protein variants from protein sequences without structure-based corrections. The reported accuracy of this three-class classification method is 43%. The training data set was rather limited, representing a few tens of proteins.

6. PERSPECTIVES

Protein Structures from Cryoelectron Microscopy and Hardware-Accelerated Calculations. Access to large and diverse data sets is a key factor in the development of new predictive methods and tools. Therefore, the applicability of force field methods to stability prediction is limited by the availability of relevant tertiary structures. At present, the PDB contains over 77 000 unique protein structures, and around 10 000 new structures are added each year. Advances in structural genomics will provide access to an additional large pool of protein structures, including previously unattainable structures of membrane-bound proteins that will be solved by cryogenic electron microscopy. A tertiary structure of a biomolecule of interest is typically required for predictions employing energy calculations. The general applicability of these methods is also hindered by their computational cost, which imposes a trade-off between accuracy and throughput. The most precise alchemical free energy calculations rely on MD simulations in which both the solute and solvent are modeled atomistically. Such calculations are too costly to be used in systematic mutagenesis campaigns with currently available computational resources. However, they could be selectively used to design mutations whose effects are poorly predicted by otherwise reliable Rosetta or FoldX calculations (e.g., substitutions that change the charge at the protein surface). Their high computational cost could be alleviated by adopting computing employing graphics processing units (GPUs), which has not yet been implemented in a number of software tools. Wider use of GPUs will enable predictions of structures and complexes that are currently too large to process using computationally demanding physical force fields.

Consistent and Balanced Stability Data Sets Are Urgently Needed. Machine learning techniques are faster than force field methods and less dependent on the availability of tertiary structures because many features used in machine-learning-based predictors can be extracted from primary sequences. However, machine learning methods are very sensitive to the size and quality of the experimental data sets available for training and validation. At present, there is a serious lack of reliable experimental data suitable for use in protein stabilization efforts. The only available database—ProTherm—is burdened by errors and contains data on fewer than 2000 single-point mutations after rigorous filtering. This number is insufficient to train reliable machine learning systems without introducing a risk of overfitting. Moreover, the ProTherm database was most recently updated in February 2013, and several protein stabilization projects have been conducted since then. Systematic mining of the scientific literature to incorporate the stability data from these projects

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

could provide valuable data resources for the training and validation of stability predictors. A new database, FireProt DB, is being established for this purpose at https://loschmidt.chemi.muni.cz/fireprotdb/. The research community should make an effort to establish validation procedures to assess the quality of predictions of protein stability and solubility. This could be done by releasing design challenges, but not experimental data, as in the well-known Critical Assessment of Protein Structure Prediction. Such a community-wide assessment is one of the most efficient ways to compare individual tools.

The Shift from Scores to Profiles and Specific Mutations in Solubility Predictions. The problem of unbalanced data sets also affects solubility predictors based on machine learning, especially those that use *k*-mer content and physicochemical properties as dominant features. The imbalance of the training data sets containing a larger number of negative samples and low diversity of protein structures limit the predictive performance and generalizability to unseen protein families. Over the short history of solubility prediction, there has been a significant and positive shift away from methods that provide single solubility scores toward alternatives that offer more detailed solubility profile predictions and even suggest mutations predicted to enhance protein solubility. However, this trend also poses problems because the quantity of relevant high-quality data decreases as the detail of the predictions increases. For single solubility score predictions, the TargetTrack database (which contains information on tens of thousands of samples) is large enough to support the development of machine learning models. For solubility profile predictions, the number of relevant samples decreases to hundreds or thousands, most of which are amyloidogenic peptides. Matters are worse still for attempts to predict the effect of mutations on protein solubility; in this case, the amount of relevant experimental data is arguably below the minimum needed to make adequate predictions. Therefore, mathematical models developed by machine learning frequently incorporate empirical components such as structure-based corrections. A mechanistic understanding of protein solubility justified by robust statistical analysis can only be expected once larger sets of experimental data become available.

High-Throughput Techniques for Highly Consistent Data Sets. We envisage that the lack of appropriate data for solubility prediction will be partially addressed by studies using novel high-throughput characterization techniques such as droplet microfluidics, fluorescence-activated cell sorting, fluorescence resonance energy transfer, deep sequencing, and deep mutational scanning. Experiments should be conducted under strictly controlled conditions to produce robust data and could employ one or more of the biomolecular and cellular systems that have recently been developed to monitor protein solubility and aggregation inside living cells. Additional high-quality data could be obtained from projects conducted by companies and other private organizations. The data generated under defined conditions need to be properly annotated, for example to report vectors, host organisms, buffers, laboratory conditions, and procedures used for protein expression, purification, and characterization. Proper controls should always be included and the statistics reported to allow a quantitative assessment of data variation. Collected data should be structured to allow processing using computers, which is for example not the case for the largest database of protein solubility data, TargetTrack. The data should be curated and stored in publicly accessible databases following the FAIR principles: Findable, Accessible, Interoperable, and Reusable. New data sets will enable the use of more sophisticated and data-intensive methods such as deep learning and allow proper external validation to be performed. Moreover, because solubility depends largely on the properties of the protein's surface, corrections based on protein structure and the inclusion of structural data in predictive tools could improve the prediction accuracy. Enhanced-sampling MD simulations of simplified molecular systems might reveal residue interactions that are important for protein folding, while advances in homology modeling and threading can complement sequence-based descriptors by providing structural information at a reasonable computational cost.

Robust Scaffolds for Directed Evolution by Phylogenetic Analyses. Whereas force field and machine learning methods are limited by a lack of data, the problem for phylogenetic approaches is different: high-throughput sequencing has made vast numbers of sequences available, allowing evolutionary analyses to be performed for the vast majority of protein families. The genomes of organisms living under extreme conditions are also becoming available, providing essential information for wider use of CD. This rapid expansion of the accessible sequence space has a downside for the ASR method, which can only use a limited number of homologous sequences for reconstruction. Therefore, large pools of potential homologues make sequence selection a challenging task. Homologue selection can be guided by annotation ontologies (e.g., molecular function, cellular component, and biological process) and other information from bioinformatics and biophysical databases. Furthermore, with increasing numbers of solved protein structures, structure-guided MSAs may displace sequence-based alternatives, and ASR may be more commonly used to generate robust scaffolds for directed evolution campaigns and de novo enzyme design. The degree of uncertainty in ASR increases the further back we go in evolutionary history. Therefore, the reliability of inference methods should be increased to more accurately predict folded, stable, and soluble ancestral proteins.

Addressing Stability−Activity Trade-Offs Using Metadata and Negative and Multistate Designs. The predictive power of computational methods has improved in recent years, with a positive impact mainly in the area of protein stabilization. A very challenging but important task is to predict thermodynamic as well as kinetic stability. There are several spectacular examples illustrating the improvement in kinetic stability by only a few mutations, but to the best of our knowledge, methods specifically targeting kinetic stability have not been developed. Connecting the design of kinetic stability with solubility within a single method could be particularly powerful. Stability−activity trade-offs are intrinsic to protein structures. Buried polar catalytic residues are suboptimal with respect to protein stability, and structural optimization of these functionally relevant regions is likely to also affect the biological activity. Mutations that stabilize regions whose conformational dynamics are important for enzyme activity can similarly be expected to negatively affect the catalytic performance. The incorporation of metadata and smart filters into engineering workflows will help preserve protein activity by enabling the identification of structurally and functionally important residues, which should be systematically excluded from mutagenesis. The incorporation of such negative designs

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

will suppress misfolding and protein aggregation. Furthermore, prediction accuracy is sometimes compromised by using a single structure in calculations. Increasing computational power and the use of GPU hardware will allow the adoption of multistate designs. Extracting multiple representative conformations and averaging results over the ensemble will further improve the robustness and accuracy of predictions.

Enhancing Accuracy by Using Metapredictors, Consensual Force Fields, and Hybrid Methods. There is a clear trend toward combining multiple fundamentally different methods within single predictors, leading to the development of metapredictors, consensual force fields, and hybrid methods. Hybrid methods offer several advantages: (i) even a simple majority voting approach over several methods yields better results than any individual method, each of which has its own strengths and weaknesses; (ii) smart filtering out of "untouchable" residues reduces the time required for calculations to a degree that permits very thorough analysis of the designable residues; (iii) the phylogenetic components of hybrid methods can incorporate both positive and negative design elements; and (iv) the availability of reliable predictions will enable the combination of substitutions to create multiple-point mutants without risking the introduction of destabilizing or antagonistic effects. Hybrid methods represent a natural step forward in the rapidly evolving field of protein stability prediction because improvements in machine learning models are limited by the availability of adequate data sets, while the application of advanced force field methods is restrained by their computational cost. It was recently demonstrated that combining phylogenetic methods and atomistic force fields can effectively optimize stability−activity trade-offs. We also envisage the future enrichment of protein stabilization methods addressing both thermodynamic and kinetic stability with tools for predicting protein solubility, aggregation propensity, and expressibility, eventually yielding all-in-one software suites capable of designing "ideal" biocatalysts.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscatal.8b03613.

Data sets for prediction of protein stability (Table S1); software tools for prediction of protein stability (Table S2); data sets for prediction of protein solubility (Table S3); software tools for prediction of protein solubility (Table S4); comparison of the existing tools with the S350 data set (Table S5) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jiri@chemi.muni.cz.

ORCID

Jiri Damborsky: 0000-0002-7848-8216

Author Contributions

‖M.M., H.K., and J.H. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the Czech Ministry of Education (LM2015051, LM2015047, LM2015055, CZ.02.1.01/0.0/0.0/16_013/0001761, CZ.02.1.01/0.0/0.0/16_019/0000868, and CZ.02.1.01/0.0/0.0/16_026/0008451) and the European Commission (720776 and 722610) for financial support. H.K. is the MSCA ITN ES-Cat Research Fellow supported by the European Commission (722610). The work of M.M. and J.H. was supported by the ICT Tools, Methods and Technologies for Smart Cities Project of the Brno University of Technology (FIT-S-17-3964).

■ REFERENCES

(1) Choi, J.-M.; Han, S.-S.; Kim, H.-S. Industrial Applications of Enzyme Biocatalysis: Current Status and Future Aspects. *Biotechnol. Adv.* **2015**, *33*, 1443−1454.

(2) Mitchell, A. C.; Briquez, P. S.; Hubbell, J. A.; Cochran, J. R. Engineering Growth Factors for Regenerative Medicine Applications. *Acta Biomater.* **2016**, *30*, 1−12.

(3) Dvořák, P.; Nikel, P. I.; Damborský, J.; de Lorenzo, V. Bioremediation 3.0: Engineering Pollutant-Removing Bacteria in the Times of Systemic Biology. *Biotechnol. Adv.* **2017**, *35*, 845−866.

(4) Vanacek, P.; Sebestova, E.; Babkova, P.; Bidmanova, S.; Daniel, L.; Dvorak, P.; Stepankova, V.; Chaloupkova, R.; Brezovsky, J.; Prokop, Z.; Damborsky, J. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catal.* **2018**, *8*, 2402−2412.

(5) Bornscheuer, U. T.; Huisman, G. W.; Kazlauskas, R. J.; Lutz, S.; Moore, J. C.; Robins, K. Engineering the Third Wave of Biocatalysis. *Nature* **2012**, *485*, 185−194.

(6) Tokuriki, N.; Stricher, F.; Serrano, L.; Tawfik, D. S. How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **2008**, *4*, e1000002.

(7) Dellus-Gur, E.; Toth-Petroczy, A.; Elias, M.; Tawfik, D. S. What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-Offs. *J. Mol. Biol.* **2013**, *425*, 2609−2621.

(8) Johansson, K. E.; Johansen, N. T.; Christensen, S.; Horowitz, S.; Bardwell, J. C. A.; Olsen, J. G.; Willemoës, M.; Lindorff-Larsen, K.; Ferkinghoff-Borg, J.; Hamelryck, T.; Winther, J. R. Computational Redesign of Thioredoxin Is Hypersensitive toward Minor Conformational Changes in the Backbone Template. *J. Mol. Biol.* **2016**, *428*, 4361−4377.

(9) Arabnejad, H.; Dal Lago, M.; Jekel, P. A.; Floor, R. J.; Thunnissen, A.-M. W. H.; Terwisscha van Scheltinga, A. C.; Wijma, H. J.; Janssen, D. B. A Robust Cosolvent-Compatible Halohydrin Dehalogenase by Computational Library Design. *Protein Eng., Des. Sel.* **2017**, *30*, 175−189.

(10) Wyganowski, K. T.; Kaltenbach, M.; Tokuriki, N. GroEL/ES Buffering and Compensatory Mutations Promote Protein Evolution by Stabilizing Folding Intermediates. *J. Mol. Biol.* **2013**, *425*, 3403−3414.

(11) Lawrence, P. B.; Gavrilov, Y.; Matthews, S. S.; Langlois, M. I.; Shental-Bechor, D.; Greenblatt, H. M.; Pandey, B. K.; Smith, M. S.; Paxman, R.; Torgerson, C. D.; Merrell, J. P.; Ritz, C. C.; Prigozhin, M. B.; Levy, Y.; Price, J. L. Criteria for Selecting PEGylation Sites on Proteins for Higher Thermodynamic and Proteolytic Stability. *J. Am. Chem. Soc.* **2014**, *136*, 17547−17560.

(12) Rueda, N.; Dos Santos, J. C. S.; Ortiz, C.; Torres, R.; Barbosa, O.; Rodrigues, R. C.; Berenguer-Murcia, Á.; Fernandez-Lafuente, R. Chemical Modification in the Design of Immobilized Enzyme Biocatalysts: Drawbacks and Opportunities. *Chem. Rec.* **2016**, *16*, 1436−1455.

(13) Stepankova, V.; Bidmanova, S.; Koudelakova, T.; Prokop, Z.; Chaloupkova, R.; Damborsky, J. Strategies for Stabilization of Enzymes in Organic Solvents. *ACS Catal.* **2013**, *3*, 2823−2836.

(14) Butt, T. R.; Edavettal, S. C.; Hall, J. P.; Mattern, M. R. SUMO Fusion Technology for Difficult-to-Express Proteins. *Protein Expression Purif.* **2005**, *43*, 1−9.

(15) LaVallie, E. R.; DiBlasio, E. A.; Kovacic, S.; Grant, K. L.; Schendel, P. F.; McCoy, J. M. A Thioredoxin Gene Fusion Expression

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

System That Circumvents Inclusion Body Formation in the *E. coli* Cytoplasm. *Nat. Biotechnol.* **1993**, *11*, 187−193.
(16) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein Stability Promotes Evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 5869−5874.
(17) Sormanni, P.; Aprile, F. A.; Vendruscolo, M. The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *J. Mol. Biol.* **2015**, *427*, 478−490.
(18) Ganesan, A.; Siekierska, A.; Beerten, J.; Brams, M.; Van Durme, J.; De Baets, G.; Van der Kant, R.; Gallardo, R.; Ramakers, M.; Langenberg, T.; Wilkinson, H.; De Smet, F.; Ulens, C.; Rousseau, F.; Schymkowitz, J. Structural Hot Spots for the Solubility of Globular Proteins. *Nat. Commun.* **2016**, *7*, 10816.
(19) Zeymer, C.; Hilvert, D. Directed Evolution of Protein Catalysts. *Annu. Rev. Biochem.* **2018**, *87*, 131−157.
(20) Starr, T. N.; Thornton, J. W. Epistasis in Protein Evolution. *Protein Sci.* **2016**, *25*, 1204−1218.
(21) Goldsmith, M.; Tawfik, D. S. Enzyme Engineering: Reaching the Maximal Catalytic Efficiency Peak. *Curr. Opin. Struct. Biol.* **2017**, *47*, 140−150.
(22) Currin, A.; Swainston, N.; Day, P. J.; Kell, D. B. Synthetic Biology for the Directed Evolution of Protein Biocatalysts: Navigating Sequence Space Intelligently. *Chem. Soc. Rev.* **2015**, *44*, 1172−1239.
(23) Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.; Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; Arrowsmith, C. H.; Baker, D. Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing. *Science* **2017**, *357*, 168−175.
(24) Sumbalova, L.; Stourac, J.; Martinek, T.; Bednar, D.; Damborsky, J. HotSpot Wizard 3.0: Web Server for Automated Design of Mutations and Smart Libraries Based on Sequence Input Information. *Nucleic Acids Res.* **2018**, *46*, W356−W362.
(25) Kuipers, R. K.; Joosten, H.-J.; van Berkel, W. J. H.; Leferink, N. G. H.; Rooijen, E.; Ittmann, E.; van Zimmeren, F.; Jochens, H.; Bornscheuer, U.; Vriend, G.; Martins dos Santos, V. A. P.; Schaap, P. J. 3DM: Systematic Analysis of Heterogeneous Superfamily Data to Discover Protein Functionalities. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2101−2113.
(26) Reetz, M. T.; Carballeira, J. D. Iterative Saturation Mutagenesis (ISM) for Rapid Directed Evolution of Functional Enzymes. *Nat. Protoc.* **2007**, *2*, 891−903.
(27) Liskova, V.; Stepankova, V.; Bednar, D.; Brezovsky, J.; Prokop, Z.; Chaloupkova, R.; Damborsky, J. Different Structural Origins of the Enantioselectivity of Haloalkane Dehalogenases toward Linear *β*-Haloalkanes: Open-Solvated versus Occluded-Desolvated Active Sites. *Angew. Chem., Int. Ed.* **2017**, *56*, 4719−4723.
(28) Bar-Even, A.; Noor, E.; Savir, Y.; Liebermeister, W.; Davidi, D.; Tawfik, D. S.; Milo, R. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **2011**, *50*, 4402−4410.
(29) Balchin, D.; Hayer-Hartl, M.; Hartl, F. U. In Vivo Aspects of Protein Folding and Quality Control. *Science* **2016**, *353*, aac4354.
(30) Colón, W.; Church, J.; Sen, J.; Thibeault, J.; Trasatti, H.; Xia, K. Biological Roles of Protein Kinetic Stability. *Biochemistry* **2017**, *56*, 6179−6186.
(31) Khersonsky, O.; Kiss, G.; Röthlisberger, D.; Dym, O.; Albeck, S.; Houk, K. N.; Baker, D.; Tawfik, D. S. Bridging the Gaps in Design Methodologies by Evolutionary Optimization of the Stability and Proficiency of Designed Kemp Eliminase KE59. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 10358−10363.
(32) Taverna, D. M.; Goldstein, R. A. Why Are Proteins Marginally Stable? *Proteins: Struct., Funct., Genet.* **2002**, *46*, 105−109.
(33) Sanchez-Ruiz, J. M. Protein Kinetic Stability. *Biophys. Chem.* **2010**, *148*, 1−15.
(34) Bommarius, A. S.; Paye, M. F. Stabilizing Biocatalysts. *Chem. Soc. Rev.* **2013**, *42*, 6534−6565.
(35) Goldenzweig, A.; Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. *Annu. Rev. Biochem.* **2018**, *87*, 105−129.
(36) Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014**, *10*, 2632−2647.
(37) Polizzi, K. M.; Bommarius, A. S.; Broering, J. M.; Chaparro-Riggers, J. F. Stability of Biocatalysts. *Curr. Opin. Chem. Biol.* **2007**, *11*, 220−225.
(38) Buck, P. M.; Kumar, S.; Wang, X.; Agrawal, N. J.; Trout, B. L.; Singh, S. K. Computational Methods To Predict Therapeutic Protein Aggregation. *Methods Mol. Biol.* **2012**, *899*, 425−451.
(39) Jaswal, S. S.; Sohl, J. L.; Davis, J. H.; Agard, D. A. Energetic Landscape of *α*-Lytic Protease Optimizes Longevity through Kinetic Stability. *Nature* **2002**, *415*, 343−346.
(40) Young, T. A.; Skordalakes, E.; Marqusee, S. Comparison of Proteolytic Susceptibility in Phosphoglycerate Kinases from Yeast and *E. coli*: Modulation of Conformational Ensembles Without Altering Structure or Stability. *J. Mol. Biol.* **2007**, *368*, 1438−1447.
(41) Shirke, A. N.; Basore, D.; Butterfoss, G. L.; Bonneau, R.; Bystroff, C.; Gross, R. A. Toward Rational Thermostabilization of Aspergillus Oryzae Cutinase: Insights into Catalytic and Structural Stability. *Proteins: Struct., Funct., Genet.* **2016**, *84*, 60−72.
(42) Liu, B.; Zhang, J.; Li, B.; Liao, X.; Du, G.; Chen, J. Expression and Characterization of Extreme Alkaline, Oxidation-Resistant Keratinase from Bacillus Licheniformis in Recombinant Bacillus Subtilis WB600 Expression System and Its Application in Wool Fiber Processing. *World J. Microbiol. Biotechnol.* **2013**, *29*, 825−832.
(43) Nguyen, V.; Wilson, C.; Hoemberger, M.; Stiller, J. B.; Agafonov, R. V.; Kutter, S.; English, J.; Theobald, D. L.; Kern, D. Evolutionary Drivers of Thermoadaptation in Enzyme Catalysis. *Science* **2017**, *355*, 289−294.
(44) Risso, V. A.; Gavira, J. A.; Gaucher, E. A.; Sanchez-Ruiz, J. M. Phenotypic Comparisons of Consensus Variants versus Laboratory Resurrections of Precambrian Proteins. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 887−896.
(45) Bednar, D.; Beerens, K.; Sebestova, E.; Bendl, J.; Khare, S.; Chaloupkova, R.; Prokop, Z.; Brezovsky, J.; Baker, D.; Damborsky, J. FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants. *PLoS Comput. Biol.* **2015**, *11*, e1004556.
(46) Babkova, P.; Sebestova, E.; Brezovsky, J.; Chaloupkova, R.; Damborsky, J. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *ChemBioChem* **2017**, *18*, 1448−1456.
(47) Goldenzweig, A.; Goldsmith, M.; Hill, S. E.; Gertman, O.; Laurino, P.; Ashani, Y.; Dym, O.; Unger, T.; Albeck, S.; Prilusky, J.; Lieberman, R. L.; Aharoni, A.; Silman, I.; Sussman, J. L.; Tawfik, D. S.; Fleishman, S. J. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **2016**, *63*, 337−346.
(48) Hammes, G. G.; Chang, Y.-C.; Oas, T. G. Conformational Selection or Induced Fit: A Flux Description of Reaction Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 13737.
(49) Kramer, R. M.; Shende, V. R.; Motl, N.; Pace, C. N.; Scholtz, J. M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophys. J.* **2012**, *102*, 1907−1915.
(50) Khow, O.; Suntrarachun, S. Strategies for Production of Active Eukaryotic Proteins in Bacterial Expression System. *Asian Pac. J. Trop. Biomed.* **2012**, *2*, 159−162.
(51) Sørensen, H. P.; Mortensen, K. K. Soluble Expression of Recombinant Proteins in the Cytoplasm of *Escherichia coli*. *Microb. Cell Fact.* **2005**, *4*, 1.
(52) Hartl, F. U.; Bracher, A.; Hayer-Hartl, M. Molecular Chaperones in Protein Folding and Proteostasis. *Nature* **2011**, *475*, 324−332.
(53) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341−346.

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

(54) Englander, S. W.; Mayne, L. The Case for Defined Protein Folding Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 8253−8258.
(55) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9(1−39). *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.
(56) Eaton, W. A.; Wolynes, P. G. Theory, Simulations, and Experiments Show That Proteins Fold by Multiple Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E9759−E9760.
(57) Yang, Y.; Niroula, A.; Shen, B.; Vihinen, M. PON-Sol: Prediction of Effects of Amino Acid Substitutions on Protein Solubility. *Bioinformatics* **2016**, *32*, 2032−2034.
(58) Broom, A.; Jacobi, Z.; Trainor, K.; Meiering, E. M. Computational Tools Help Improve Protein Stability but with a Solubility Tradeoff. *J. Biol. Chem.* **2017**, *292*, 14349−14361.
(59) Cabantous, S.; Waldo, G. S. *In Vivo* and *in Vitro* Protein Solubility Assays Using Split GFP. *Nat. Methods* **2006**, *3*, 845−854.
(60) Niwa, T.; Ying, B.-W.; Saito, K.; Jin, W.; Takada, S.; Ueda, T.; Taguchi, H. Bimodal Protein Solubility Distribution Revealed by an Aggregation Analysis of the Entire Ensemble of *Escherichia coli* Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 4201−4206.
(61) Eijsink, V. G.; Vriend, G.; van den Burg, B.; van der Zee, J. R.; Veltman, O. R.; Stulp, B. K.; Venema, G. Introduction of a Stabilizing 10 Residue Beta-Hairpin in Bacillus Subtilis Neutral Protease. *Protein Eng., Des. Sel.* **1992**, *5*, 157−163.
(62) Lee, C.; Levitt, M. Accurate Prediction of the Stability and Activity Effects of Site-Directed Mutagenesis on a Protein Core. *Nature* **1991**, *352*, 448−451.
(63) Buß, O.; Muller, D.; Jager, S.; Rudat, J.; Rabe, K. S. Improvement in the Thermostability of a *β*-Amino Acid Converting *ω*-Transaminase by Using FoldX. *ChemBioChem* **2018**, *19*, 379−387.
(64) Modarres, H. P.; Mofrad, M. R.; Sanati-Nezhad, A. Protein Thermostability Engineering. *RSC Adv.* **2016**, *6*, 115252−115270.
(65) Pace, C. N.; Scholtz, J. M.; Grimsley, G. R. Forces Stabilizing Proteins. *FEBS Lett.* **2014**, *588*, 2177−2184.
(66) Lazaridis, T.; Karplus, M. Effective Energy Functions for Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139−145.
(67) Seeliger, D.; de Groot, B. L. Protein Thermostability Calculations Using Alchemical Free Energy Simulations. *Biophys. J.* **2010**, *98*, 2309−2316.
(68) 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.
(69) Wickstrom, L.; Gallicchio, E.; Levy, R. M. The Linear Interaction Energy Method for the Prediction of Protein Stability Changes Upon Mutation. *Proteins: Struct., Funct., Genet.* **2012**, *80*, 111−125.
(70) 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.
(71) Mendes, J.; Guerois, R.; Serrano, L. Energy Estimation in Protein Design. *Curr. Opin. Struct. Biol.* **2002**, *12*, 441−446.
(72) Dehouck, Y.; Gilis, D.; Rooman, M. A New Generation of Statistical Potentials for Proteins. *Biophys. J.* **2006**, *90*, 4010−4017.
(73) Dehouck, Y.; Kwasigroch, J. M.; Gilis, D.; Rooman, M. PoPMuSiC 2.1: A Web Server for the Estimation of Protein Stability Changes upon Mutation and Sequence Optimality. *BMC Bioinf.* **2011**, *12*, 151.
(74) Liu, H. On Statistical Energy Functions for Biomolecular Modeling and Design. *Quant. Biol.* **2015**, *3*, 157−167.
(75) Kumar, M. D. 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*, D204−206.
(76) Pucci, F.; Bourgeas, R.; Rooman, M. High-Quality Thermodynamic Data on the Stability Changes of Proteins Upon Single-Site Mutations. *J. Phys. Chem. Ref. Data* **2016**, *45*, 023104.
(77) 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.
(78) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res.* **2005**, *33*, W382−388.
(79) Kepp, K. P. Towards a "Golden Standard" for Computing Globin Stability: Stability and Structure Sensitivity of Myoglobin Mutants. *Biochim. Biophys. Acta, Proteins Proteomics* **2015**, *1854*, 1239−1248.
(80) Christensen, N. J.; Kepp, K. P. Accurate Stabilities of Laccase Mutants Predicted with a Modified FoldX Protocol. *J. Chem. Inf. Model.* **2012**, *52*, 3028−3042.
(81) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(82) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656−1676.
(83) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13*, 3031−3048.
(84) Davey, J. A.; Damry, A. M.; Euler, C. K.; Goto, N. K.; Chica, R. A. Prediction of Stable Globular Proteins Using Negative Design with Non-Native Backbone Ensembles. *Structure* **2015**, *23*, 2011−2021.
(85) Ó Conchúir, S.; Barlow, K. A.; Pache, R. A.; Ollikainen, N.; Kundert, K.; O'Meara, M. J.; Smith, C. A.; Kortemme, T. A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. *PLoS One* **2015**, *10*, e0130433.
(86) Trainor, K.; Broom, A.; Meiering, E. M. Exploring the Relationships between Protein Sequence, Structure and Solubility. *Curr. Opin. Struct. Biol.* **2017**, *42*, 136−146.
(87) Das, R. Four Small Puzzles That Rosetta Doesn't Solve. *PLoS One* **2011**, *6*, e20044.
(88) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Role of Conformational Sampling in Computing Mutation-Induced Changes in Protein Structure and Stability. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 830−838.
(89) Musil, M.; Stourac, J.; Bendl, J.; Brezovsky, J.; Prokop, Z.; Zendulka, J.; Martinek, T.; Bednar, D.; Damborsky, J. FireProt: Web Server for Automated Design of Thermostable Proteins. *Nucleic Acids Res.* **2017**, *45*, W393−W399.
(90) Bush, J.; Makhatadze, G. I. Statistical Analysis of Protein Structures Suggests That Buried Ionizable Residues in Proteins Are Hydrogen Bonded or Form Salt Bridges. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 2027−2032.
(91) Stranges, P. B.; Kuhlman, B. A Comparison of Successful and Failed Protein Interface Designs Highlights the Challenges of Designing Buried Hydrogen Bonds. *Protein Sci.* **2013**, *22*, 74−82.
(92) Beerens, K.; Mazurenko, S.; Kunka, A.; Marques, S. M.; Hansen, N.; Musil, M.; Chaloupkova, R.; Waterman, J.; Brezovsky, J.; Bednar, D.; Prokop, Z.; Damborsky, J. Evolutionary Analysis Is a Powerful Complement to Energy Calculations for Protein Stabilization. *ACS Catal.* **2018**, *8*, 9420−9428.
(93) Wijma, H. J.; Floor, R. J.; Jekel, P. A.; Baker, D.; Marrink, S. J.; Janssen, D. B. Computationally Designed Libraries for Rapid Enzyme Stabilization. *Protein Eng., Des. Sel.* **2014**, *27*, 49−58.

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

(94) Thiltgen, G.; Goldstein, R. A. Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency. *PLoS One* **2012**, *7*, e46084.
(95) Buß, O.; Rudat, J.; Ochsenreither, K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? *Comput. Struct. Biotechnol. J.* **2018**, *16*, 25−33.
(96) Allen, B. D.; Nisthal, A.; Mayo, S. L. Experimental Library Screening Demonstrates the Successful Application of Computational Protein Design to Large Structural Ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 19838−19843.
(97) Barlow, K. A.; Ó Conchúir, S.; Thompson, S.; Suresh, P.; Lucas, J. E.; Heinonen, M.; Kortemme, T. Flex DdG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *J. Phys. Chem. B* **2018**, *122*, 5389−5399.
(98) Ludwiczak, J.; Jarmula, A.; Dunin-Horkawicz, S. Combining Rosetta with Molecular Dynamics (MD): A Benchmark of the MD-Based Ensemble Protein Design. *J. Struct. Biol.* **2018**, *203*, 54−61.
(99) Davis, I. W.; Arendall, W. B.; Richardson, D. C.; Richardson, J. S. The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. *Structure* **2006**, *14*, 265−274.
(100) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **2016**, *116*, 6516−6551.
(101) Fan, H.; Mark, A. E. Relative Stability of Protein Structures Determined by X-Ray Crystallography or NMR Spectroscopy: A Molecular Dynamics Simulation Study. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 111−120.
(102) Kuzmanic, A.; Pannu, N. S.; Zagrovic, B. X-Ray Refinement Significantly Underestimates the Level of Microscopic Heterogeneity in Biomolecular Crystals. *Nat. Commun.* **2014**, *5*, 3220.
(103) Karshikoff, A.; Nilsson, L.; Ladenstein, R. Rigidity versus Flexibility: The Dilemma of Understanding Protein Thermal Stability. *FEBS J.* **2015**, *282*, 3899−3917.
(104) Der, B. S.; Kluwe, C.; Miklos, A. E.; Jacak, R.; Lyskov, S.; Gray, J. J.; Georgiou, G.; Ellington, A. D.; Kuhlman, B. Alternative Computational Protocols for Supercharging Protein Surfaces for Reversible Unfolding and Retention of Stability. *PLoS One* **2013**, *8*, e64363.
(105) Chan, P.; Curtis, R. A.; Warwicker, J. Soluble Expression of Proteins Correlates with a Lack of Positively-Charged Surface. *Sci. Rep.* **2013**, *3*, 3333.
(106) Rezaie, E.; Mohammadi, M.; Sakhteman, A.; Bemani, P.; Ahrari, S. Application of Molecular Dynamics Simulations To Design a Dual-Purpose Oligopeptide Linker Sequence for Fusion Proteins. *J. Mol. Model.* **2018**, *24*, 313.
(107) Folkman, L.; Stantic, B.; Sattar, A.; Zhou, Y. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. *J. Mol. Biol.* **2016**, *428*, 1394−1405.
(108) Teng, S.; Srivastava, A. K.; Wang, L. Sequence Feature-Based Prediction of Protein Stability Changes upon Amino Acid Substitutions. *BMC Genomics* **2010**, *11* (Suppl 2), S5.
(109) Huang, L.-T.; Gromiha, M. M.; Ho, S.-Y. IPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes upon Mutations. *Bioinformatics* **2007**, *23*, 1292−1293.
(110) Paladin, L.; Piovesan, D.; Tosatto, S. C. E. SODA: Prediction of Protein Solubility from Disorder and Aggregation Propensity. *Nucleic Acids Res.* **2017**, *45*, W236−W240.
(111) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2*, 18−22.
(112) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(113) Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLoS One* **2017**, *12*, e0177678.
(114) Ling, C. X.; Sheng, V. S. Cost-Sensitive Learning and the Class Imbalance Problem. In *Encyclopedia of Machine Learning*; Sammut, C., Ed.; Springer: New York, 2007.
(115) Rao, R.; Fung, G.; Rosales, R. On the Dangers of Cross-Validation. An Experimental Evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*; Society for Industrial and Applied Mathematics: Philadelphia, PA, 2008; pp 588−596.
(116) Stephens, Z. D.; Lee, S. Y.; Faghri, F.; Campbell, R. H.; Zhai, C.; Efron, M. J.; Iyer, R.; Schatz, M. C.; Sinha, S.; Robinson, G. E. Big Data: Astronomical or Genomical? *PLoS Biol.* **2015**, *13*, e1002195.
(117) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(118) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(119) Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics* **1998**, *14*, 755−763.
(120) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM−HMM Alignment. *Nat. Methods* **2012**, *9*, 173−175.
(121) Pearson, W. R. An Introduction to Sequence Similarity ("Homology") Searching. *Curr. Protoc. Bioinf.* **2013**, *42*, 3.1.1−3.1.8.
(122) Rost, B. Twilight Zone of Protein Sequence Alignments. *Protein Eng., Des. Sel.* **1999**, *12*, 85−94.
(123) Fletcher, W.; Yang, Z. The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. *Mol. Biol. Evol.* **2010**, *27*, 2257−2267.
(124) Vialle, R. A.; Tamuri, A. U.; Goldman, N. Alignment Modulates Ancestral Sequence Reconstruction Accuracy. *Mol. Biol. Evol.* **2018**, *35*, 1783−1797.
(125) Chowdhury, B.; Garai, G. A Review on Multiple Sequence Alignment from the Perspective of Genetic Algorithm. *Genomics* **2017**, *109*, 419−431.
(126) Taly, J.-F.; Magis, C.; Bussotti, G.; Chang, J.-M.; Di Tommaso, P.; Erb, I.; Espinosa-Carrasco, J.; Kemena, C.; Notredame, C. Using the T-Coffee Package to Build Multiple Sequence Alignments of Protein, RNA, DNA Sequences and 3D Structures. *Nat. Protoc.* **2011**, *6*, 1669−1682.
(127) Pei, J.; Grishin, N. V. PROMALS3D: Multiple Protein Sequence Alignment Enhanced with Evolutionary and Three-Dimensional Structural Information. *Methods Mol. Biol.* **2014**, *1079*, 263−271.
(128) Steipe, B.; Schiller, B.; Plückthun, A.; Steinbacher, S. Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *J. Mol. Biol.* **1994**, *240*, 188−192.
(129) Sullivan, B. J.; Nguyen, T.; Durani, V.; Mathur, D.; Rojas, S.; Thomas, M.; Syu, T.; Magliery, T. J. Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability. *J. Mol. Biol.* **2012**, *420*, 384−399.
(130) Lehmann, M.; Kostrewa, D.; Wyss, M.; Brugger, R.; D'Arcy, A.; Pasamontes, L.; van Loon, A. P. From DNA Sequence to Improved Functionality: Using Protein Sequence Comparisons to Rapidly Design a Thermostable Consensus Phytase. *Protein Eng., Des. Sel.* **2000**, *13*, 49−57.
(131) Magliery, T. J. Protein Stability: Computation, Sequence Statistics, and New Experimental Methods. *Curr. Opin. Struct. Biol.* **2015**, *33*, 161−168.
(132) Porebski, B. T.; Buckle, A. M. Consensus Protein Design. *Protein Eng., Des. Sel.* **2016**, *29*, 245−251.
(133) Jäckel, C.; Bloom, J. D.; Kast, P.; Arnold, F. H.; Hilvert, D. Consensus Protein Design without Phylogenetic Bias. *J. Mol. Biol.* **2010**, *399*, 541−546.
(134) Goyal, V. D.; Magliery, T. J. Phylogenetic Spread of Sequence Data Affects Fitness of SOD1 Consensus Enzymes: Insights from Sequence Statistics and Structural Analyses. *Proteins: Struct., Funct., Genet.* **2018**, *86*, 609−620.
(135) Vázquez-Figueroa, E.; Chaparro-Riggers, J.; Bommarius, A. S. Development of a Thermostable Glucose Dehydrogenase by a

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

Structure-Guided Consensus Concept. *ChemBioChem* **2007**, *8*, 2295−2301.
(136) Parthasarathy, S.; Murthy, M. R. Protein Thermal Stability: Insights from Atomic Displacement Parameters (B Values). *Protein Eng., Des. Sel.* **2000**, *13*, 9−13.
(137) Cole, M. F.; Gaucher, E. A. Exploiting Models of Molecular Evolution to Efficiently Direct Protein Engineering. *J. Mol. Evol.* **2011**, *72*, 193−203.
(138) Hochberg, G. K. A.; Thornton, J. W. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu. Rev. Biophys.* **2017**, *46*, 247−269.
(139) Aerts, D.; Verhaeghe, T.; Joosten, H.-J.; Vriend, G.; Soetaert, W.; Desmet, T. Consensus Engineering of Sucrose Phosphorylase: The Outcome Reflects the Sequence Input. *Biotechnol. Bioeng.* **2013**, *110*, 2563−2572.
(140) Trudeau, D. L.; Kaltenbach, M.; Tawfik, D. S. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Mol. Biol. Evol.* **2016**, *33*, 2633−2641.
(141) Wheeler, L. C.; Lim, S. A.; Marqusee, S.; Harms, M. J. The Thermostability and Specificity of Ancient Proteins. *Curr. Opin. Struct. Biol.* **2016**, *38*, 37−43.
(142) Yang, Z. PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood. *Bioinformatics* **1997**, *13*, 555−556.
(143) Stamatakis, A. RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics* **2006**, *22*, 2688−2690.
(144) Huelsenbeck, J. P.; Ronquist, F.; Nielsen, R.; Bollback, J. P. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* **2001**, *294*, 2310−2314.
(145) Goldstein, R. A.; Pollard, S. T.; Shah, S. D.; Pollock, D. D. Nonadaptive Amino Acid Convergence Rates Decrease over Time. *Mol. Biol. Evol.* **2015**, *32*, 1373−1381.
(146) Williams, P. D.; Pollock, D. D.; Blackburne, B. P.; Goldstein, R. A. Assessing the Accuracy of Ancestral Protein Reconstruction Methods. *PLoS Comput. Biol.* **2006**, *2*, e69.
(147) Eick, G. N.; Bridgham, J. T.; Anderson, D. P.; Harms, M. J.; Thornton, J. W. Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty. *Mol. Biol. Evol.* **2016**, *34*, 247−261.
(148) Gaucher, E. A.; Govindarajan, S.; Ganesh, O. K. Palaeotemperature Trend for Precambrian Life Inferred from Resurrected Proteins. *Nature* **2008**, *451*, 704−707.
(149) Akanuma, S. Characterization of Reconstructed Ancestral Proteins Suggests a Change in Temperature of the Ancient Biosphere. *Life (Basel, Switz.)* **2017**, *7*, 33.
(150) Gumulya, Y.; Baek, J.-M.; Wun, S.-J.; Thomson, R. E. S.; Harris, K. L.; Hunter, D. J. B.; Behrendorff, J. B. Y. H.; Kulig, J.; Zheng, S.; Wu, X.; Wu, B.; Stok, J. E.; De Voss, J. J.; Schenk, G.; Jurva, U.; Andersson, S.; Isin, E. M.; Bodén, M.; Guddat, L.; Gillam, E. M. J. Engineering Highly Functional Thermostable Proteins Using Ancestral Sequence Reconstruction. *Nat. Catal.* **2018**, *1*, 878.
(151) Dehouck, Y.; Grosfils, A.; Folch, B.; Gilis, D.; Bogaerts, P.; Rooman, M. Fast and Accurate Predictions of Protein Stability Changes upon Mutations Using Statistical Potentials and Neural Networks: PoPMuSiC-2.0. *Bioinformatics* **2009**, *25*, 2537−2543.
(152) Khatun, J.; Khare, S. D.; Dokholyan, N. V. Can Contact Potentials Reliably Predict Stability of Proteins? *J. Mol. Biol.* **2004**, *336*, 1223−1238.
(153) 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.
(154) Yin, S.; Ding, F.; Dokholyan, N. V. Eris: An Automated Estimator of Protein Stability. *Nat. Methods* **2007**, *4*, 466−467.
(155) 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*, 3−4.
(156) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845−854.
(157) de Groot, B. L.; van Aalten, D. M.; Scheek, R. M.; Amadei, A.; Vriend, G.; Berendsen, H. J. C. Prediction of Protein Conformational Freedom from Distance Constraints. *Proteins: Struct., Funct., Genet.* **1997**, *29*, 240−251.
(158) Hoppe, C.; Schomburg, D. Prediction of Protein Thermostability with a Direction- and Distance-Dependent Knowledge-Based Potential. *Protein Sci.* **2005**, *14*, 2682−2692.
(159) Pucci, F.; Bourgeas, R.; Rooman, M. Predicting Protein Thermal Stability Changes upon Point Mutations Using Statistical Potentials: Introducing HoTMuSiC. *Sci. Rep.* **2016**, *6*, 23257.
(160) 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.
(161) Cheng, J.; Randall, A.; Baldi, P. Prediction of Protein Stability Changes for Single-Site Mutations Using Support Vector Machines. *Proteins: Struct., Funct., Genet.* **2006**, *62*, 1125−1132.
(162) Wainreb, G.; Wolf, L.; Ashkenazy, H.; Dehouck, Y.; Ben-Tal, N. Protein Stability: A Single Recorded Mutation Aids in Predicting the Effects of Other Mutations in the Same Amino Acid Site. *Bioinformatics* **2011**, *27*, 3286−3292.
(163) Li, Y.; Fang, J. PROTS-RF: A Robust Model for Predicting Mutation-Induced Protein Stability Changes. *PLoS One* **2012**, *7*, e47247.
(164) Quang, D.; Chen, Y.; Xie, X. DANN: A Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants. *Bioinformatics* **2015**, *31*, 761−763.
(165) Wang, Y.; Mao, H.; Yi, Z. Protein Secondary Structure Prediction by Using Deep Learning Method. *Knowl.-Based Syst.* **2017**, *118*, 115−123.
(166) Ivakhnenko, A. G. Polynomial Theory of Complex Systems. *IEEE Trans. Syst., Man, Cybern.* **1971**, *SMC-1*, 364−378.
(167) Bengio, Y.; Boulanger-Lewandowski, N.; Pascanu, R. Advances in Optimizing Recurrent Networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE: New York, 2013; pp 8624−8628.
(168) Cang, Z.; Wei, G.-W. TopologyNet: Topology Based Deep Convolutional and Multi-Task Neural Networks for Biomolecular Property Predictions. *PLoS Comput. Biol.* **2017**, *13*, e1005690.
(169) Laimer, J.; Hofer, H.; Fritz, M.; Wegenkittl, S.; Lackner, P. MAESTRO - Multi Agent Stability Prediction upon Point Mutations. *BMC Bioinf.* **2015**, *16*, 116.
(170) Khan, S.; Vihinen, M. Performance of Protein Stability Predictors. *Hum. Mutat.* **2010**, *31*, 675−684.
(171) 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.
(172) Montanucci, L.; Martelli, P. L.; Ben-Tal, N.; Fariselli, P. A Natural Upper Bound to the Accuracy of Predicting Protein Stability Changes upon Mutations. 2018, arXiv:1809.10389 [q-bio.BM]. arXiv.org e-Print archive. https://arxiv.org/abs/1809.10389.
(173) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276−277.
(174) Lu, G.; Moriyama, E. N. Vector NTI, a Balanced All-in-One Sequence Analysis Suite. *Briefings Bioinf.* **2004**, *5*, 378−388.
(175) Bendl, J.; Stourac, J.; Sebestova, E.; Vavra, O.; Musil, M.; Brezovsky, J.; Damborsky, J. HotSpot Wizard 2.0: Automated Design of Site-Specific Mutations and Smart Libraries in Protein Engineering. *Nucleic Acids Res.* **2016**, *44*, W479−487.
(176) Stamatakis, A. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* **2014**, *30*, 1312−1313.
(177) Ashkenazy, H.; Penn, O.; Doron-Faigenboim, A.; Cohen, O.; Cannarozzi, G.; Zomer, O.; Pupko, T. FastML: A Web Server for

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

Probabilistic Reconstruction of Ancestral Sequences. *Nucleic Acids Res.* **2012**, *40*, W580−584.
(178) Diallo, A. B.; Makarenkov, V.; Blanchette, M. Ancestors 1.0: A Web Server for Ancestral Sequence Reconstruction. *Bioinformatics* **2010**, *26*, 130−131.
(179) Westesson, O.; Barquist, L.; Holmes, I. HandAlign: Bayesian Multiple Sequence Alignment, Phylogeny and Ancestral Reconstruction. *Bioinformatics* **2012**, *28*, 1170−1171.
(180) Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D. L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M. A.; Huelsenbeck, J. P. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Syst. Biol.* **2012**, *61*, 539−542.
(181) Finn, R. D.; Clements, J.; Eddy, S. R. HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res.* **2011**, *39*, W29−37.
(182) Altschul, S. F.; Gertz, E. M.; Agarwala, R.; Schäffer, A. A.; Yu, Y.-K. PSI-BLAST Pseudocounts and the Minimum Description Length Principle. *Nucleic Acids Res.* **2009**, *37*, 815−824.
(183) Whitehead, T. A.; Chevalier, A.; Song, Y.; Dreyfus, C.; Fleishman, S. J.; De Mattos, C.; Myers, C. A.; Kamisetty, H.; Blair, P.; Wilson, I. A.; Baker, D. Optimization of Affinity, Specificity and Function of Designed Influenza Inhibitors Using Deep Sequencing. *Nat. Biotechnol.* **2012**, *30*, 543−548.
(184) Shimizu, Y.; Inoue, A.; Tomari, Y.; Suzuki, T.; Yokogawa, T.; Nishikawa, K.; Ueda, T. Cell-Free Translation Reconstituted with Purified Components. *Nat. Biotechnol.* **2001**, *19*, 751−755.
(185) Niwa, T.; Kanamori, T.; Ueda, T.; Taguchi, H. Global Analysis of Chaperone Effects Using a Reconstituted Cell-Free Translation System. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 8937−8942.
(186) Berman, H. M.; Gabanyi, M. J.; Kouranov, A.; Micallef, D. I.; Westbrook, J. Protein Structure Initiative - TargetTrack 2000−2017 - All Data Files. DOI: 10.5281/zenodo.821654.
(187) Price, W. N.; Handelman, S. K.; Everett, J. K.; Tong, S. N.; Bracic, A.; Luff, J. D.; Naumov, V.; Acton, T.; Manor, P.; Xiao, R.; Rost, B.; Montelione, G. T.; Hunt, J. F. Large-Scale Experimental Studies Show Unexpected Amino Acid Effects on Protein Expression and Solubility in Vivo in *E. coli*. *Microb. Inf. Exp.* **2011**, *1*, 6.
(188) Hirose, S.; Kawamura, Y.; Yokota, K.; Kuroita, T.; Natsume, T.; Komiya, K.; Tsutsumi, T.; Suwa, Y.; Isogai, T.; Goshima, N.; Noguchi, T. Statistical Analysis of Features Associated with Protein Expression/Solubility in an in Vivo *Escherichia coli* Expression System and a Wheat Germ Cell-Free Expression System. *J. Biochem.* **2011**, *150*, 73−81.
(189) Pawlicki, S.; Le Béchec, A.; Delamarche, C. AMYPdb: A Database Dedicated to Amyloid Precursor Proteins. *BMC Bioinf.* **2008**, *9*, 273.
(190) Thompson, M. J.; Sievers, S. A.; Karanicolas, J.; Ivanova, M. I.; Baker, D.; Eisenberg, D. The 3D Profile Method for Identifying Fibril-Forming Segments of Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 4074−4078.
(191) Beerten, J.; Van Durme, J.; Gallardo, R.; Capriotti, E.; Serpell, L.; Rousseau, F.; Schymkowitz, J. WALTZ-DB: A Benchmark Database of Amyloidogenic Hexapeptides. *Bioinformatics* **2015**, *31*, 1698−1700.
(192) Wozniak, P. P.; Kotulska, M. AmyLoad: Website Dedicated to Amyloidogenic Protein Fragments. *Bioinformatics* **2015**, *31*, 3395−3397.
(193) Sastry, A.; Monk, J.; Tegel, H.; Uhlen, M.; Palsson, B. O.; Rockberg, J.; Brunk, E. Machine Learning in Computational Biology to Accelerate High-Throughput Protein Expression. *Bioinformatics* **2017**, *33*, 2487−2495.
(194) Thangakani, A. M.; Nagarajan, R.; Kumar, S.; Sakthivel, R.; Velmurugan, D.; Gromiha, M. M. CPAD, Curated Protein Aggregation Database: A Repository of Manually Curated Experimental Data on Protein and Peptide Aggregation. *PLoS One* **2016**, *11*, e0152949.
(195) Tian, Y.; Deutsch, C.; Krishnamoorthy, B. Scoring Function To Predict Solubility Mutagenesis. *Algorithms Mol. Biol.* **2010**, *5*, 33.
(196) Wilkinson, D. L.; Harrison, R. G. Predicting the Solubility of Recombinant Proteins in *Escherichia coli*. *Nat. Biotechnol.* **1991**, *9*, 443−448.
(197) Davis, G. D.; Elisee, C.; Newham, D. M.; Harrison, R. G. New Fusion Protein Systems Designed to Give Soluble Expression in *Escherichia coli*. *Biotechnol. Bioeng.* **1999**, *65*, 382−388.
(198) Magnan, C. N.; Randall, A.; Baldi, P. SOLpro: Accurate Sequence-Based Prediction of Protein Solubility. *Bioinformatics* **2009**, *25*, 2200−2207.
(199) Smialowski, P.; Doose, G.; Torkler, P.; Kaufmann, S.; Frishman, D. PROSO II—A New Method for Protein Solubility Prediction. *FEBS J.* **2012**, *279*, 2192−2200.
(200) Agostini, F.; Cirillo, D.; Livi, C. M.; Delli Ponti, R.; Tartaglia, G. G. CcSOL Omics: A Webserver for Solubility Prediction of Endogenous and Heterologous Expression in *Escherichia coli*. *Bioinformatics* **2014**, *30*, 2975−2977.
(201) Khurana, S.; Rawi, R.; Kunji, K.; Chuang, G.-Y.; Bensmail, H.; Mall, R. DeepSol: A Deep Learning Framework for Sequence-Based Protein Solubility Prediction. *Bioinformatics* **2018**, *34*, 2605−2613.
(202) Chang, C. C. H.; Li, C.; Webb, G. I.; Tey, B.; Song, J.; Ramanan, R. N. Periscope: Quantitative Prediction of Soluble Protein Expression in the Periplasm of *Escherichia coli*. *Sci. Rep.* **2016**, *6*, 21844.
(203) Hirose, S.; Noguchi, T. ESPRESSO: A System for Estimating Protein Expression and Solubility in Protein Expression Systems. *Proteomics* **2013**, *13*, 1444−1456.
(204) Hon, J.; Marusiak, M.; Martinek, T.; Zendulka, J.; Bednar, D.; Damborsky, J. SoluProt: Prediction of Protein Solubility. *Nucleic Acids Res.* **2018**, in preparation.
(205) DuBay, K. F.; Pawar, A. P.; Chiti, F.; Zurdo, J.; Dobson, C. M.; Vendruscolo, M. Prediction of the Absolute Aggregation Rates of Amyloidogenic Polypeptide Chains. *J. Mol. Biol.* **2004**, *341*, 1317−1326.
(206) Tartaglia, G. G.; Pawar, A. P.; Campioni, S.; Dobson, C. M.; Chiti, F.; Vendruscolo, M. Prediction of Aggregation-Prone Regions in Structured Proteins. *J. Mol. Biol.* **2008**, *380*, 425−436.
(207) Conchillo-Solé, O.; de Groot, N. S.; Avilés, F. X.; Vendrell, J.; Daura, X.; Ventura, S. AGGRESCAN: A Server for the Prediction and Evaluation of "Hot Spots" of Aggregation in Polypeptides. *BMC Bioinf.* **2007**, *8*, 65.
(208) Fernandez-Escamilla, A.-M.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Prediction of Sequence-Dependent and Mutational Effects on the Aggregation of Peptides and Proteins. *Nat. Biotechnol.* **2004**, *22*, 1302−1306.
(209) Maurer-Stroh, S.; Debulpaep, M.; Kuemmerer, N.; Lopez de la Paz, M.; Martins, I. C.; Reumers, J.; Morris, K. L.; Copland, A.; Serpell, L.; Serrano, L.; Schymkowitz, J. W. H.; Rousseau, F. Exploring the Sequence Determinants of Amyloid Structure Using Position-Specific Scoring Matrices. *Nat. Methods* **2010**, *7*, 237−242.
(210) Walsh, I.; Seno, F.; Tosatto, S. C. E.; Trovato, A. PASTA 2.0: An Improved Server for Protein Aggregation Prediction. *Nucleic Acids Res.* **2014**, *42*, W301−307.
(211) Bryan, A. W.; Menke, M.; Cowen, L. J.; Lindquist, S. L.; Berger, B. BETASCAN: Probable Beta-Amyloids Identified by Pairwise Probabilistic Analysis. *PLoS Comput. Biol.* **2009**, *5*, e1000333.
(212) Garbuzynskiy, S. O.; Lobanov, M. Y.; Galzitskaya, O. V. FoldAmyloid: A Method of Prediction of Amyloidogenic Regions from Protein Sequence. *Bioinformatics* **2010**, *26*, 326−332.
(213) Goldschmidt, L.; Teng, P. K.; Riek, R.; Eisenberg, D. Identifying the Amylome, Proteins Capable of Forming Amyloid-like Fibrils. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 3487−3492.
(214) Ahmed, A. B.; Znassi, N.; Château, M.-T.; Kajava, A. V. A Structure-Based Approach to Predict Predisposition to Amyloidosis. *Alzheimer's Dementia* **2015**, *11*, 681−690.
(215) Krogh, A.; Vedelsby, J. Neural Network Ensembles, Cross Validation and Active Learning. In *Proceedings of the 7th International*

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054

Conference on Neural Information Processing Systems (NIPS'94); MIT Press: Cambridge, MA, 1994; pp 231−238.
(216) Maclin, R.; Opitz, D. Popular Ensemble Methods: An Empirical Study. *J. Artif. Intell. Res.* **1999**, *11*, 169−198.
(217) Tsolis, A. C.; Papandreou, N. C.; Iconomidou, V. A.; Hamodrakas, S. J. A Consensus Method for the Prediction of "Aggregation-Prone" Peptides in Globular Proteins. *PLoS One* **2013**, *8*, e54175.
(218) Emily, M.; Talvas, A.; Delamarche, C. MetAmyl: A METa-Predictor for AMYLoid Proteins. *PLoS One* **2013**, *8*, e79722.
(219) Zambrano, R.; Jamroz, M.; Szczasiuk, A.; Pujols, J.; Kmiecik, S.; Ventura, S. AGGRESCAN3D (A3D): Server for Prediction of Aggregation Properties of Protein Structures. *Nucleic Acids Res.* **2015**, *43*, W306−313.
(220) De Baets, G.; Van Durme, J.; van der Kant, R.; Schymkowitz, J.; Rousseau, F. Solubis: Optimize Your Protein. *Bioinformatics* **2015**, *31*, 2580−2582.
(221) Van Durme, J.; De Baets, G.; Van Der Kant, R.; Ramakers, M.; Ganesan, A.; Wilkinson, H.; Gallardo, R.; Rousseau, F.; Schymkowitz, J. Solubis: A Webserver To Reduce Protein Aggregation through Mutation. *Protein Eng., Des. Sel.* **2016**, *29*, 285−289.

DOI: 10.1021/acscatal.8b03613
ACS Catal. 2019, 9, 1033−1054