doi:10.2533/chimia.2023.116 Chimia 77 (2023) 116–121 © D. Patsch, R. Buller

Improving Enzyme Fitness with Machine Learning

David Patsch[ab] and Rebecca Buller[a*]

Abstract: The combinatorial composition of proteins has triggered the application of machine learning in enzyme engineering. By predicting how protein sequence encodes function, researchers aim to leverage machine learning models to select a reduced number of optimized sequences for laboratory measurement with the aim to lower costs and shorten timelines of enzyme engineering campaigns. In this review, we highlight successful algorithm-aided protein engineering examples, including work carried out within NCCR Catalysis. In this context, we will discuss the underlying computational methods developed to improve enzyme properties such as enantioselectivity, regioselectivity, activity, and stability. Considering the rapid maturing of computational techniques, we expect that their continued application in enzyme engineering campaigns will be key to deliver additional powerful biocatalysts for sustainable chemical synthesis.

Keywords: Bioinformatics · Enzyme engineering · Halogenase · Industrial biocatalysis · Machine learning

David Patsch studied biology and received his BSc from the University of Innsbruck. He obtained his MSc degree in biotechnology from the Management Center Innsbruck. Since 2019 he is pursuing his PhD in the group of Rebecca Buller at the ZHAW.

Rebecca Buller is a biological chemist and Professor for Biotechnological Methods, Systems and Processes at the Zurich University of Applied Sciences. Rebecca Buller studied chemistry at the Westfälische – Wilhelms Universität Münster (D) and the University of California Santa Barbara (US). After completing her PhD with a focus on enzyme engineering at ETH Zurich (CH), Rebecca Buller accepted a position as laboratory head at the flavour and fragrance company Firmenich (CH). In 2015, she relocated to the Zurich University of Applied Sciences where she founded the Competence Center for Biocatalysis (CCBIO). Research in Rebecca Buller's laboratory focusses on the expansion of the biocatalytic toolbox by sourcing and engineering enzymes for synthetic applications.

1. Introduction

In optimal settings, enzymes can facilitate complex reactions with extraordinary specificity and selectivity.[1,2] However, practical reality usually differs from this ideal as wildtype enzymes are often just marginally stable in the selected reaction conditions[3] and perform at scales well below what is required to drive an industrial process. However, as enzymes are combinatorically composed from a limited set of simple building blocks, improved catalysts can be constructed in the laboratory by applying enzyme engineering strategies, among them the directed evolution of proteins. Consequently, engineered enzymes are harnessed in many industrial fields ranging from the fine chemical to the pharmaceutical sectors.[4–6]

Over the last decades, the technique of directed evolution has developed into a powerful tool (Nobel prize for chemistry 2018)[7] and today, it is routinely applied to tailor critical protein properties.[4,8] Directed evolution mimics nature's selection process in the laboratory through iterative cycles of gene diversification and selection of the encoded protein variants generating enzyme lineages with new or improved functions.[9] However, unlike nature, which selects for survival or reproduction, directed evolution can be used to precisely tailor desired protein traits.[10] In this context, astounding improvements in target biological functions for several different enzyme families have been achieved, including activity,[11–13] stereoselectivity,[14,15] thermostability,[16] and solvent tolerance.[17] Strikingly, these studies screened only a relatively small fraction of the target protein's underlying sequence space, raising the question of whether better sequence solutions would, in principle, exist for the function of interest. Unfortunately, such a question cannot easily be answered experimentally: Full randomization of a small protein consisting, for example, of 100 amino acids leads to a search space of sequences that is larger than the estimated number of atoms in the universe.[18] Even the targeted randomization of predefined positions within a protein quickly leads to a screening bottleneck: While replacing a single amino acid position with all other natural amino acids yields an experimentally manageable library size of 20^1 variants, combinatorically investigating as little as five sites in a protein already leads to a library size of 20^5. Clearly, it is difficult to experimentally screen such large libraries exhaustively, even when using advanced automation. In addition, most mutations introduced into a protein are either neutral or unfavorable,[19] leading to an even more inefficient sampling of the sequence space. To address the numbers problem in protein engineering, researchers are increasingly interested in implementing computational techniques, such as molecular dynamics simulations,[20] phylogeny, docking,[21,22] and, more recently, machine learning (ML) (Fig. 1).[23]

ML, in particular, has emerged as a powerful and versatile tool for various applications, many of which affect our daily lives, such as translating languages[24] or recommending what movies

Correspondence: Prof. R. Buller[a], E-Mail: rebecca.buller@zhaw.ch

[a]Institute of Chemistry and Biotechnology, Zurich University of Applied Sciences, CH-8820 Wädenswil, Switzerland; [b]Institute of Biochemistry, Dept. of Biotechnology & Enzyme Catalysis, Greifswald University, Felix-Hausdorff-Strasse 4, D17487 Greifswald, Germany

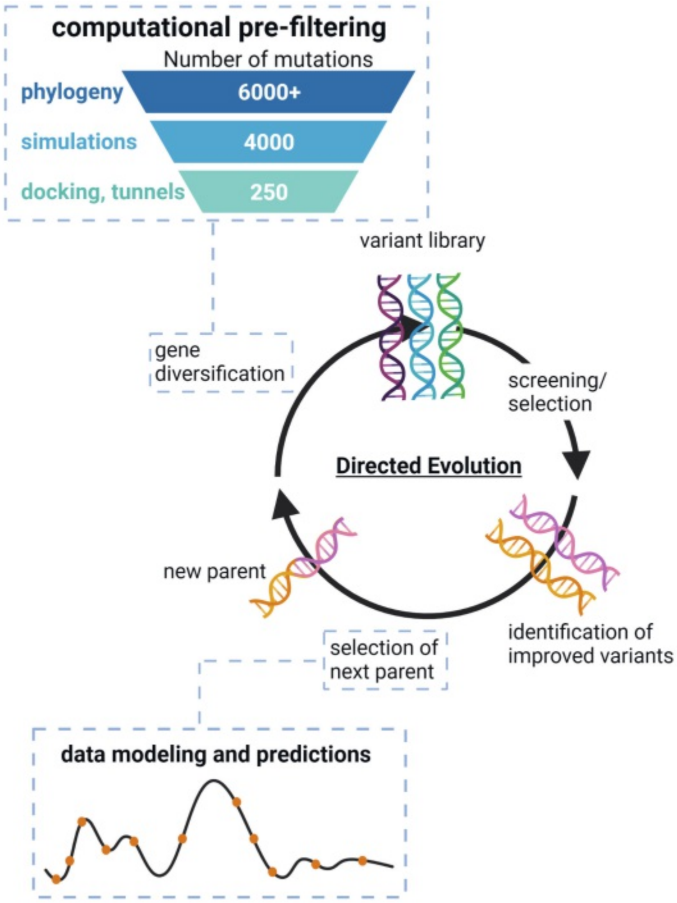

Fig. 1. Integration of *in silico* tools into directed evolution of proteins. As the random generation of genetic diversity is often inefficient when targeting to improve a desired function, information from various bioinformatic sources, such as phylogeny, docking, tunnels, and ML tools, can be used to build ‘smart’ enzyme libraries. Additionally, ML methods might be able to learn the underlying enzyme fitness landscape and suggest improved variants which have not yet been experimentally screened. Image created with BioRender.com.

to watch next.[25] Looking forward, ML is expected to profoundly impact the field of protein engineering as well. In contrast to traditional directed evolution, which discards information except if related to the most beneficial mutations, ML techniques can rely on all generated data to speed up the evolution process. This acceleration might be achieved by learning a function representing the underlying protein landscape from a set of sequence-fitness pairs. Based on this function, additional variants can be evaluated computationally, allowing exploration of the sequence space at a scale that cannot be achieved through laboratory experiments alone.[26] The potential benefits of ML make it an attractive research objective, and multiple attempts to apply it to protein engineering have been made. This report is by no means meant to cover them exhaustively but instead focuses on work related to research carried out in the frame of NCCR Catalysis.

2. ML-aided Optimization of Enzyme Stereoselectivity

From an organic chemist’s perspective, facilitating the tailoring of the stereo- and regioselectivity of enzymes might be one of the most exciting applications of ML in protein engineering.[27] In this context, a first ML-driven study to improve enantioselectivity for the selective ring opening of a racemic mixture of glycidyl phenyl ether catalyzed by an epoxide hydrolase from *Aspergillus niger* (ANEH) was published in 2012.[21,28] More selective ANEH variants were predicted through the adaptive substituent reordering algorithm (ASRA),[29,30] which focuses on finding beneficial regions in a combinatorial enzyme library with minimal screening effort. The underlying principle of the approach is to first evaluate a small subset of all possible variants of a combinatorial enzyme library experimentally before reordering the amino-acid pairs to maximize the smoothness of the measured property landscape (Fig. 2).[28] Unlike the traditional quantitative structure–activity relationships (QSAR), ASRA does not make explicit assumptions about linearity, additivity, or specific relationships between structure and function. It only relies on the hypothesis that the underlying protein landscape is, to some extent, smooth.[31,32] This is an assumption ASRA shares with most, if not all, computational approaches and, from our experience, represents a valid bias in protein engineering in most cases. Within the ANEH study,[28] ASRA was shown to be a powerful tool for obtaining reliable estimates about areas of interest within the sizable sequence space that arises from evaluating variants combinatorially. Notably, ASRA did not rely on complex protein/residue descriptors making the algorithm a compelling starting point for protein engineering campaigns.

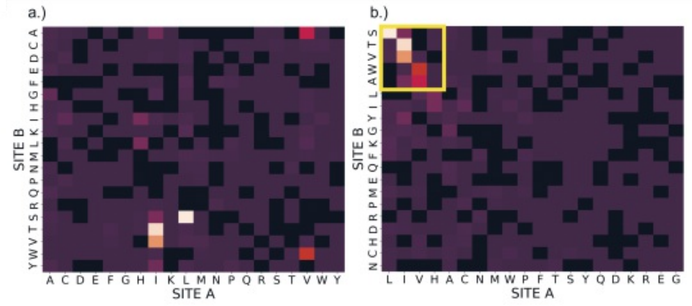

Fig. 2. Application of the ASRA algorithm: First, a random subset of a two-site combinatorial library is screened (left). Next, the amino acid pairs are rearranged to maximize the smoothness of the fitness landscape (right). This rearrangement highlights beneficial regions (yellow box in right plot) to explore in a library of reduced size. In this representation, black squares denote amino acid combinations which were not experimentally measured, whereas a colored filling indicates variants that have been measured for activity.

Following this first example, a second study on ML-aided directed evolution for stereoselectivity was published in 2018. Interestingly, it builds upon the same experimental platform as the previous example, namely the enantioselectivity of epoxide hydrolase from ANEH.[33] Starting from only nine experimentally evaluated single-point mutants, the researchers built a model and predicted the enantioselectivity of all combinations of these initial changes (2^9). The algorithm, which was used to predict the new sequences, dubbed innov’SAR, was developed by PEACCEL, a France-based biotechnology start-up focusing on enzyme engineering and drug discovery.[34] Innov’SAR only requires sequence information and experimental protein fitness values for training and subsequent inference. Overall, the applied process can be summarized in four steps: 1) The entire protein sequence is encoded based on each amino acid’s physicochemical and biochemical properties; 2) from this numerical protein representation, a protein spectrum is calculated through digital signal processing techniques; 3) the protein signals and their respective fitness values are used to construct a regression model; 4) this regression model finally predicts the properties of all possible variant permutations. Applying these steps to the epoxide hydrolase from ANEH led to predicted sequences which, when evaluated experimentally, revealed enzyme variants with improved enantioselectivity.

3. ML-aided Optimization of Enzyme Activity

Complementing the above-described applications of ML to boost enzyme stereoselectivity, we set out to explore algorithm-aided engineering of regioselectivity and activity. Interested in the late-stage functionalization of complex molecules by direct enzymatic CH activation, we explored the potential of Fe(II)/α-ketoglutarate dependent halogenases for the selective halogenation of soraphen A,[35] a potent anti-fungal agent and a target of pharmaceutical interest.[36] We identified a suitable starting sequence capable of catalyzing the desired halogenation reaction in a previously engineered variant of the halogenase WelO5* from *Hapalopsiphon welwitschii* IC-52-2.[37] Notably, we found that while the wildtype enzyme did not accept the bulky substrate, variants that had been specifically engineered to have a broader substrate spectrum exhibited activity.[37] Based on this initial reference and additional docking studies, we selected three critical residues (V81/A88/I161) for complete randomization, *e.g.,* replacement of each amino acid by all other 19 amino acids. As delineated above, the theoretical size of such a library calculates to 20^3. However, due to the redundancy of the genetic code and sampling reasons, the actual screening effort required to cover all combinations exhaustively increases. Specifically, if the screening aim is to cover at least 95% of all encoded variants in a library, a three-fold oversampling should be targeted,[38] challenging experimentalists.

In our halogenase engineering project, we thus opted to explore ML methods to reduce the experimental screening burden and accelerate the identification of beneficial mutations. Notably, our study considered two main engineering objectives: Firstly, we targeted to increase the overall chlorination activity of the enzyme variants, and secondly, we aimed to control the regioselectivity of the halogenation reaction, which would allow the analysis of several derivatized macrolides in structure-function relationship assays.[35]

As a first step, we experimentally confirmed 504 unique halogenase sequence–function pairs, corresponding to 6.3% of the theoretical library. Based on previous applications of ML in protein engineering,[39–41] we then explored the remaining sequence space *in silico*. Toward this goal, we first represented each variant numerically by concatenating the physicochemical and biochemical properties of the amino acids at each mutation site. Multiple amino acid descriptors exist, such as the very comprehensive AAindex[42] or the T-scale descriptor.[43] In our case, combining the T-scale descriptor and selected additional amino acid characteristics[44] produced the best results. With this representation in hand, we trained a Gaussian process. Gaussian processes have received increased attention in the ML community and have also been applied successfully to protein engineering.[39,40] They are accurate and flexible methods for regression and classification and can give a reliable estimate of their own uncertainty. Following training, our model was then used to make activity and regioselectivity predictions on the library's unexplored sequence space. The best-predicted variants were synthesized and experimentally assayed toward their activity and regioselectivity. Gratifyingly, all seven variants predicted towards increased activity performed well, with four halogenases outperforming the previous best variant (Fig. 3). Similarly, the variants predicted towards selectivity exhibited the desired enzyme trait: While seven out of eight produced halogenases showed high selectivity toward the chlorinated soraphen regioisomer **1b**, variant 'LHG' exhibited not only absolute regio-selectivity but also a doubled activity compared to the previous best **1b** producing variant.[37]

Overall, the algorithm-aided evolution process generated halogenase variants capable of synthesizing three distinct chlorinated species from soraphen A and its derivative soraphen C in quantities sufficient for biological testing. In the phenotypic tests, which were carried out on six key pathogens in crop protection, we found that soraphen A derivative **1b** showed an overall better performance than **1a** whereas a chlorinated soraphen C derivative displayed higher species selectivity than the other investigated compounds.[35]

A further successful computational technique in protein engineering is the analysis of protein sequence activity relationships (ProSAR), which has been successfully applied to construct sev-

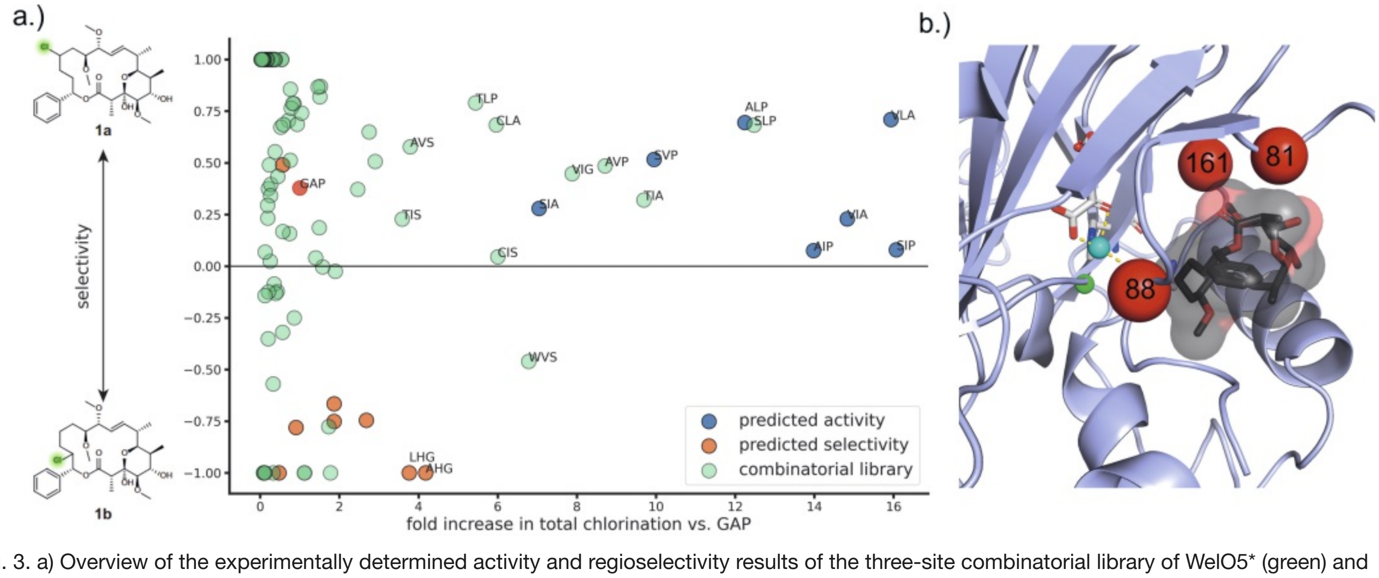

Fig. 3. a) Overview of the experimentally determined activity and regioselectivity results of the three-site combinatorial library of WelO5* (green) and the predicted variants towards activity (blue) and selectivity (orange). Halogenase variants were capable to produce two chlorinated products of soraphen A (**1a** and **1b**). The y-axis shows the regioselectivity of chlorination. The selectivity (S) is calculated using the formula $S = (SIM_{1a} – SIM_{1b})/(SIM_{1a} + SIM_{1b})$. Activity data is normalized to a reference variant (GAP), which was included as an internal reference on each measured 96-well plate. Each variant with a fold-improvement greater than 3.5 is highlighted with a three-letter code representative of the introduced mutations compared to wildtype. For example, V81V/A88L/I161A is shortened to VLA. b) Docking of soraphen A (black) into a model of variant WelO5* V81G/I161P (light blue). The enzyme model was generated using SWISS-MODEL[68] and the crystal structure of WelO5 (PDB ID: 5J4R) as a template. The macrolide soraphen A was docked using AutoDock Vina.[69] The red spheres indicate the targeted positions for the full randomization of the library.

eral highly optimized enzyme variants.[45,46] This technique, which was first published in 2005 by the US-based enzyme engineering company Codexis, facilitated the development of a halohydrin dehalogenase (HHDH) for the industrial production of ethyl (*R*)-4-cyano-3-hydroxybutyrate (HN), improving the enzyme's activity by ~4,000 fold compared to the initial wildtype enzyme (Fig. 4a). To achieve this goal, more than 18 rounds of evolution were carried out, during which 35 distinct mutations were introduced into the wildtype scaffold.[47] In later studies, ProSAR was also employed to increase the stability of a carbonic anhydrase (CA), translating to a 4,000,000-fold improvement over the wildtype in terms of compounded thermostability and alkali tolerance (Fig. 4c).[16] Furthermore, ProSAR enabled the development of a 140,000-fold improved Baeyer-Villiger monooxygenase for the commercial manufacture of esomeprazole used in the blockbuster drug Nexium® by engineering the natural biocatalyst over 19 rounds of evolution.[48] Very recently, ProSAR aided in identifying beneficial mutations in the evolution campaign of an amine transaminase, highly optimized for the efficient production of a chiral sacubitril precursor, a key component of a critical heart failure drug (Fig. 4b).[49]

The multivariate optimization strategy fueling the examples above is an iterative process consisting of diversity generation and statistical modeling. During diversity generation, potentially interesting mutations are generated from various methods, such as rational design, homology modeling, and random mutagenesis. These mutations are then evaluated in combinatorial libraries of varying sizes and screened for activity. A small fraction of this library is sequenced, typically in the order of 3*N, where N is the number of diverse mutations. The generated sequence data then serves as the training set for the statistical analysis. In ProSAR, the statistical modeling step is based on the PLS variable regression technique,[45] which projects the sequence representations to a space of reduced dimensionality to fit a linear model.[50,51] The regression coefficients assigned to each variable represent the impact of a mutation on fitness and are used to decide whether mutations should be retained, discarded, or evaluated again in a different context.[47] Notably, it is not necessarily a priority of ProSAR to find the best variant in each round but rather to rapidly identify beneficial mutations for recombination to reach fitness targets.[16]

As delineated above, the ProSAR-driven approach focuses on parallelized, fast, and efficient iterations in short timeframes. However, not all biocatalysts can be assayed with high throughput at a large scale, and consequently it might be necessary to identify optimal sequences with minimized experimental burden. Such a case was recently described by Greenhalgh *et al.* who targeted an acyl-ACP reductase to produce fatty alcohols *in vivo*.[52] The researchers relied on only 20 sequence–function pairs to initialize an iterative process consisting of *in silico* prediction and experimental evaluations. Rather than predicting which sequences were expected to show the highest activity and evaluating only these variants, the next engineering round was built on an upper-confidence bound criterion. This criterion balances exploration and exploitation,[53,54] by simultaneously exploring areas of uncertainty within the sequence space and assessing possibly improved variants. Such an approach is particularly effective in minimizing the number of evaluations of expensive experiments.[55] The researchers iterated over ten design-test-learn cycles, sampling 10–12 sequences at each iteration, and saw gradual improvements in fatty alcohol titers, cumulating in enzymes that produce above two-fold more fatty alcohols than the wildtype sequences.[52] In our opinion, this Bayesian-type optimization nicely contrasts the ProSAR approach, highlighting how project constraints define the optimization strategy to be used.

4. ML-aided Optimization of Enzyme Stability

Of course, there are other protein properties that researchers attempt to engineer with computational methods, including enzyme stability.[56,57] Notably, a study on the ML-aided engineering of hydrolases for PET depolymerization[58] has recently managed to garner mainstream media attention. Even though more active PET degrading enzymes have previously been developed,[59] the approach is worth highlighting. The involved researchers relied on MutCompute,[60] a 3D self-supervised convolutional neural network, to predict stabilizing mutations. The neural network was trained on a large set of experimentally determined structures from the protein data bank to associate amino acids with neighboring chemical microenvironments with the goal to identify novel gain-of-function mutations.[60] MutCompute was then used to predict which amino acids are not in an optimal configuration for their local environments, effectively performing a single-site saturation scan across all residues in the protein computationally. Sites which the algorithm identified as 'abnormal' were then optimized according to predicted probabilities. This technique was applied to the PET-hydrolysing enzyme (PHE) from *Ideonella sakaiensis* (PETase),[61] and previously engineered variants ThermoPETase[62] and DuraPETase.[63] Validation of the predicted changes revealed scaffolds with improved thermostability (up to 10 °C ΔT_m compared to the respective reference variant), increased protein yield (up to 3.8 fold increase), as well as enhanced catalytic activity (up to 29 fold at selected temperatures).[58]

It should be noted that the MutCompute-type approach is quite different from the examples highlighted above. Rather than learning from a subset of the theoretically available data and predicting fitness within a defined sequence space, biological information is extracted from vast and ever-growing protein databases harnessing the fact that evolution seems to record information about structure and function into evolutionary patterns.[64] This information can be captured, to some extent, by these models and help guide decisions in downstream tasks,[65] complementing and improving the representations used to build models in other machine-learning projects.

5. Conclusion and Outlook

ML is having a notable impact on the biological sciences. Just a few years ago, determining a single protein structure could be a month to year-long process; now, structures can be predicted with similar accuracy within seconds.[64,66] As first engineering examples suggest (*vide supra*), the information contained within the vast sequence and structure datasets already collected might be able to facilitate meaningful predictions even from a few experimentally determined data points. However, not all aspects of protein engineering will benefit equally from ML. The additional costs incurred by sequencing variants, synthesizing the predicted genes, and the time and resources needed to ensure that high-quality data is being provided to train the algorithms must be weighed carefully with the advantages ML provides compared to simply combining beneficial mutations with additive effects.[67] Currently, no clear benchmarks to assess such a benefit exist, as ML accelerated protein engineering examples are scarce, and validating algorithms are restricted to only a handful of datasets.[35,60] Yet, as the field of algorithm-aided enzyme evolution is being more firmly anchored into the biocatalysis sector and gene synthesis and sequencing technologies mature further, we are confident that the *in silico* techniques will evolve into a key element to help address the numbers problem in directed evolution.

Acknowledgements

This work was created as part of NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation (Grant number 180544).

Received: December 14, 2022

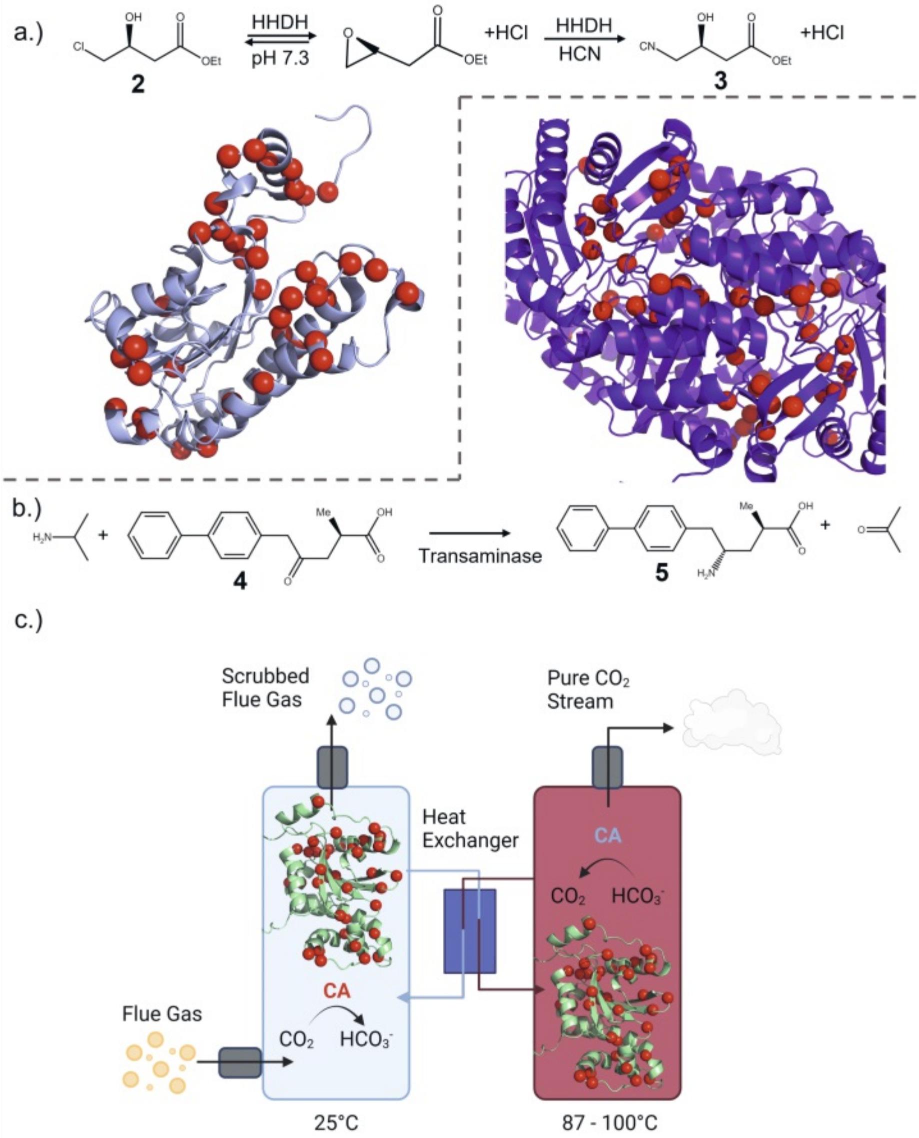

Fig. 4. Overview of successful ProSAR applications. a) HHDH catalyzes a single-vessel enzymatic conversion of ethyl (*S*)-4-chloro-3-hydroxybutyrate (**2**) to ethyl (*R*)-4-cyano-3-hydroxybutyrate (**3**). Variants with ~4,000 fold improvements over wildtype were identified after screening approximately 60,000 variants.[47] The evolved protein structure is depicted as a cartoon and mutated residues are visualized as red spheres. b) Engineering of an amine transaminase for the efficient production of (2*R*,4*S*)-5-biphenyl-4-amino-2-methylpentanoic acid (**5**), a precursor to a critical component in the blockbuster heart failure drug Entresto®. The final transaminase variant, obtained after 11 rounds of evolution, enables an economic conversion of ketone **4** with high yield and purity.[49] The evolved transaminase homodimer is shown as a cartoon with mutated residues highlighted as red spheres. c) An engineered carbonic anhydrase for efficient carbon capture from flue gas. The evolved protein, depicted in green with mutations shown as red spheres, is employed in an absorber column (blue pillar) where CO_2 chemisorbs into an amine solvent. The HCO_3^- containing amine solvent and the evolved enzyme are then transferred to a second column, where CO_2 is stripped at elevated temperatures (red pillar). The depicted carbon capture system represents one of the most challenging industrial environments applied to enzymes.[16] Image created with BioRender.com.

[1] M. Leisola, O. Turunen, *Appl. Microbiol. Biotechnol.* **2007**, *75*, 1225, https://doi.org/10.1007/s00253-007-0964-2.

[2] A. Schmid, J. S. Dordick, B. Hauer, A. Kiener, M. Wubbolts, B. Witholt, *Nature* **2001**, *409*, 258, https://doi.org/10.1038/35051736.

[3] T. J. Magliery, *Curr. Opin. Struct. Biol.* **2015**, *33*, 161, https://doi.org/10.1016/j.sbi.2015.09.002.

[4] S. Wu, R. Snajdrova, J. C. Moore, K. Baldenius, U. T. Bornscheuer, *Angew. Chem. Int. Ed.* **2021**, *60*, 88, https://doi.org/10.1002/anie.202006648.

[5] R. Buller, K. Hecht, M. A. Mirata, H. P. Meyer, in 'RSC Catalysis Series', *Vol. 2018-January*, Royal Society Of Chemistry, **2018**, pp. 3, https://doi.org/10.1039/9781782629993-00001.

[6] K. Hecht, H. P. Meyer, R. Wohlgemuth, R. Buller, *Catalysts* **2020**, *10*, 1, https://doi.org/10.3390/catal10121420.

[7] F. Arnold, *Angew. Chem. Int. Ed.* **2018**, *57*, 4143, https://doi.org/10.1002/ange.201708408.

[8] I. Victorino da Silva Amatto, N. Gonsales da Rosa-Garzon, F. Antônio de Oliveira Simões, F. Santiago, N. Pereira da Silva Leite, J. Raspante Martins, H. Cabral, *Biotechnol. Appl. Biochem.* **2022**, *69*, 389, https://doi.org/10.1002/bab.2117.

[9] S. Lutz, *Curr. Opin. Biotechnol.* **2010**, *21*, 734, https://doi.org/10.1016/j.copbio.2010.08.011.

[10] Y. Wang, P. Xue, M. Cao, T. Yu, S. T. Lane, H. Zhao, *Chem. Rev.* **2021**, *121*, 12384, https://doi.org/10.1021/acs.chemrev.1c00260.

[11] R. Blomberg, H. Kries, D. M. Pinkas, P. R. E. Mittl, M. G. Grütter, H. K. Privett, S. L. Mayo, D. Hilvert, *Nature* **2013**, *503*, 418, https://doi.org/10.1038/nature12623.

[12] F. Meyer, R. Frey, M. Ligibel, E. Sager, K. Schroer, R. Snajdrova, R. Buller, *ACS Catal.* **2021**, *11*, 6261, https://doi.org/10.1021/acscatal.1c00678.

[13] M. Eichenberger, S. Hüppi, D. Patsch, N. Aeberli, R. Berweger, S. Dossenbach, E. Eichhorn, F. Flachsmann, L. Hortencio, F. Voirol, S. Vollenweider, U. T. Bornscheuer, R. Buller, *Angew. Chem. Int. Ed.* **2021**, *60*, 26080, https://doi.org/10.1002/anie.202108037.

[14] M. T. Reetz, L. W. Wang, M. Bocola, *Angew. Chem. Int. Ed.* **2006**, *45*, 1236, https://doi.org/10.1002/anie.200502746.

[15] M. Voss, S. Hüppi, D. Schaub, T. Hayashi, M. Ligibel, E. Sager, K. Schroer, R. Snajdrova, R. M. U. Buller, *ChemCatChem* **2022**, https://doi.org/10.1002/cctc.202201115.

[16] O. Alvizo, L. J. Nguyen, C. K. Savile, J. A. Bresson, S. L. Lakhapatri, E. O. P. Solis, R. J. Fox, J. M. Broering, M. R. Benoit, S. A. Zimmerman, S. J. Novick, J. Liang, J. J. Lalonde, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 16436, https://doi.org/10.1073/pnas.1411461111.

[17] M. T. Reetz, P. Soni, L. Fernández, Y. Gumulya, J. D. Carballeira, *Chem. Commun.* **2010**, *46*, 8657, https://doi.org/10.1039/c0cc02657c.

[18] N. J. Turner, *Nat. Chem. Biol.* **2009**, *5*, 567, https://doi.org/10.1038/nchembio.203.

[19] J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 5869, https://doi.org/10.1073/pnas.0510098103.

[20] R. D. Lewis, M. Garcia-Borràs, M. J. Chalkley, A. R. Buller, K. N. Houk, S. B. Jennifer Kan, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 7308, https://doi.org/10.1073/pnas.1807027115.

[21] M. Reetz, *ChemBioChem* **2022**, https://doi.org/10.1002/cbic.202200049.

[22] B. T. Porebski, A. M. Buckle, *Protein Eng., Des. Select.* **2016**, *29*, 245, https://doi.org/10.1093/protein/gzw015.

[23] Z. Wu, S. B. Jennifer Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 8852, https://doi.org/10.1073/pnas.1901979116.

[24] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, S. Jain, in '2017 International Conference on Computer, Communications and Electronics (Comptelix)', **2017**, pp. 162, https://doi.org/10.1109/COMPTELIX.2017.8003957.

[25] F. Furtado, A. Singh, *Int. J. Res. Ind. Eng.* **2020**, *9*, 84, https://doi.org/10.22105/riej.2020.226178.1128.

[26] B. J. Wittmann, K. E. Johnston, Z. Wu, F. H. Arnold, *Curr. Opin. Struct. Biol.* **2021**, *69*, 11, https://doi.org/10.1016/j.sbi.2021.01.008.

[27] G. Li, Y. Dong, M. T. Reetz, *Adv. Syn. Catal.* **2019**, *361*, 2377, https://doi.org/10.1002/adsc.201900149.

[28] X. Feng, J. Sanchis, M. T. Reetz, H. Rabitz, *Chem. Eur. J.* **2012**, *18*, 5646, https://doi.org/10.1002/chem.201103811.

[29] F. Liang, X. J. Feng, M. Lowry, H. Rabitz, *J. Phys. Chem. B* **2005**, *109*, 5842, https://doi.org/10.1021/jp045926y.

[30] N. Shenvi, J. M. Geremia, H. Rabitz, *J. Phys. Chem. A* **2003**, *107*, 2066, https://doi.org/10.1021/jp021932n.

[31] K. W. Moore, A. Pechen, X. J. Feng, J. Dominy, V. J. Beltrani, H. Rabitz, *Phys. Chem. Chem. Phys.* **2011**, *13*, 10048, https://doi.org/10.1039/c1cp20353c.

[32] K. W. Moore, A. Pechen, X. J. Feng, J. Dominy, V. Beltrani, H. Rabitz, *Chem. Sci.* **2011**, *2*, 417, https://doi.org/10.1039/c0sc00425a.

[33] F. Cadet, N. Fontaine, G. Li, J. Sanchis, M. Ng Fuk Chong, R. Pandjaitan, I. Vetrivel, B. Offmann, M. T. Reetz, *Sci. Rep.* **2018**, *8*, https://doi.org/10.1038/s41598-018-35033-y.

[34] B. Offmann, F. Cadet, P. Charton, WO Patent Appl. No. WO2016166253A1, **2016**.

[35] J. Büchler, S. H. Malca, D. Patsch, M. Voss, N. J. Turner, U. T. Bornscheuer, O. Allemann, C. le Chapelain, A. Lumbroso, O. Loiseleur, R. Buller, *Nat. Commun.* **2022**, *13*, https://doi.org/10.1038/s41467-022-27999-1.

[36] A. Naini, F. Sasse, M. Brönstrup, *Nat. Prod. Rep.* **2019**, *36*, 1394, https://doi.org/10.1039/c9np00008a.

[37] T. Hayashi, M. Ligibel, E. Sager, M. Voss, J. Hunziker, K. Schroer, R. Snajdrova, R. Buller, *Adv. Mater.* **2019**, *131*, 18706, https://doi.org/10.1002/ANGE.201907245.

[38] M. T. Reetz, D. Kahakeaw, R. Lohmer, *ChemBioChem* **2008**, *9*, 1797, https://doi.org/10.1002/cbic.200800298.

[39] P. A. Romero, A. Krause, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, https://doi.org/10.1073/pnas.1215251110.

[40] Y. Saito, M. Oikawa, H. Nakazawa, T. Niide, T. Kameda, K. Tsuda, M. Umetsu, *ACS Synth. Biol.* **2018**, *7*, 2014, https://doi.org/10.1021/acssynbio.8b00155.

[41] T. Vornholt, F. Christoffel, M. M. Pellizzoni, S. Panke, T. R. Ward, M. Jeschek, *Sci. Adv.* **2021**, *7*, 1, https://doi.org/10.1126/sciadv.abe4208.

[42] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, *Nucleic Acids Res.* **2008**, *36*, https://doi.org/10.1093/nar/gkm998.

[43] F. Tian, P. Zhou, Z. Li, *J. Mol. Struct.* **2007**, *830*, 106, https://doi.org/10.1016/j.molstruc.2006.07.004.

[44] Z. O. Ibraheem, R. Abd Majid, S. M. Noor, H. M. Sedik, R. Basir, *Malar. Res. Treat.* **2014**, *2014*, https://doi.org/10.1155/2014/950424.

[45] R. Fox, A. Roy, S. Govindarajan, J. Minshull, C. Gustafsson, J. T. Jones, R. Emig, *Protein Eng.* **2003**, *16*, 589, https://doi.org/10.1093/protein/gzg077.

[46] R. Fox, *J. Theor. Biol.* **2005**, *234*, 187, https://doi.org/10.1016/j.jtbi.2004.11.031.

[47] R. J. Fox, S. C. Davis, E. C. Mundorff, L. M. Newman, V. Gavrilovic, S. K. Ma, L. M. Chung, C. Ching, S. Tam, S. Muley, J. Grate, J. Gruber, J. C. Whitman, R. A. Sheldon, G. W. Huisman, *Nat. Biotechnol.* **2007**, *25*, 338, https://doi.org/10.1038/nbt1286.

[48] Y. K. Bong, S. Song, J. Nazor, M. Vogel, M. Widegren, D. Smith, S. J. Collier, R. Wilson, S. M. Palanivel, K. Narayanaswamy, B. Mijts, M. D. Clay, R. Fong, J. Colbeck, A. Appaswami, S. Muley, J. Zhu, X. Zhang, J. Liang, D. Entwistle, *J. Org. Chem.* **2018**, *83*, 7453, https://doi.org/10.1021/acs.joc.8b00468.

[49] S. J. Novick, N. Dellas, R. Garcia, C. Ching, A. Bautista, D. Homan, O. Alvizo, D. Entwistle, F. Kleinbeck, T. Schlama, T. Ruch, *ACS Catal.* **2021**, *11*, 3762, https://doi.org/10.1021/acscatal.0c05450.

[50] K. K. Yang, Z. Wu, F. H. Arnold, *Nat. Methods* **2019**, *16*, 687, https://doi.org/10.1038/s41592-019-0496-6.

[51] P. Geladi, B. R. Kowalski, *Analytica Chim. Acta* **1986**, *185*, 1, https://doi.org/https://doi.org/10.1016/0003-2670(86)80028-9.

[52] J. C. Greenhalgh, S. A. Fahlberg, B. F. Pfleger, P. A. Romero, *Nat. Commun.* **2021**, *12*, https://doi.org/10.1038/s41467-021-25831-w.

[53] P. Auer, *J. Mach. Learn. Res.* **2002**, *3*, 397, https://doi.org/10.1162/153244303321897663.

[54] N. Srinivas, A. Krause, S. M. Kakade, M. Seeger, in *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3250-3265, May **2012**, https://doi.org/10.1109/TIT.2011.2182033.

[55] J. Snoek, H. Larochelle, R. P. Adams, *Adv. Neural Inf. Process Syst.* **2012**, *4*, 2951, https://doi.org/10.48550/ARXIV.1206.2944.

[56] Y. Li, D. A. Drummond, A. M. Sawayama, C. D. Snow, J. D. Bloom, F. H. Arnold, *Nat. Biotechnol.* **2007**, *25*, 1051, https://doi.org/10.1038/nbt1333.

[57] J. R. Klesmith, J.-P. Bacik, E. E. Wrenbeck, R. Michalczyk, T. A. Whitehead, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2265, https://doi.org/10.1073/pnas.1614437114.

[58] H. Lu, D. J. Diaz, N. J. Czarnecki, C. Zhu, W. Kim, R. Shroff, D. J. Acosta, B. R. Alexander, H. O. Cole, Y. Zhang, N. A. Lynd, A. D. Ellington, H. S. Alper, *Nature* **2022**, *604*, 662, https://doi.org/10.1038/s41586-022-04599-z.

[59] V. Tournier, C. M. Topham, A. Gilles, B. David, C. Folgoas, E. Moya-Leclair, E. Kamionka, M. L. Desrousseaux, H. Texier, S. Gavalda, M. Cot, E. Guémard, M. Dalibey, J. Nomme, G. Cioci, S. Barbe, M. Chateau, I. André, S. Duquesne, A. Marty, *Nature* **2020**, *580*, 216, https://doi.org/10.1038/s41586-020-2149-4.

[60] R. Shroff, A. W. Cole, D. J. Diaz, B. R. Morrow, I. Donnell, A. Annapareddy, J. Gollihar, A. D. Ellington, R. Thyer, *ACS Synth. Biol.* **2020**, *9*, 2927, https://doi.org/10.1021/acssynbio.0c00345.

[61] S. Yoshida, K. Hiraga, T. Takehana, I. Taniguchi, H. Yamaji, Y. Maeda, K. Toyohara, K. Miyamoto, Y. Kimura, K. Oda, *Science* **2016**, *351*, 1196, https://doi.org/10.1126/science.aad6359.

[62] H. F. Son, I. J. Cho, S. Joo, H. Seo, H. Y. Sagong, S. Y. Choi, S. Y. Lee, K. J. Kim, *ACS Catal.* **2019**, *9*, 3519, https://doi.org/10.1021/acscatal.9b00568.

[63] Y. Cui, Y. Chen, X. Liu, S. Dong, Y. Tian, Y. Qiao, R. Mitra, J. Han, C. Li, X. Han, W. Liu, Q. Chen, W. Wei, X. Wang, W. Du, S. Tang, H. Xiang, H. Liu, Y. Liang, K. N. Houk, B. Wu, *ACS Catal.* **2021**, *11*, 1340, https://doi.org/10.1021/acscatal.0c05126.

[64] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, Alexander Rives, *BioRxiv* **2021**, https://doi.org/10.1101/2022.07.20.500902.

[65] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, W. Yu, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *14*, 1, https://doi.org/10.1109/TPAMI.2021.3095381.

[66] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583, https://doi.org/10.1038/s41586-021-03819-2.

[67] E. J. Ma, E. Siirola, C. Moore, A. Kummer, M. Stoeckli, M. Faller, C. Bouquet, F. Eggimann, M. Ligibel, D. Huynh, G. Cutler, L. Siegrist, R. A. Lewis, A. C. Acker, E. Freund, E. Koch, M. Vogel, H. Schlingensiepen, E. J. Oakeley, R. Snajdrova, *ACS Catal.* **2021**, *11*, 12433, https://doi.org/10.1021/acscatal.1c02786.

[68] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, *Nucleic Acids Res.* **2018**, *46*, W296, https://doi.org/10.1093/nar/gky427.

[69] O. Trott, A. J. Olson, *J. Comput. Chem.* **2009**, *31*, 455, https://doi.org/10.1002/jcc.21334.

License and Terms