Advanced Synthesis & Catalysis

Accepted Article

Title: Can Machine Learning Revolutionize Directed Evolution of Selective Enzymes?

Authors: Guangyue Li, Yijie Dong, and Manfred Reetz

To be cited as: *Adv. Synth. Catal.* 10.1002/adsc.201900149

Link to VoR: http://dx.doi.org/10.1002/adsc.201900149

10.1002/adsc.201900149

VERY IMPORTANT PUBLICATION - VIP

REVIEW

DOI: 10.1002/adsc.201((will be filled in by the editorial staff))

Can Machine Learning Revolutionize Directed Evolution of Selective Enzymes?

Guangyue Li[a], Yijie Dong[a] and Manfred T. Reetz[b,c]*

[a] State Key Laboratory for Biology of Plant Diseases and Insect Pests/Key Laboratory of Control of Biological Hazard Factors (Plant Origin) for Agri-product Quality and Safety, Ministry of Agriculture, Institute of Plant Protection, Chinese Academy of AgriculturalSciences, Beijing 100081, China

[b] Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany.

[c] Fachbereich Chemie der Philipps-Universität, Hans-Meerwein-Strasse, 35032 Marburg, Germany

Correspondence should be addressed to (M.T.R.) reetz@mpi-muelheim.mpg.de

Abstract. Machine learning as a form of artificial intelligence consists of algorithms and statistical models for improving computer performance for different tasks. Training data is utilized for making decisions and predictions. Since directed evolution of enzymes produces huge amounts of potential training data, machine learning seems to be ideally suited to support this protein engineering technique. Machine learning has been used in protein science for a long time with different purposes. This mini- review focuses on the utility of machine learning as an aid in directed evolution of selective enzymes. Recent studies have shown that the algorithms ASRA and Innov'SAR are well suited as guides when performing saturation mutagenesis at sites lining the binding pocket for enhancing stereoselectivity and activity.

Keywords: machine learning; directed evolution; stereoselectivity; enzymes; saturation mutagenesis

1 Introduction

Directed evolution of stereo-, regio- and chemoselective enzymes as catalysts in organic chemistry and biotechnology constitutes a prolific source of catalysts for asymmetric transformations in organic chemistry and biotechnology.[1,2] It consists of recursive cycles of gene mutagenesis, expression and screening. The most common gene mutagenesis techniques in directed evolution are error-prone polymerase chain reaction (epPCR, a shotgun technique), saturation mutagenesis (focused randomization restricted to rationally chosen sites), and DNA shuffling (a recombinant technique). Proof of principle of directed evolution of stereoselectivity was reported in 1997, in which 4 cycles of epPCR were applied for enhancing the stereoselectivity of a lipase 10-fold.[3] However, further epPCR led only to minor improvements. Since screening for enantioselectivity is the labor-intensive step (bottleneck),[4] many studies have focused on developing efficient methods for generating high(est)-quality mutant libraries that require a minimum of analytical work.[1,2] A major step forward was the development of structure-guided generation of focused mutant libraries as a viable alternative to epPCR and DNA shuffling. Saturation mutagenesis at residues lining the binding pocket for evolving enantioselectivity (and/or activity), first reported in 2001[5a] and systematized in 2005 with the emergence of the Combinatorial Active-Site Saturation Test (CAST),[5b] has proven to be an exceptionally efficient strategy. Shortly thereafter it was extended to include Iterative Saturation Mutagenesis (ISM),[6a] useful in directed evolution of selective enzymes in general,[1,6] and also when engineering the selectivity of artificial metalloenzymes.[7] ISM has also been used in the B-FIT technique for enhancing thermostability.[6c-d] The CAST/ISM approach is a knowledge-based semi-rational strategy which is most successful if X-ray structural analyses or homology models and consensus data are used as guides, flanked by theoretical techniques such as molecular dynamics (MD) simulations and/or QM/MM computations.[1,6,7] Recent methodology developments focus on ways to further increase mutant library quality, specifically by minimizing or eliminating amino acid bias in directed evolution,[1, 2] either by novel molecular biological techniques[8a] or by means of on-chip solid-phase gene synthesis.[8b]

Guangyue Li studied protein engineering at Institute of Plant Protection, Chinese Academy of Agricultural Sciences, obtaining his master degree in 2010. He then joined the group of Prof. Dunming zhu at the Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, obtaining a doctoral degree in 2014 in the area of enzyme engineering. Thereafter, he performed postdoctoral work in the group of M. T. Reetz focusing on directed evolution of enzymes (2014-2017). Since 2017, He became a full professor at Institute of Plant Protection (IPP), Chinese Academy of Agricultural Sciences (CAAS). His main research interest is directed evolution of enzymes and research and development of the natural products from *Xenorhabdus nematophila.*

Yi-jie Dong was born in Hebei Province (China). He is currently performing his postdoctoral research at Institute of Plant Protection (IPP), Chinese Academy of Agricultural Sciences (CAAS). His main research interest is identification of the natural products from *Xenorhabdus nematophila*.

Manfred T. Reetz obtained his doctoral degree in 1969 with Ulrich Schöllkopf at Göttingen University and then performed postdoctoral work in the group of Reinhard W. Hoffmann at Marburg University. Following an appointment at Bonn University, he became Full Professor of organic chemistry back at Marburg University (1980–1991) before joining the Max-Planck-Institut für Kohlenforschung in Mülheim/Germany, where he served as Managing Director from 1993 to 2002. Following formal retirement as Director in 2011, he accepted the offer to become the first Hans-Meerwein-Research-Professor in the Chemistry Department of Marburg University while also being external (emeritus) group leader of the Mülheim Max-Planck Institute.

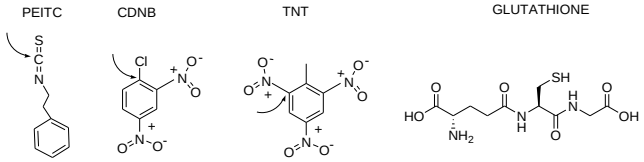
An alternative to directed evolution is rational design based on site-specific mutagenesis.[9] It has been used especially to increase protein thermostability,[9d] but less so for enhancing or inverting stereoselectivity, which appears to be a more difficult task. Moreover, increasing activity without tradeoffs in thermostability or stereoselectivity (when relevant) remains a difficult task when restricting protein engineering to rational design.

In principle, another way to improve efficacy when aiming to generate small(est) and high(est)-quality mutant libraries, either by rational design or by directed evolution, is machine learning.[10] As a type of artificial intelligence, it is actually an "old" science, which consists of algorithms and statistical models needed for improving computer performance in a variety of differently defined tasks. Accordingly, training data is used for making decisions and predictions. Today, machine learning is applied in many different areas of modern society, a research field that continues to develop at a rapid pace, as shown, inter alia, by a journal that carries its name.[10c] Whatever the goal may be when applying machine learning, success depends on the quality of the data that is fed into the system.

Directed evolution produces huge amounts of data,[1,2] an ideal situation for exploiting machine learning. However, it likewise requires high-quality data. This also applies to the use of machine learning when rationally designing protein properties. Parts of the older protein engineering literature feature studies in which the term "machine learning" is used, but also those in which it was essentially practiced, but not specified as machine learning. As early as 1992, King et al reported the use of logic-based machine learning for protein secondary structure prediction,[11] and in 2008 Shen, Bai and Vihinen published a study showing how a physicochemical feature-based classification of amino acid mutations can be used in machine learning.[12] In further studies, other important developments are listed here.[13]

In a seminal 2007 contribution, Minshull and coworkers applied machine learning in the quest to enhance the activity of proteinase K.[14] Based on sequence alignments, they chose 24 amino acid substitutions and designed 59 proteinase K variants characterized by different mutational combinations. The 59 variants were used as catalysts in the hydrolysis of a tetrapeptide following heat treatment at 68°C for 5 minutes. Sequence and activity data were then analyzed by 8 different machine learning algorithms, leading to the identification of the most potent amino acid exchanges. Then, machine learning was applied once more for making decisions which mutations to combine for optimal effects. Following 3 design cycles, a total of 95 mutants were synthesized and evaluated for activity. Some of them showed up to 20-fold activity enhancements.[14] Small libraries generated with the help of machine learning have also been reported by other groups using alternative techniques.[13]

In a recent and likewise enlightening study, Mannervik et al exploited machine learning for exploring sequence-function space of a poplar glutathione transferase.[15] Glutathione transferases catalyze the detoxification of xenobiotic substrates via conjugation with glutathione by nucleophilic attack of its thiol-moiety. A rational approach based on the use of "infologs" was chosen, which were defined as "synthetic sequences with specific substitutions capturing maximal sequence information derived from the evolutionary history of the protein family". Following the synthesis of 95 infolog genes and expression in *Escherichia coli*, the variants were tested in the reactions of phenethyl isothiocyanate (PEITC),



Scheme 1. Substrates used in protein engineering of a poplar glutathione transferase based on machine learning, the arrows indicating the points of nucleophilic substitution by the thiol-moiety of glutathione.[15]

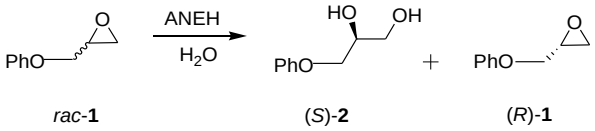
1-chloro-2, 4-dinitrobenzene (CDNB) and 2, 4, 6-trinitrotoluene (TNT) (Scheme 1).[15]

On the basis of the observed sequence-activity relationships obtained from the infologs, a second set of 47 infologs was designed. About 90% of the members showed distinctly improved properties relative to wildtype (WT). Especially two mutants, C2 (V55I/E95D/D108E/A160V) and G5 (F13L/C70A /G122E), proved to be particularly effective as catalysts in reactions of substrates PEITC and CDNB (Scheme 1) and other compounds, but not of TNT, as shown by specific activities measured in µmol/min/mg [Substrate PEITC, WT: 0.43; variant C2: 6.4; variant G5: 3.5; substrate CDNB, WT 3.32; variant C2: 17.1; variant G5: 43.1]. Expression efficiency was also improved. Interestingly, the mutations that increase activity occur at residues remote from the active site, but due to the specific design, this does not surprise.[15] Application of infologs to sites lining the binding pocket of enzymes, e.g., as an aid in controlling stereoselectivity by CAST/ISM could be a rewarding endeavor in future studies.

This Essay consists of three parts: 1) Application of the ASRA algorithm in the directed evolution of the epoxide hydrolase ANEH as catalyst in the hydrolytic kinetic resolution of a racemic epoxide. 2) Application of the innovSAR algorithm in the same directed evolution experimental platform. 3) Other applications of machine learning in protein science.

2 The ASRA Algorithm as an Aid in Directed Evolution of Stereoselectivity

As organic chemists, the Reetz group was particularly interested to see if machine learning can be of any use in protein engineering of enzyme stereoselectivity. The collaboration with Herschel Rabitz, who had previously developed the Adaptive Substituent Reordering Algorithm (ASRA),[16] was a welcome opportunity to test machine learning as a possible aid in CAST/ISM-based directed evolution of enzyme stereoselectivity.[17] For this purpose, we chose the hydrolytic kinetic resolution of glycidyl phenyl ether (*rac*-**1**) as the model asymmetric reaction, catalyzed by the epoxide hydrolase from *Aspergillus niger* (ANEH) (Scheme 2).[17] WT ANEH is slightly (*S*)-selective in favor of *(S)*-2 (*E* = 4.6). Earlier directed evolution work employing CAST/ISM had shown that

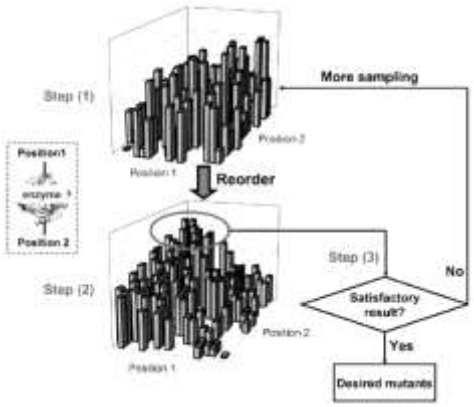


Scheme 2. ANEH-catalyzed kinetic resolution used for the purpose of testing machine learning in directed evolution of enzyme stereoselectivity.[17]

(*S*)-selectivity can be improved to *E* = 115 by the final variant LW202, but this required the screening of ~20,000 transformants using an isotopically labeled substrate (**1**) and an expensive multiplexing mass spectrometry instrument for determining enantioselectivity values.[6a] A mechanistic study encompassing X-ray data, kinetics, and molecular dynamics (MD) computations revealed the origin of enhanced enantioselectivity.[18]

ASRA was originally developed for discovering and optimizing new molecules for new properties. [16] It is independent of any structural information and therefore quite different from classical Quantitative Structure-Activity Relationship (QSAR). When extending it to directed evolution of enzyme stereoselectivity based on CAST/ISM,[17] we considered an enzyme library with N substitution positions (i.e., residues in the amino acid sequence where mutations occur) and S_i substituents (i.e., different types of amino acids, usually S_i = 20 corresponding to the 20 canonical amino acids) on the i-th position (I = 1, 2,….N). ASRA treats the property y (in the present case: enantioselectivity) of an enzyme in the library as a function of an unknown form with N independent variables $y=f(X_1,\ldots X_i,\ldots,X_N)$, in which $X_i \in [1, Si]$ is a distinct integer assigned to each substituent on the i-th position. This ensures that each protein in the mutant library is uniquely associated with an integer vector $\boldsymbol{X}=\langle X_1, \ldots, X_i, \ldots, X_N\rangle$. The collection of all variants in the library and their corresponding property values form a discrete N-dimensional property landscape, in this case an "enantioselectivity landscape". Upon choosing two substitution positions as mutation targets, *inter alia*, each amino acid at every position was assigned a random distinct integer between 1 and 20, in this case the total number of mutants being 20^2 = 400. In step 1 of the ASRA process (Scheme 3),[17] a small subset of the 400 variants characterized by substitutions on both positions were prepared, and their respective enantioselectivity measured. Due to the random integer assignment, it does not surprise that the initial selectivity landscape is irregular with little or no predictive power. In the crucial step 2, the identification of the optimal integer assignment for each amino acid at each position occurs, so that the property landscape becomes as regular as possible. Note that when an amino acid at position 1 (or position 2) "moves" by changing its integer assignment from a certain value to another one, all 20 amino acids on the other position move along with it in order to maintain consistent indexing. In step 3, the location of the best variants is predicted based on the geometric features of the re-ordered enantioselectivity landscape

10.1002/adsc.201900149



Scheme 3. The basic steps when applying ASRA to CAST/ISM-based directed evolution of enzyme stereoselectivity (see text for explanation of each step).[17]

(Scheme 3).[17] As an illustrative example, the circle can be expected to be a desired area because of the monotonic landscape geometry. Finally, the identified mutants are used in the next virtual ISM cycle by returning to step 1. The two selected CAST positions of ANEH are shown in Figure 1.[17]

The optimal reordering of the enantioselectivity landscapes as defined by *E*-values is shown in Figure 2.[17] Based on these results, a second set of 45 new mutants were generated and tested to evaluate the reliability of the ASRA predictions. Out of the chosen mutants, 34 are in the 7x7 box at the most interesting lower-right corner in the reordered landscape (Fig. 2B). The other 11 variants occur randomly in other areas of

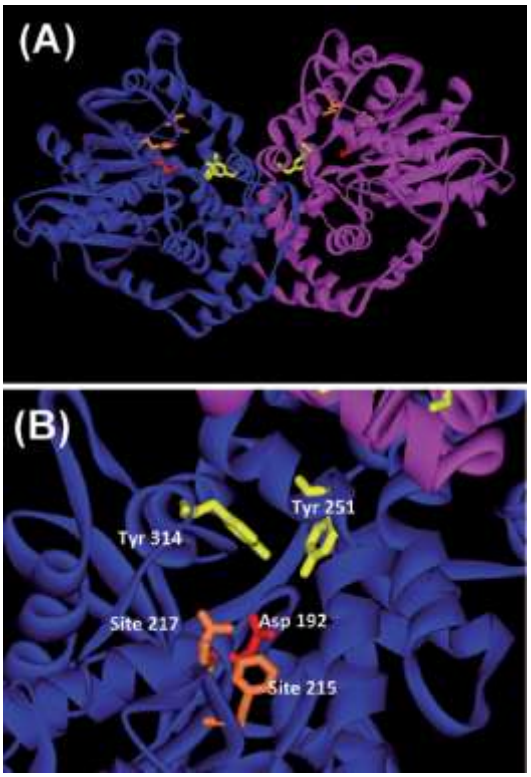


Figure 1. (A) Crystal structure of the dimer of ANEH mutant LW202. (B) Active site of ANEH featuring the catalytically active residues Asp192, Tyr251 and Tyr314, in addition to the two substitution positions 1 (residue 215) and 2 (residue 217) used in the ASRA-study.[17]

the landscape for background comparison. When these mutants are put in the *E*-value landscape employing the optimal ordering obtained from the 95 random mutants, all those with $E \geq 40$ are found inside the box (Fig. 2C). The other boxes in Figure 2 are explained by the respective legends. The results demonstrate the reliability of ASRA prediction. With a great deal of effort, we could have generated all 400 theoretically possible mutants for testing the absolute predictive power of ASRA. But it must be remembered that the purpose of ASRA is to find good regions in an enzyme library space by means of minimal sampling effort. The previous CAST/ISM study utilized not 2, but 15 residues, and 20,000 transformants had to be screened for enantioselectivity,[6a] which means that a direct comparison is difficult. All in all, only 45 new mutants were synthesized in order to ensure a sound evaluation of ASRA's prediction of good versus bad regions, which was achieved. Moreover, the selection of the new mutants demonstrates how the predictive step 3 can be performed in practice. A number of mutants were found with $E \geq 50$. In principle, further iterations of ASRA can be performed on the identified good

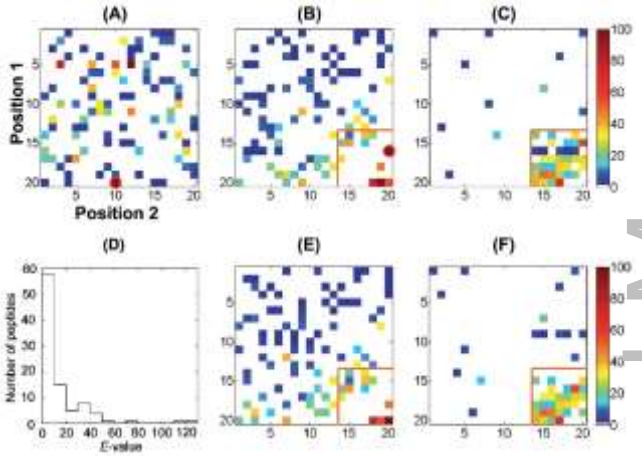


Figure 2. Optimal reordering of the *E*-value landscapes with 60 min reaction time.[17] A) Color heat map for the *E*-value landscape of 95 randomly sampled mutants plotted with a random amino acid ordering. Each color square represents one mutant with red indicating a high *E*-value and blue corresponding to a low *E*-value (see color bar on the far right). White squares are unsampled proteins. B) *E*-value landscape of the 95 mutants using the ASRA-identified optimal amino acid ordering. The result predicts that proteins with high *E*-values are most likely located in the lower right corner. The mutant at position 16/20 (circled in red in both A and B) of the reordered landscape turned out to be the same as the mutant at position 20/19; the wrong protein was accidentally placed in this position in the experiment. C) *E*-value landscape for 45 newly sampled mutants, guided by the ordering in B. D) *E*-value distribution for the 95 initial random mutants. E) Reordered *E*-value landscape for the 94 mutants (excluding the erroneous mutant at position 16/20 in B). F) *E*-value landscape for the 45 newly sampled mutants, based on the ordering in step E.

mutants in order to find the absolute best variant, if so desired. ASRA was also performed with optimized individual reaction times, likewise leading to astounding predictability. Further details can be found in the original study.[17]

In summary, the use of ASRA in CAST/ISM-based directed evolution does not require assumptions of linearity, additivity or any functional form of structure-property relationships. This machine learning approach does not need descriptors, and X-ray data is unnecessary. In fact, the only requirement is global regularity of the underlying property landscape.[17] As long as the location of each amino acid on each substitution position can be consistently indexed and followed, ASRA will be successful. It is also compatible with Pareto optimization methods for simultaneously optimizing multiple properties of the same protein library. Another positive attribute is the low cost, the major part being the limited experimental work. Nevertheless, ASRA has not been routinely used in directed evolution, possibly because some experimentally oriented biotechnologists and chemists shy away from getting acquainted with computational techniques.

3 The Innov'SAR Algorithm as an Aid in Directed Evolution of Stereo− selectivity

In our second project designed for testing machine learning as an aid in directed evolution of enzyme stereoselectivity, we again used the same experimental platform as in the ASRA-study, i.e., the ANEH-catalyzed hydrolytic kinetic resolution of epoxide **1** (Scheme 2). This time we collaborated with Frederic Cadet and coworkers, who had developed the Innov'SAR algorithm.[19]It is an innovative sequence-activity relationship approach, based on digital signal processing which combines wet-lab experimentation and computational protein design. A predictive model was developed for finding the desired enzyme property, in this case enantioselectivity as measured by the selectivity factor *E*. (Fig. 3)[20] When *n* single point mutations are permutated, 2^n combinations are possible. When applying Innov'SAR with the aim of finding highly enantioselective ANEH variants, only sequence information and the experimental results of a small set of mutants were necessary.[20] Possible epistatic interactions[21] do not prevent success, and in fact, they can be predicted by the algorithm. The 3 phases of Innov'SAR can be summarized as follows: 1) the encoding phase; 2) the modelling phase; and 3) the predictive phase. In the first phase, utilizing solely sequence data, the algorithm utilizes the idices of the AAindex database for encoding the primary protein sequence into a numerical chain, each letter of the amino acid being replaced by a letter. The data base lists >500 numerical indices corresponding to various physicochemical and biochemical properties for the 20 canoncial amino acids, and correlations between these indices are also provided. This is followed by Fast Fourier Transformation (FFT) of the encoded sequences. FFT comprises a digital signal processing technique that converts numerical signals into an energy versus frequency representation according to Eq. 1[20]:

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-2i\pi k \frac{n}{N}} \quad \text{.....Eq.1}$$

where *S* is the output spectrum (complex numbers), *s* is the input signal (encoded sequence) of length *N*, *n* is the position in the input signal, *k* is the frequency in the spectrum, and *i* is the complex number so that $i^2 = -1$. The modeling approach in Innov'SAR is based on digital signal processing using Fourier transform (FFT). Figure 3 shows a schematic representation of the Innov'SAR technique that was used in this study.[20]

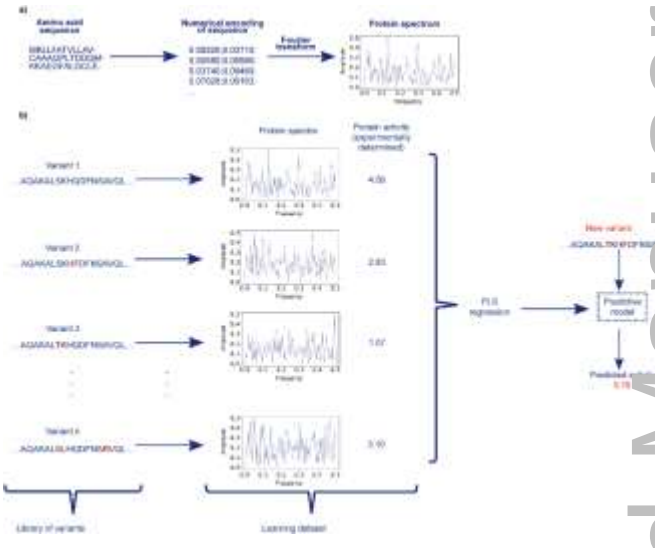


Figure 3. Schematic illustration of the Innov'SAR method as applied in the present enantioselectivity study.[20] a) A protein sequence is encoded in two steps, with a numerical encoding based on an index of AAindex dsatabase, followed by a Fast Fourier Transform for converting the encoded sequence into a protein (enzyme) spectrum; b) The different phases of Innov'SAR. An encoding phase transforms the primary sequences of the initial dataset into protein spectra. The modelling phase uses the protein spectra and the protein activity as a learning dataset in order to construct a regression model. The construction of the model is based on a partial least square regression method, PLS regression, in the modelling of the epoxide hydrolase ANEH by Innov'SAR. Then the predictive phase uses the regression model and the enzyme spectra of new ANEH variants to have their predicted activity.

When constructing the PLS (Fig. 3), the number of latent variables depends upon the chosen model. In general, the operator chooses the number of latent variables according to the minimum root mean squared error (RMSE) that is obtained in cross validation (cvRMSE). For example, in the case of [9 mutants + WT] of the present investigation, 8 latent variables are involved.[22] For building a robust PLS model, about 80 protein spectra suffice.[22] In this study, *n* = 9 single

Table 1. Performance of new ANEH mutants as catalysts in the hydrolytic kinetic resolution of *rac*-1 as predicted by Innov'SAR.

Variant	Mutations	Predicted ΔΔG‡ (kcal/mol)	Predicted *E-value*	Experimental *E-value*
WT		−1.07	6	4.6
P1	A217N_R219S_L249Y	−1.18	7	6
P2	A217N_L249Y_T317W_M329P_L330Y_C350V	−1.98	27	15
P3	L215F_A217N_R219S_L249Y_T317W_T318V_ M329P_C350V	−2.86	117	96
P4	L215F_A217N_L249Y_T317W_T318V_M329P_ C350V	−3.10	175	253
P5	L215F_A217N_R219S_L249Y_T317W_T318V_L 330Y_C350V	−3.14	185	228

CAST point mutations of ANEH were first experimentally assessed for enantioselectivity with sufficient activity, a crucial exploratory step.[20] The 9 mutants are L215F, A217N, R219S, L249Y, T317W, T318V, M329P, L330Y and C350V. It is interesting to point out that something similar is used when applying the recently improved up-to-date version of CAST/ISM, namely exploratory NNK-based saturation mutagenesis at selected CAST residues, the data then forming the basis for the design of appropriate reduced amino acid alphabets as combinatorial building blocks at multi-residue randomization sites.[1a,7] In the

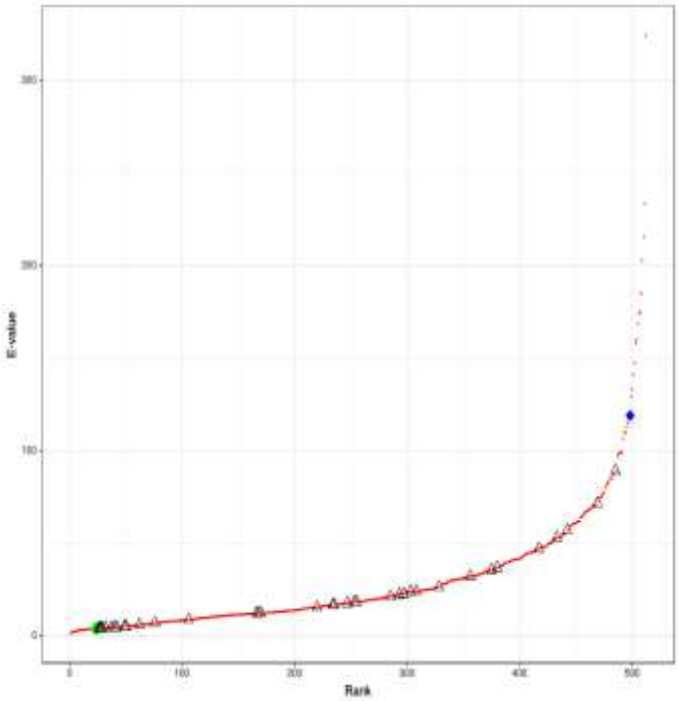


Figure 4. Ranking of the *E*-values for the 512 possible variants of ANEH as catalysts in the hydrolytic kinetic resolution of epoxide **1**. (Δ): *E*-value measured for WT and 37 single and multiple point mutants. (•): *E*-value predicted for all 512 possible variants. (♦): Best mutant previously identified: LW201(*E*=115) [6a].

Innov'SAR study, predictions were made for all combinations of the 9 single point mutations, ($2^9 = 512$ variants). Among these, a dozen variants were identified, showing notably higher enantioselectivity relative to the previous best variant LW202 ($E = 115$) evolved by CAST/ISM (Fig. 4).[20] Furthermore, *E* values of five variants randomly chosen from Innov'SAR prediction were measured by experiments, the results showing high consistent with prediction (Table. 1). These results underscore the reliability of Innov'SAR as an efficient machine learning technique. It should be noted that the steps in the Innov'SAR process constitute machine learning, and that, in principle, this technique can be applied repeatedly for guiding the individual steps of iterative saturation mutagenesis (ISM).

4 Recent Examples of Machine Learning in Protein Science with other Goals

While the focus of this mini-review concerns the utility of machine learning for engineering enzyme stereoselectivity, we emphasize that other research groups have focused on other goals.[13] As already pointed out in the Introduction (Section 1), machine learning has been applied to protein engineering for a long time with a variety of different purposes. Just a few selected examples from the more recent literature are mentioned here for illustrative purposes. One important contribution was provided by Ebrahimi and coworkers, who resorted to artificial neural networks with the aim of predicting thermostability from amino acid attributes.[23] A combination of clustering with attribute weighting formed the basis of this machine learning approach. A different goal was reached by Wang, and Tang and coworkers, who used machine learning to engineer the oleaginous yeast *Yarrowia lipolytica* in the successful attempt to produce ß-ionone on a large scale.[24] In another notable contribution,

Saito et al constructed a machine-learning-guided mutagenesis platform for engineering the properties of the green fluorescent protein (GFP).[25] In a type of virtual ISM, they combined molecular evolution with machine learning. Using 155 and 78 variants for the initial and second-round libraries, respectively, a number of mutants were discovered that showed strong yellow fluorescence.[25] It will be interesting to see whether in the future this platform can be used as a guide in CAST/ISM-based directed evolution of stereoselectivity.

Yet another goal when using machine learning was defined by Arnold and coworkers, namely designing integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization.[26] This study demonstrates that machine learning on training sequences provided by structure-guided SCHEMA recombination allows the prediction of rare sequences in a diverse library of channelrhodopsins that express and localize to the plasma membrane of mammalian cells.

5 Conclusions and Perspectives

As shown by numerous older and recent studies and reviews,[13-15, 17, 20, 23-26] machine learning[10] has reached an impressive status in protein science. This mini-review focuses on the role of machine learning in protein engineering of stereoselectivity using rational design or directed evolution. These two protein engineering techniques are already in the process of fusing,[1c,7,27] and machine learning will accelerate this process. When aiming for stereo-, regio- and/or chemoselectivity, the CAST/ISM method has emerged as a particularly effective information-based approach, [1c,2,7] and machine learning can be expected to serve as an additional guide in future work, as suggested in recent proof of concept studies.[17, 20]

Two machine learning algorithms have been used successfully in protein engineering of enzyme enantioselectivity, ASRA and Innov'SAR, but at this stage it is difficult to say which is more useful and efficient. More comparative studies are necessary before the best option can be defined. In the ASRA-study (Section 2), several steps of iterative saturation mutagenesis (ISM) were supported by the machine learning process.[17]In the Innov'SAR-study (Section 3), the immediate aim was the enhancement of enzyme stereoselectivity, which was achieved upon generating only a mini-library of mutants.[20] As an alternative, it is also possible to choose a strategy that fully combines Innov'SAR with CAST/ISM as follows: Following exploratory NNK-based saturation mutagenesis at single CAST residues in the laboratory,[7, 27b] and learning from the results, perform one round of ISM, and then utilize the data for applying Innov'SAR. A server for user-friendly application of Innov'SAR still needs to be constructed. In such an approach, platforms other than Innov'SAR and ASRA are alternatives yet to be tested. Machine learning can also be expected to be useful when aiming to engineer multiple enzyme properties simultaneously, such as stereo- and regioselectivity and activity, one of the primary on-going challenges in protein engineering.[27b, 28]

As a final assessment, it is currently unclear whether the techniques of machine learning can be further developed to a point where they will actually revolutionize directed evolution of selective enzymes as catalysts in organic chemistry and biotechnology. However, a promising start has been made in the quest to *support* either rational design or directed evolution.

Acknowledgements

MTR thanks the Max-Planck-Society for generously supporting his emeritus group 2011-2016 in Marburg. G.L. thanks the Chinese Academy of Agriculture Science for the fund of Elite Youth Program and the National Natural Science Foundation of China (*Grant No. 21807111*)

Conflict of interest

The authors declare no conflict of interest.

References

[1] Reviews of directed evolution with focus on enzyme stereoselectivity: a) M. T. Reetz, *Recent Advances in Directed Evolution of Stereoselective Enzymes*, In *Directed Enzyme Evolution: Advances and Applications*, M. Alcalde (ed), Springer, Stuttgart, **2017**; b) M. T. Reetz, *J. Am. Chem. Soc.* **2013**, *135*, 12480-12496; c) M. T. Reetz, *Angew. Chem. Int. Ed.* **2011**, *50*, 138-174.

[2] General reviews of enzyme directed evolution: a) C. Zeymer, D. Hilvert, *Annu. Rev. Biochem.* **2018**, *87*,131-157; b) M. Alcalde, *Directed Enzyme Evolution: Advances and Applications*, Springer, Stuttgart, **2017**; c) M. T. Reetz, *Directed Evolution of Selective Enzymes: Catalysts for Organic Chemistry and Biotechnology*, Wiley-VCH, Weinheim, **2016**; d) S. C. Hammer, A. M. Knight, F. H. Arnold, *Curr. Opin. Green Sustain. Chem.* **2017**, *7*, 23-30; e) M. C. Ebert, J. N. Pelletier, *Curr. Opin. Chem. Biol.* **2017**, *37*, 89-96; f) M. Goldsmith, D. S. Tawfik, *Curr. Opin. Struct. Biol.* **2017**, *47*, 140-150; g) C. P. Badenhorst, U. T. Bornscheuer, *Trends. Biochem. Sci.* **2018**, *43*, 180-198; h) H. Sun, H. Zhang, E. L. Ang, H. Zhao, *Biorg. Med. Chem.* **2018**, *26*, 1275-1284.

[3] Directed evolution of a lipase for enhancing the stereoselectivity via epPCR: a) M. T. Reetz, A. Zonta, K. Schimossek, K. Liebeton, K. E. Jaeger, *Angew. Chem. Int. Ed.* **1997**, *36*, 2830-2832; *Angew. Chem.*

1997, *109*, 2961-2963; b) M. T. Reetz, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5716-5722.

[4] Screening techniques developed for directed evolution of stereoselective enzymes: a) M. T. Reetz, *Methods Molec. Biol*. **2003**, *230*, 283-290; b) J.-L. Reymond, *Enzyme assays: high-throughput screening, genetic selection and fingerprinting*, Wiley−VCH, Weinheim, **2006**; c) M. Wójcik, A. Telzerow, W. J. Quax, Y. L. Boersma, *Int. J. Mol. Sci.* **2015**, *16*, 24918-24945; d) H. Xiao, Z. Bao, H. Zhao, *Ind. Eng. Chem. Res.* **2014**, *54*, 4011-4020; e) C. G. Acevedo-Rocha, R. Agudo, M. T. Reetz, *J. Biotechnol.* **2014**, *191*, 3-10.

[5] Early work on the design and systematization of Combinatorial Active-Site Saturation Test (CAST) for evolving enantioselectivity: a) M. T. Reetz, S. Wilensek, D. Zha, K. E. Jaeger, *Angew. Chem. Int. Ed.* **2001**, *113*, 3701-3703; *Angew. Chem.* **2001**, *113*, 3701-3703; b) M. T. Reetz, M. Bocola, J. D. Carballeira, D. Zha, A. Vogel, *Angew. Chem. Int. Ed.* **2005**, *44*, 4192-4196; *Angew. Chem.* **2005**, *117*, 4264-4268.

[6] Iterative Saturation Mutagenesis (ISM) as a useful method for evolving stereo- and/or regioselective and thermostable enzymes: a) M. T. Reetz, L. W. Wang, M. Bocola, *Angew. Chem. Int. Ed.* **2006**, *45*, 1236-1241; *Angew. Chem*. **2006**, *118*, 1258-1263; b) Y. Gumulya, M. T. Reetz, *ChemBioChem* **2011**, *12*, 2502-2510; c) M. T. Reetz, J. D. Carballeira, *Nat. Protoc*. **2007**, *2*, 891-903; d) Z. Sun, Q. Liu, G. Qu, Y. Feng, M. T. Reetz, *Chem. Rev*. **2019**, *119*, 1626-1665.

[7] Selected key studies and reviews of directed evolution of artificial metalloenzymes: a) M. T. Reetz, J. J.-P. Peyralans, A. Maichele, Y. Fu, M. Maywald, *Chem. Commun*. **2006**, 4318-4320; b) M. T. Reetz, *Acc. Chem. Res.* **2019**, 52, 336-344; c) M. Creus, A. Pordea, T. Rossel, A. Sardo, C. Lecondor, A. Ivanova, I. Letrong, R. E. Stenkamp, T. R. Ward, *Angew. Chem. Int. Ed*. **2008**, *47*, 1400-1404; *Angew. Chem*. **2008**, *120*, 1422-1427; d) F. Schwizer, Y. Okamoto, T. Heinisch, Y. Gu, M. M. Pellizzoni, V. Lebrun, R. Reuter, V. Köhler, J. Lewis, T. R. Ward, *Chem. Rev*. **2018**, *118*, 142-231; e) H. Renata, Z. J. Wang, F. H. Arnold, *Angew. Chem. Int. Ed.* **2015**, *54*, 3351-3367; *Angew. Chem*. **2015**, *127*, 3408-3426; f) A. Chandgude, R. Fasan, *Angew. Chem. Int. Ed.* **2018**, *57*, 15852-15856; *Angew. Chem*. **2018**, *130*, 16078-16082; g) J. C. Lewis, *Acc. Chem. Res.* **2019**; DOI: 10.1021/acs.accounts.8b00625.

[8] a) A. Li, C. G. Acevedo-Rocha, M. T. Reetz, *Appl. Microbiol. Biotechnol.* **2018**, *102*, 6095-6103; b) A. Li, Z. Sun, M. T. Reetz, *ChemBioChem.* **2018**, *19*, 2023-2032.

[9] Reviews of rational design based on site-specific mutagenesis: a) J. Pleiss, *Rational Design of Enzymes*, In *Enzyme catalysis in organic synthesis*, 3rd Edition, K. Drauz, H. Gröger, O. May (eds.), Wiley-VCH: Weinheim, **2012**, pp 89-117; b) T. Ema, Y. Nakano, D. Yoshida, S. Kamata, T. Sakai, *Org. Biomol. Chem.* **2012**, *10*, 6299-6308; c) K. Steiner, H. Schwab, *Comput. Struct. Biotechnol. J.* **2012**, *2*, e201209010; d) V. G. H. Eijsink, A. Bjork, S. Gaseidnes, R. Sirevag, B. Synstad, B. van den Burg, G. Vriend, *J. Biotechnol*. **2004**, *113*, 105-120.

[10] General reviews of machine learning: a) Bishop, M. Christopher, *Pattern Recognition and Machine Learning*, Springer, Stuttgart, **2006**; b) T. M. Mitchell, *Machine Learning*, McGraw Hill, New York, **1997**; c) P. Langley, *Machine Learning* **2011**, *82*, 275-279.

[11] S. Muggleton, R. D. King, M. J. Stenberg, *Prot. Eng. Des. Sel.* **1992**, *5*, 647-657.

[12] B. Shen, J. Bai, M. Vihinen, *Prot. Eng. Des. Sel.* **2007**, *21*, 37-44.

[13] Selected studies in which machine learning was used, although not always designated as such: a) R. Fox, *J. Theoret. Biol.* **2005**, *234*, 187-199; b) E. Capriotti, P. Fariselli, R. Casadio, *Nucleic. Acids. Res.* **2005**, *33*, W306-W310; c) J. Cheng, A. Randall, P. Baldi, *Proteins* **2006**, *62*, 1125-1132; d) J. Ehren, S. Govindarajan, B. Morón, J. Minshull, C. Khosla, *Prot. Eng. Des. Sel.* **2008**, *21*, 699-707; e) F. A. Buske, R. Their, E. M. Gillam, M. Bodén, *Proteins* **2009**, *77*, 111-120; f) Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, M. Rooman, *Bioinformatics* **2009**, *25*, 2537-2543; g) J. Tian, N. Wu, X. Chu, Y. Fan, *BMC bioinformatics* **2010**, *11*, 370; h) J. Liu, X. Kang, *BMC bioinformatics* **2012**, *13*, 44; i) K. S. Midelfort, R. Kumar, S. Han, M. J. Karmilowicz, K. McConnell, D. K. Gehlhaar, A. Mistry, J. S. Chang, M. Anderson, A. Villalobos, *Prot. Eng. Des. Sel.* **2012**, *26*, 25-33; j) D. E. Pires, D. B. Ascher, T. L. Blundell, *Bioinformatics* **2013**, *30*, 335-342; k) M. Giollo, A. J. Martin, I. Walsh, C. Ferrari, S. C. Tosatto, *BMC genomics* **2014**, *15*, S7; l) S. Govindarajan, B. Mannervik, J. A. Silverman, K. Wright, D. Regitsky, U. Hegazy, T. J. Purcell, M. Welch, J. Minshull, C. Gustafsson, *ACS Synth. Biol.* **2014**, *4*, 221-227; m) L. Jia, R. Yarlagadda, C. C. Reed, *PloS one* **2015**, *10*, e0138022; n) E. Jokinen, M. Heinonen, H. Lähdesmäki, *Bioinformatics* **2018**, *34*, i274-i283; o) Y. Yang, S. Urolagin, A. Niroula, X. Ding, B. Shen, M. Vihinen, *Int. J. Mol. Sci.* **2018**, *19*, 1009; p) K. K. Yang, Z. Wu, C. N. Bedbrook, F. H. Arnold, *Bioinformatics* **2018**, *1*, 7; q) M. H. Barley, N. J. Turner, R. Goodacre, *J. Chem. Inf. Model.* **2018**, *58*, 234-243.

[14] J. Liao, M. K. Warmuth, S. Govindarajan, J. E. Ness, R. P. Wang, C. Gustafsson, J. Minshull, *BMC Biotechnol.* **2007**, *7*, 16.

[15] Y. Musdal, S. Govindarajan, B. Mannervik, *Prot. Eng Des. Sel.* **2017**, *30*, 543-549.

[16] Development of the Adaptive Substituent Reordering Algorithm (ASRA): a) F. Liang, X.-j. Feng, M. Lowry, H. Rabitz, *J. Phys. Chem. B.* **2005**, *109*, 5842-5854; b) S. R. McAllister, X.-J. Feng, P. A. DiMaggio Jr, C. A. Floudas, J. D. Rabinowitz, H. Rabitz, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 5967-5970; c) N. Shenvi, J. M. Geremia, H. Rabitz, *J. Phys. Chem. A.* **2003**, *107*, 2066-2074.

[17] X. Feng, J. Sanchis, M. T. Reetz, H. Rabitz, *Chem-Eur. J.* **2012**, *18*, 5646-5654.

10.1002/adsc.201900149

[18] M. T. Reetz, M. Bocola, L.-W. Wang, J. Sanchis, A. Cronin, M. Arand, J. Zou, A. Archelas, A.-L. Bottalla, A. Naworyta, *J. Am. Chem. Soc.* **2009**, *131*, 7334-7343.

[19] Patent application by PEACCEAL, Protein Engineering Accelerator, Paris/France: N. Fontaine, F. Cadet, "Method and electronic system for predicting at least one fitness value of a protein, related computer program product". **2018**.

[20] F. Cadet, N. Fontaine, G. Li, J. Sanchis, M. N. F. Chong, R. Pandjaitan, I. Vetrivel, B. Offmann, M. T. Reetz, *Sci. Rep.* **2018**, *8*, 16757.

[21] Mini-review of additive versus more-than-additive (cooperative) mutational interactions in directed evolution: a) M. T. Reetz, *Angew. Chem. Int. Ed.* **2013**, *52*, 2658-2666; *Angew. Chem.* **2013**, *125*, 2720-2729.

[22] Private communication of F. Cadet to M. T. Reetz, 13. March 2019.

[23] M. Ebrahimi, A. Lakizadeh, P. Agha-Golzadeh, E. Ebrahimie, M. Ebrahimi, *PloS one* **2011**, *6*, e23146.

[24] J. J. Czajka, J. A. Nathenson, V. T. Benites, E. E. Baidoo, Q. Cheng, Y. Wang, Y. J. Tang, *Microb. Cell Fact.* **2018**, *17*, 136-142.

[25] A. Pustogow, A. McLeod, Y. Saito, D. Basov, M. Dressel, *Sci. Adv.* **2018**, *4*, eaau9123.

[26] C. N. Bedbrook, K. K. Yang, A. J. Rice, V. Gradinaru, F. H. Arnold, *PLOS Comput. Biol.* **2017**, *13*, e1005786.

[27] Typical studies emphasizing the fusion of directed evolution with rational design: a) H. Zhou, B. Wang, F. Wang, X. Yu, L. Ma, A. Li, M. T. Reetz, *Angew. Chem. Int. Ed.* **2019**, *58*, 764-768; *Angew. Chem.* **2019**, *131*, 774-778; b) C. G. Acevedo-Rocha, C. Gamble, R. Lonsdale, A. Li, N. Nett, S. Hoebenreich, J. B. Lingnau, C. Wirtz, C. Fares, H. Hinrichs, A. Deege, A. J. Mulholland, Y. Nov, D. Leys, K. J. McLean, A. W. Munro, M. T. Reetz, *ACS Catal.* **2018**, *8*, 3395-3401.

[28] G. Li, H. Zhang, Z. Sun, X. Liu, M. T. Reetz, *ACS. Catal.* **2016**, *6*, 3679-3687.

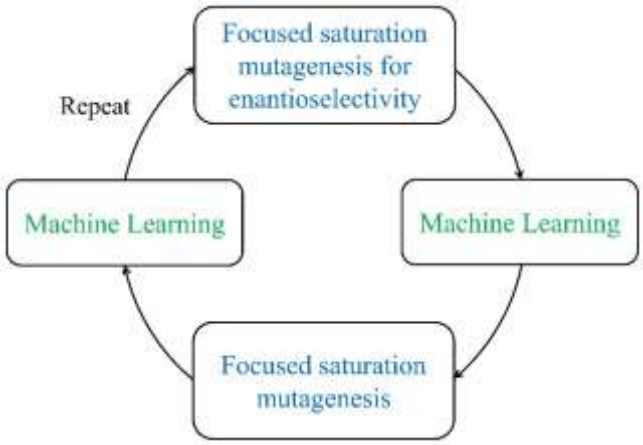
10.1002/adsc.201900149

REVIEW

Can Machine Learning Revolutionize Directed Evolution of Selective Enzymes?

Adv. Synth. Catal. **Year**, *Volume*, Page – Page

Guangyue Li[a], Yijie Dong[a] and Manfred T. Reetz[b,c]*



Never stop learning: Machine learning as a form of artificial intelligence is an excellent aid when performing directed evolution of stereoselective enzymes based on focused saturation mutagenesis.