

# SCIENTIFIC REPORTS



OPEN

## A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes

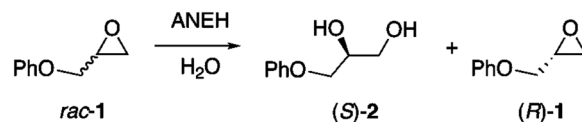
Frédéric Cadet<sup>1</sup>, Nicolas Fontaine<sup>1</sup>, Guangyue Li<sup>2</sup>, Joaquin Sanchis<sup>3</sup>, Matthieu Ng Fuk Chong<sup>1</sup>, Rudy Pandjaitan<sup>1</sup>, Iyanar Vetrivel<sup>1</sup>, Bernard Offmann<sup>4</sup> & Manfred T. Reetz<sup>2,5</sup>

Directed evolution is an important research activity in synthetic biology and biotechnology. Numerous reports describe the application of tedious mutation/screening cycles for the improvement of proteins. Recently, knowledge-based approaches have facilitated the prediction of protein properties and the identification of improved mutants. However, epistatic phenomena constitute an obstacle which can impair the predictions in protein engineering. We present an innovative sequence-activity relationship (innov'SAR) methodology based on digital signal processing combining wet-lab experimentation and computational protein design. In our machine learning approach, a predictive model is developed to find the resulting property of the protein when the  $n$  single point mutations are permuted ( $2^n$  combinations). The originality of our approach is that only sequence information and the fitness of mutants measured in the wet-lab are needed to build models. We illustrate the application of the approach in the case of improving the enantioselectivity of an epoxide hydrolase from *Aspergillus niger*.  $n=9$  single point mutants of the enzyme were experimentally assessed for their enantioselectivity and used as a learning dataset to build a model. Based on combinations of the 9 single point mutations ( $2^9$ ), the enantioselectivity of these 512 variants were predicted, and candidates were experimentally checked: better mutants with higher enantioselectivity were indeed found.

Directed protein evolution is a relatively tedious and time-consuming endeavour. Originally starting from purely random mutagenesis approaches<sup>1,2</sup>, protein engineering has advanced to a more and more information driven effort (site directed, structure based)<sup>3-7</sup>. Directed evolution of stereoselectivity continues to be a central issue of significant importance in organic and medicinal chemistry as well as biotechnology. Such techniques as rationally chosen saturation mutagenesis at sites lining the binding pocket as part of the combinatorial active-site saturation test (CAST)<sup>8</sup> and iterative saturation mutagenesis using reduced amino acids for minimizing the screening effort<sup>9-13</sup> constitute important advances. Nevertheless, the screening effort remains the bottleneck of directed evolution of stereo- and regioselectivity, which calls for support and guidance by in silico techniques.

Machine learning algorithms in protein science were developed as early as 1992, in that case for secondary structure prediction<sup>14</sup>. Thereafter new versions of machine learning for predicting structure, folding, binding, and even catalytic activity appeared with the aim of processing the accumulating information about mutants and their properties<sup>15-21</sup>. The “big data” serves as training set for these algorithms to facilitate the prediction of new and improved variants, thereby aiding experimental efforts in protein engineering based on site-specific mutagenesis or directed evolution<sup>6,22</sup>. However, these methods are mainly based on addition of the activities of the characterized single mutations. Hence, the non-additivity of functional mutations can lead to inaccurate identification of

<sup>1</sup>PEACCEL, Protein Engineering Accelerator, Paris, France. <sup>2</sup>Department of Chemistry, Philipps-University, 35032, Marburg, Germany. <sup>3</sup>Faculty of Pharmacy and Pharmaceutical Sciences, Monash University, Parkville, Australia. <sup>4</sup>UFIP, UMR 6286 CNRS, UFR Sciences et Techniques, Université de Nantes, Nantes, France. <sup>5</sup>Max-Planck-Institut fuer Kohlenforschung, 45470, Mülheim, Germany. Correspondence and requests for materials should be addressed to F.C. (email: [frederic.cadet@peaccel.com](mailto:frederic.cadet@peaccel.com))



**Figure 1.** Hydrolytic kinetic resolution of an epoxide (*rac*-1) catalysed by the epoxide hydrolase from *Aspergillus niger* (ANEH).

the best performing engineered protein using *in silico* approaches. Epistasis phenomena can impair the predictions in protein engineering and screening. In this sense, machine learning has not been applied in order to evolve enzyme mutants with enhanced or inverted stereoselectivity.

Digital Signal Processing (DSP) techniques are analytic procedures, which decompose and process signals in order to reveal information embedded in them<sup>23</sup>. The signals may be continuous (unending) or discrete such as the protein residues. In proteins, Fourier transform methods have been used for: biosequence (DNA and protein) comparison<sup>24</sup>, characterization of protein families and pattern recognition<sup>25–27</sup>; classification and other structure based studies such as analysis of symmetry and repeating structural units or patterns, prediction of secondary/tertiary structure, prediction of hydrophobic core motifs, conserved domains, prediction of membrane proteins<sup>28–31</sup>, prediction of conserved regions<sup>32</sup>, prediction of protein subcellular location<sup>33</sup>, for the study of the secondary structure content in amino acids sequence<sup>34</sup> and for the detection of periodicity in protein<sup>35</sup>. More recently new methods for the detection of solenoids domains in protein structures were proposed<sup>36,37</sup>.

Cosić has developed the most known approach using DSP and called Resonant Recognition Model (RRM). Digital Signal Processing techniques have helped analyse protein interactions<sup>26,38</sup> and made biological functionalities calculable. In these approaches, protein residues are first converted into numerical sequences using one of the available AAindex from this database<sup>39,40</sup>, representing a biochemical property or physicochemical parameter for each amino acid. These numerical sequences are then processed by means of Discrete Fourier Transform (DFT) to present the biological characteristics of the proteins in the form of Informational Spectrum Method (ISM)<sup>41</sup>. ISM procedure has been used to investigate principal arrangement in Calcium binding protein<sup>25</sup> and Influenza viruses<sup>42,43</sup>. A variant of the ISM, which engages amino acids parameter called Electron-Ion Interaction Potential (EIIP) is referred as Resonant Recognition Model (RRM). In this procedure, biological functionalities are presented as spectral characteristics. This physico-mathematical process is based on the fact that biomolecules with same biological characteristics recognize and bio-attach to themselves when their valence electrons oscillate and then reverberate in an electromagnetic field<sup>26,44</sup>. RRM involves four steps<sup>45</sup>: (i) the conversion of the Protein Residues into Numerical Values of EIIP Parameter; (ii) a zero-padding/up-sampling (iii) the generation of protein spectrum using Fast Fourier Transform (FFT). FFT processed by means DFT to yield Spectral Characteristics (SC) and point-wise multiplied to generate the Cross Spectral (CS) features during the last step. (iv) Cross-Spectral (CS) analysis represents the point-wise multiplication of the Spectral Characteristics (SC). In this approach, a consensus spectrum (defined as a CS of a large group of sequences that share one or more common biological functions) is the final outcome of the method and the starting point for spectral characterization of protein families<sup>46</sup>.

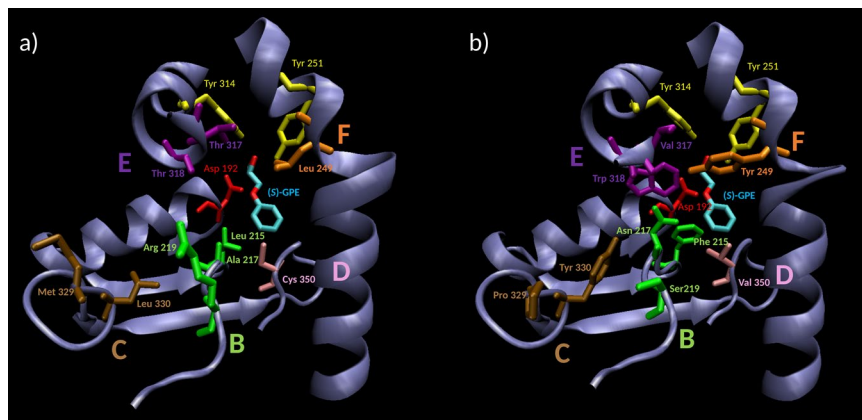
But up to now, the energy spectra obtained after FFT has never been used to go through statistical modelling and predict the effect of mutations on the fitness of an amino acid sequence. The energy spectra have never been used to explore protein sequence-activity or protein sequence/fitness relationship. We propose a new approach based only on the amino acids sequence and using digital signal processing (FFT) and protein spectrum to modelling and predict the biological activity/fitness and for identifying combination of single points amino acid substitutions with improved fitness. There is no report of such method to the best of our knowledge.

Recently, the machine-learning innov'SAR (innovative Sequence-Activity Relationship) methodology appeared in the patent literature, a structure independent mutant library screening approach developed by PEACCEL<sup>47</sup>. It relies on the representation of proteins as spectra based on the physico-chemical properties of the amino acids that constitute the protein. The protein spectrum that comes out from FFT treatment is the starting point.

In the current study, we demonstrate the potential of innov'SAR methodology (a form of artificial intelligence) in efficiently identifying enantioselective mutants of the epoxide hydrolase from *Aspergillus niger* (ANEH). We also investigate innov'SARs' capacity to predict mutational epistasis and to identify mutants with improved biological activity.

Epoxide hydrolases (EC 3.3.2.3) are enzymes that catalyse the hydrolytic kinetic resolution of racemic epoxides or the desymmetrisation of meso-epoxides, resulting in the formation of the respective chiral diols in enantiomerically enriched or pure form. Epoxide hydrolases have implications in both the fine chemical and pharmaceutical industries<sup>48</sup>. Sometimes wildtype (WT) epoxide hydrolases are highly stereoselective. In those many cases in which enantioselectivity is poor, directed evolution has been applied successfully, especially using saturation mutagenesis at sites lining the binding pocket (CASTing) and if necessary iterative saturation mutagenesis. For example, the group of Reetz carried out studies on the 398-residue ANEH<sup>49–51</sup> that is a biocatalyst for the enantioselective hydrolytic kinetic resolution of glycidyl phenyl ether (*rac*-1, GPE), (Fig. 1). The WT ANEH has an enantioselectivity factor (E-value) of 4.6, in slight favour of the (S)-2 enantiomer.

In the CAST-study aimed at enhancing the enantioselectivity in the reaction of *rac*-1, the ANEH crystal structure served as a guide for identifying randomization sites used in saturation mutagenesis. Five sites (B-F) were chosen for combinatorial randomization using NNK codon degeneracy encoding all 20 canonical amino acids, inducing in each saturation mutagenesis cycle either single, double or triple point mutations (Fig. 2 and Table S1).



**Figure 2.** Active site of ANEH. ANEH catalytic triad comprises of Tyr 251 and 314 (in yellow), which orientates the GPE substrate *rac-1*, and Asp 192 (in red), responsible for the nucleophilic GPE-ring opening. After a CAST evolutionary process, 9 amino acids from ANEH WT (a), gathered in sites B–F, were mutated to yield LW202 (b).

After 5 iterative rounds of saturation mutagenesis, a total of nine single point mutations led to an improved variant (LW202) with an E-value of 115. Even for this successful approach, this improvement of 25-fold involved at the time a tedious screening effort of around 20000 clones. Although CASTing and iterative saturation mutagenesis (ISM) have since been improved requiring notably less screening<sup>52</sup>, we were interested in exploring the possibility of applying machine learning for efficient directed evolution of enantioselectivity using the same model reaction.

To develop a really useful sequence-based statistical predictor for a biological system, one should observe the following five steps<sup>53</sup>: (i) how to construct or select a valid benchmark dataset to train and test the predictor; (ii) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) how to properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) and as far as possible, how to establish a user-friendly web-server for the predictor that is accessible to the public. Below, we describe how to deal with these steps one-by-one.

## Results

### Modelling approach based on Digital Signal Processing using Fast Fourier Transform (FFT).

innov'SAR requires only, as input data, an initial dataset with the primary sequences of protein variants and their values for a biological activity, to generate a predictive model. This model can be used to find the activity of new mutants, outside of the initial dataset. In contrast to many other rational methods used to find new interesting mutants, innov'SAR can be used without knowledge of the 3D structure. Therefore, innov'SAR can be used for proteins where the crystal structure could not be obtained. 3D structural data could still be used to validate the results from innov'SAR.

innov'SAR consists of 3 phases, namely, the encoding phase, the modelling phase and the predictive phase.

In the first phase, the encoding phase, innov'SAR must encode the alphabetic protein sequence into a numerical sequence, understandable by the modelling tools. This phase takes into account only the protein sequences.

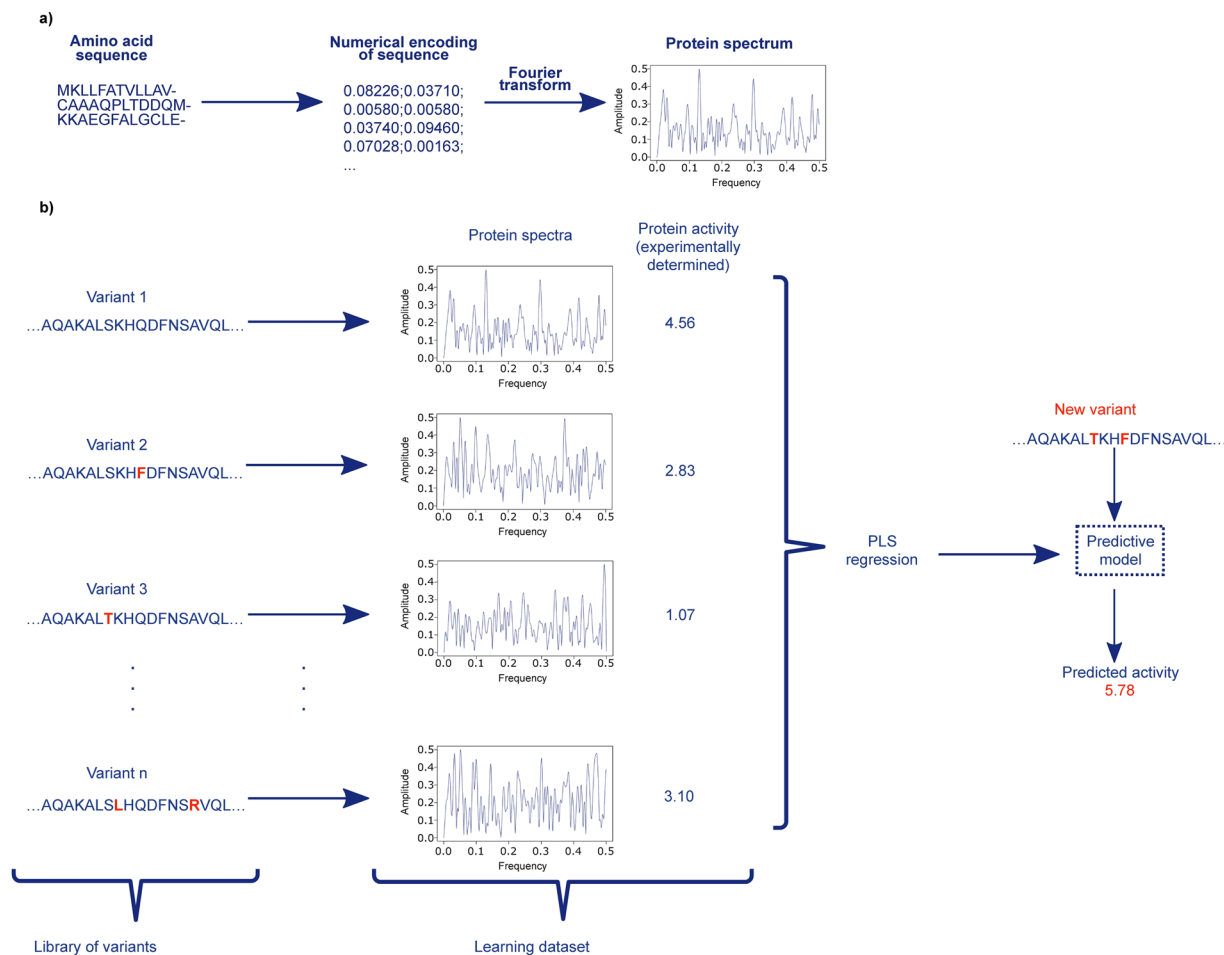
innov'SAR uses two steps for the encoding (Fig. 3). First, innov'SAR uses the indexes of the AAindex database [20] to encode the primary protein sequence into a numerical chain, where each letter of amino acid is replaced by a value. This database holds more than 500 numerical indices representing various physicochemical and biochemical properties for the 20 standard amino acids and correlations between these indices are also listed.

The second step comprises a Fast Fourier Transformation (FFT) of the encoded sequences from the first step. FFT is a digital signal processing technique that is used to convert numerical signals into an Energy versus frequency representation (equation 1). After this step, a spectral form of the protein, called the protein spectrum, is generated. The use of FFT and *protein spectra* are the cornerstones for the prediction of biological activity by innov'SAR approach.

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

With  $s$  the input signal (encoded sequence) of length  $N$ ,  $S$  the output spectrum (complex numbers),  $n$  the position in the input signal,  $k$  the frequency in the spectrum and  $i^2 = -1$ .

A representation of a protein spectrum is shown in Fig. 3a as a plot  $Energy = f(\text{frequency})$ . The protein spectrum allows to take into account the impact of mutations on the whole spectrum and does not focus on local fitness alone. Therefore, a single point mutation impacts the whole protein spectrum in a similar fashion a single point mutation can impact the whole structure of a protein. At the end of the encoding phase all the variants from the initial dataset will have a protein spectrum.



**Figure 3.** Schematic illustration of innov'SAR methodology. (a) A protein sequence is encoded in two steps: i. with a numerical encoding based on an index of AAindex database, i.e. a Fast Fourier transform is applied to convert the encoded sequence into a protein 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 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 by innov'SAR. Then the predictive phase uses the regression model and the protein spectra of new variants to have their predicted activity.

Only the next phase, the modelling phase, will use the experimental values of the target activity, in conjunction with these protein spectra, in order to identify a predictive model (Fig. 3b). The model is constructed by the application of standard regression approaches based on a learning step and a validation step. innov'SAR used a partial least square regression, PLS, as algorithm of regression to do the model for the predictions of the enantioselectivity of epoxide hydrolase. The protein spectra and the values of the activity are inputs of this regression method for the construction of a model. The goal of the regression method is an attempt to learn and to analyse possible associations between frequencies of protein spectra and the activity values, during the learning step. This learning step leads to the construction of a model. The validation step consists to test the accuracy of the model in order to check if the learning step was efficient. The root mean squared error (RMSE) and the coefficient of determination ( $R^2$ ) are the performance parameters to assess a regression model, during the validation step. RMSE values varies between 0 and  $+\infty$ .  $R^2$  value varies between 0 and 1. An accurate regression model has an RMSE close to 0 and a  $R^2$  close to 1.

One particularity of innov'SAR approach, in the modelling phase, is to evaluate multiple encoding indexes to find the best for the construction of models. innov'SAR uses the initial dataset (training set) to construct a predictive model for each encoding index. For each model, innov'SAR calculates the value of the performance parameters in two stages. The first stage is a standard cross validation. The next stage is a modelling integrating the full set in the learning step. The performances from the two stages are analysed to evaluate and to check the robustness and the validity of a model. In the first stage, the cross-validation stage, the initial dataset is split into  $k$  equal portions. The number  $k$  varies according to the size of the initial dataset. We use low  $k$  value if the dataset size is high and high  $k$  value in the opposite case. We use  $k-1$  portions as the learning dataset and the remaining one as the test dataset. The procedure is repeated  $k$  more times until each portion is used as the testing dataset

once. The cross-validation allows to avoid potential overfitting problem and to optimize some modelling parameters. The method of cross-validation used for this study is the Leave-One-Out Cross-Validation (LOOCV), where  $k$  is equal to the number of sequences.

In the second stage, the full set stage, the whole initial dataset is used as a learning dataset and a test dataset will be tested with the optimized parameters from the first stage. This stage checks the accuracy of the predictions for learned sequences.

At the end of the modelling phase, a set of accurate models and their associated encoding indexes are selected and kept.

In the predictive phase, the sequences of the new variants are pre-treated by an encoding phase with a selected encoding index, determined in the modelling phase. Once all the new variants have a protein spectrum, innov'SAR employs a model associated to the encoding index and selected in the modelling phase. Next, the model predicts the values of the activity of the new variants from their protein spectrum (see Fig. 3b).

All steps of the innov'SAR approach were implemented on a workstation equipped with Intel(R) Xeon(R) E5-2650 v4 2.20 GHz processor and 16 GB of RAM. Using this hardware, we were able to handle up to 30 single mutations, generate all combinations of mutations ( $2^{30}$ ) and predict their values of the activity in less than 48 hours. The developed algorithm made use of all of the 12 cores of the CPU to speed up the calculation, but the memory was the main limiting factor to generate and predict the activity of new variants. This methodology can scale to larger machines like High Performance Computing (HPC) clusters to either reduce the computation time or to increase the number of mutations that can be handled at once and ultimately the number of new variants that can be predicted.

**Modelling of the training set of ANEH with innov'SAR methodology.** *Prediction of enantioselectivity with the  $\Delta\Delta G^\ddagger$  for multiple point mutants using single point mutations or recombinations thereof.* Like in most machine learning approaches, innov'SAR needs a training set to learn the correlation between sequence space and experimental values. We wanted to know if it is possible to establish a predictive model with a train set containing only few single point mutations, i.e., without any combinations of mutations allowing to capture epistasis. As noted above, in the original 2006-study, nine single point mutations of ANEH were experimentally evolved and assessed for their impact on the enantioselectivity<sup>49</sup>.

The E-value can be transformed in  $\Delta\Delta G^\ddagger$  (kcal/mol) by the relation  $\Delta\Delta G^\ddagger = -RT \ln(E)$ . We used the  $\Delta\Delta G^\ddagger$  values for the construction of innov'SAR models, in order to follow and determine the enantioselectivity of the ANEH variants of this study. The energy of the 9-single point mutations and the WT constitute a first learning dataset of protein sequences that was used by innov'SAR and named **dataset A** (Table S2).

A first model, named **model DSA\_FFT**, was generated with **dataset A** to predict the  $\Delta\Delta G^\ddagger$  of double, triple and quadruple point mutants and to appraise the modelling done by innov'SAR with only single point mutants. After the evaluation of multiple encoding indexes by innov'SAR, described previously, the encoding index ZHOH040101<sup>54</sup> was identified as one of the best to generate an accurate model by innov'SAR and selected to build the **model DSA\_FFT**.

The  $\Delta\Delta G^\ddagger$  values of all possible combinations of these single point mutations,  $2^9 = 512$  possible variants, were thus predicted. We created a validation dataset with the 28 multiple point mutants (Table S3), including the mutant LW202, that were previously reported (among a total of 512)<sup>50</sup>, in order to check the accuracy of our models based on only 9 single point mutations. The high value of the coefficient of determination,  $R^2 = 0.81$  from the prediction of the validation dataset, indicates the good accuracy of our model (Fig. 4).

The model uses a learning dataset based only on single point mutations, with relatively high value of  $\Delta\Delta G^\ddagger$ , i.e., low values of enantioselectivity. However, it can determine the multiple point mutants with high experimental values of E-value such as the mutant LW202.

The incorporation of the 28-multiple mutation variants, including the mutant LW202, into the dataset A should improve the accuracy of the model for the prediction of the 512 possible variants. This incorporation formed a new dataset, named **dataset B**, by gathering all the 37 mutants and the WT (Table S4). A new model, named **model DSB\_FFT**, was constructed with the **dataset B**. After the evaluation of multiple indices from the AAindex database, the index RACS820104<sup>55</sup> was used as encoding index due to its higher modelling performance on the new learning dataset.

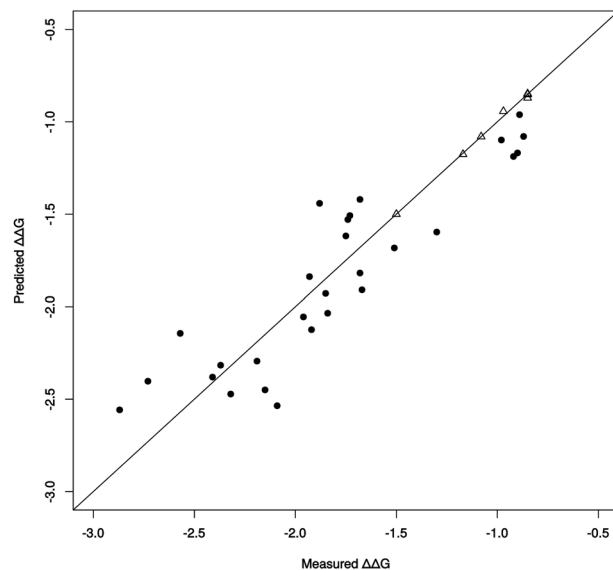
We compared the performances of the **model DSA\_FFT** and the **model DSB\_FFT** for the prediction of the 28 multiple point mutants. As the **model DSB\_FFT** used already the 28 multiple mutants in the learning step, we performed a Leave-One-Out Cross-Validation (LOOCV) for the validation of the second model. In the LOOCV, the predicted value of each mutant was calculated by removing the mutant from the learning dataset.

We obtained an excellent  $R^2$  value (0.96) for the LOOCV of the **model DSB\_FFT**, based on the multiple point mutants. Hence, the addition of multiple point mutants to the learning dataset includes some information about the epistasis between mutations and thus improves the accuracy of the prediction model (Fig. 5).

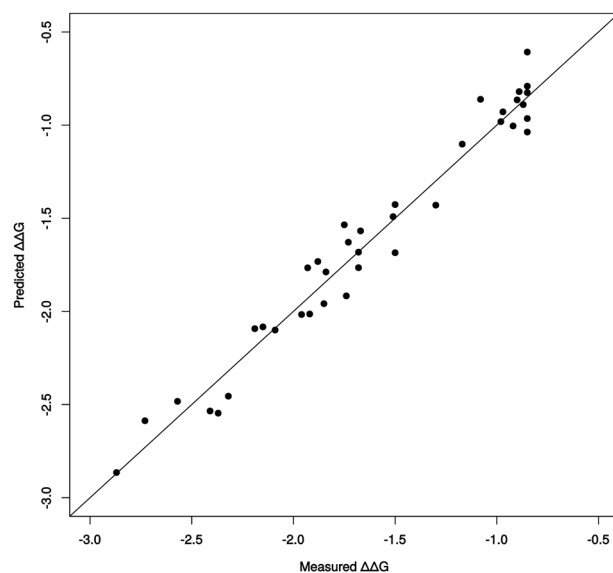
*Analysis of mutants demonstrating epistatic interactions.* It was reported that the recombination of mutations resulted in various non-linear improvements in the desired properties of the mutant ANEH<sup>51</sup>. This cooperativity reflects the effect of epistatic interactions between mutations of ANEH. In a mutant demonstrating epistasis, the association of two or several mutations may result in a different activity than the one resulting from pure addition of each single mutation<sup>56–59</sup>. The effect of combining mutations can be either an addition of mutations where the resulting activity represents the sum of the single activities, or a positive epistasis, with an increase of activity compared to the addition, or a negative epistasis with decreased activity compared to the addition<sup>56–59</sup>.

Hence, it is crucial to determine the capacity for identifying these epistatic interactions of any predictive algorithm, so that mutants with positive epistasis or negative epistasis can be distinguished. We therefore first calculated the theoretical value of  $\Delta\Delta G^\ddagger$ , for the case of pure additivity of mutations, for each of the 28 recombinants,





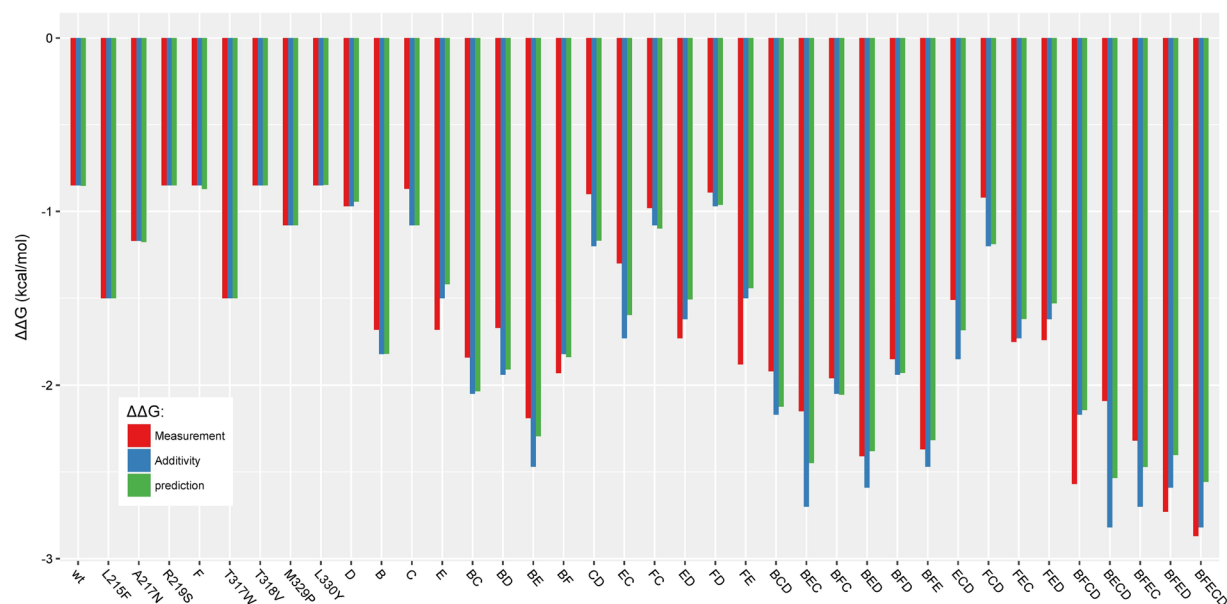
**Figure 4.** Predictions from model DSA\_FFT based solely on experimental values of 9 single point mutants and the WT of ANEH as learning set. The  $\Delta$  points are the variants for the learning set, the 9 single point mutants and the WT, already learned by the model. The  $\bullet$  points are the variants for the validation set, comprising the 28 multiple point mutants, not learned by the model.  $R^2$  for the learning dataset ( $\Delta$ ):  $R^2 = 0.99$ .  $R^2$  for the validation dataset ( $\bullet$ ):  $R^2 = 0.81$ .



**Figure 5.** Leave-One-Out Cross-Validation (LOOCV) predictions from model DSB\_FFT, based on all the 37 single and multiple point mutants, including LW202 mutant, and the WT of ANEH as learning set. LOOCV performances of prediction:  $R^2 = 0.96$ .

by simply adding the corresponding values of the single mutations (Table S4). The comparison between these theoretical additive values and the measured values allow the evaluation of the impact of epistatic interactions, while revealing whether positive epistasis or negative epistasis is involved. It turned out that all of the 28 variants with multiple mutations have a positive or a negative epistasis: 9 with positive epistasis and 19 with negative epistasis (Table S4).

After the calculation of the theoretical values, we used them to evaluate the accuracy of predictions based only on the addition of mutations to find the  $\Delta\Delta G^{\ddagger}$  of the 28 recombinants. We obtained a  $R^2$  of 0.80. We already know that the presence of epistasis prevents a perfect accuracy if we consider only the addition of mutations. But for these 28 variants, the predictions based on the addition of mutations provide already a correct accuracy. The **model DSA\_FFT**, based on single point mutants, has a performance of predictions ( $R^2 = 0.81$ ) close to the performance when the addition of mutations is used. The performances are close, but our model is not based only on



**Figure 6.** Predictions of  $\Delta\Delta G^\ddagger$  values for multiple point mutation mutants from model DSA\_FFT, based on the WT and the 9 single point mutants. Experimental values are in red, theoretical values for predictions based only on addition of mutations are in blue, and innov'SAR predictions are in green.

the addition of mutations. Indeed, in a next step we compared the predicted  $\Delta\Delta G^\ddagger$  of the 28 recombinants based on only the addition of mutation and the predicted  $\Delta\Delta G^\ddagger$  of the **model DSA\_FFT** (Fig. 6).

Among the 28 multiple point mutants, the model DSA\_FFT gives better results for 18 mutants than the prediction based only on the addition of mutations (Fig. 6 and Table 1). The innov'SAR model allows a distinct capture of epistasis for these mutants, which is a powerful characteristic of this machine learning approach.

We then tested if **model DSB\_FFT**, with the incorporation of the 28-multiple point mutants into the 9-single point mutant model could bring improvements to the prediction of epistasis for the other mutants where the **model DSA\_FFT** fails to fully capture epistasis. Among the 28 multiple point mutants, the **model DSB\_FFT** generates more accurate predictions for 22 mutants, better results than the one based only on the addition of mutations (see Fig. 7 and Table 1). An improvement is observed in the ability of innov'SAR to capture epistatic effects compared to the use of only single mutants in the learning dataset. As already pointed out, this fact is also shown by the higher  $R^2$  value (0.96) for the LOOCV training set. The LOOCV shows that the model identifies more epistatic effects from the learning dataset.

*The generation of protein spectra is a key feature of innov'SAR methodology.* To get a better idea why innov'SAR approach could predict epistatic effects, we ran different kinds of modelling, with either single and/or multiple mutants and with or without FFT during the encoding phase of innov'SAR (see Fig. 3a).

Models were built without applying FFT during the encoding phase. When a model based on **dataset A** and without FFT, named **model DSA\_noFFT**, the  $\Delta\Delta G^\ddagger$  predictions perfectly fit with the pure additive values (Fig. 8). In this case, we obtained the same performance as when only the addition of mutations is used to predict the value of  $\Delta\Delta G^\ddagger$  ( $R^2 = 0.80$ ). Without FFT, the **model DSA\_noFFT** supposes only the addition of mutation and could not reproduce the epistasis for the prediction.

Next, we then performed the modelling with the dataset B and without FFT to generate the **model DSB\_noFFT**. This model has 0.88 as value of  $R^2$ . Figure 9 shows that the **model DSB\_noFFT** can predict some epistatic interactions, but with lower accuracy than the **model DSB\_FFT**, using FFT in the encoding phase (Fig. 7). Finally, as shown in Fig. 7, a model including FFT improves the accuracy of the predictions of the epistatic interactions of the multiple point mutations. *FFT is an improving factor for the predictions of mutants with epistasis in the models generated by innov'SAR.*

*Prediction of new improved ANEH mutants.* The best experimental mutant previously described<sup>49</sup>, named LW202, contains all 9 single point mutations in its sequence and has an E-value of 115. In an attempt to identify better mutants, we generated computationally all combinations of the 9 single point mutations ( $2^9$  variants). This approach resulted in  $2^9 - 38 = 474$  new variants with multiple point mutations. Figure 10 shows the predictions of E-value for all recombinants from the **model DSB\_FFT** and our model was able to identify candidates with better predicted enantioselectivity than LW202.

**Experimental characterization of predicted ANEH mutants.** In a next step, we made an arbitrary selection of the predicted improved mutants and tested them in wet lab for enantioselectivity. The predicted and experimental values are listed in Table 2.

<i>Mutant</i>	model DSA_FFT	model DSB_FFT
<i>B</i>	I	I
<i>C</i>	I	I
<i>E</i>	A	I
<i>BC</i>	I	I
<i>BD</i>	I	I
<i>BE</i>	I	I
<i>BF</i>	I	A
<i>CD</i>	I	I
<i>EC</i>	I	I
<i>FC</i>	A	I
<i>ED</i>	A	I
<i>FD</i>	I	I
<i>FE</i>	A	I
<i>BCD</i>	I	I
<i>BEC</i>	I	I
<i>BFC</i>	A	I
<i>BED</i>	I	I
<i>BFD</i>	I	A
<i>BFE</i>	I	A
<i>ECD</i>	I	I
<i>FCD</i>	I	I
<i>FEC</i>	A	A
<i>FED</i>	A	A
<i>BFCD</i>	A	I
<i>BECD</i>	I	I
<i>BFEC</i>	I	I
<i>BFED</i>	A	A
<i>BFECD</i>	A	I

**Table 1.** Comparison of the 28 multiple point mutants between the predictions based on addition of mutation and the prediction from the innov'SAR model DSA\_FFT and model DSB\_FFT. I means better prediction by innov'SAR model and A means better predictions by using only the addition of mutations.

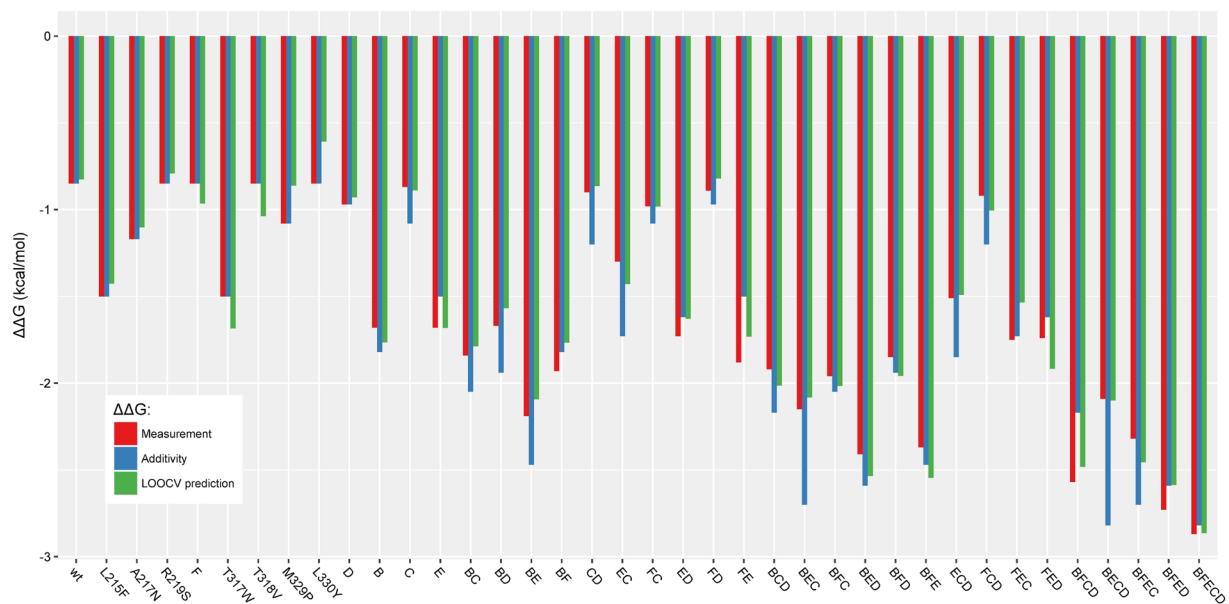
We decided to select 5 mutants comprising a large spectrum of E-values ranging from E-values that are close to the WT and to mutants with E-values that are outside the highest value of the training set. We were able to demonstrate that the 5 mutants have properties that corresponded quite well to the predictions. We also observed that innov'SAR not only allows accurate predictions of E-values that are inside the learning dataset but can also be used to predict and identify superior variants of ANEH. Gratifyingly, a  $R^2$  value of 0.94 was obtained. With only 37 mutants, our model was able to predict the potential outcome of 512, which means that on the basis of only 7% of the mutants, a whole fitness landscape was predicted. We determined the predictions of E-values of these 5 mutants only when considering the addition of mutations and when no FFT is used with the **model DSB\_noFFT** (Fig. 11). We found 0.62 and 0.64 as  $R^2$  values, respectively, for the predictions based on the addition of mutations and for the predictions generated without FFT. These results show that innov'SAR model performs better for the predictions of these new mutants. It can also be seen that the predicted E-values correspond very well to the experimentally determined enantioselectivity. Especially impressive is the performance of the predicted mutants P4(8) and P5(9), the experimental E-values amounting to 253 and 228, respectively (Table 2). These are distinctly better than mutant LW202 ( $E = 115$ ) evolved in our original study.

It is important to note that the predictions based on only the addition of mutations cannot find any better mutants than the mutant LW202, because this mutant comprises already all 9 possible mutations and each of these single mutants has an activity equal or better than the WT. The prediction based solely on descriptors without FFT also failed to give predictions of enantioselectivity better than mutant LW202 (Fig. 11). Both models would have therefore not resulted in the identification of improved new recombinants and are not able to predict recombinants which are outside of the training set.

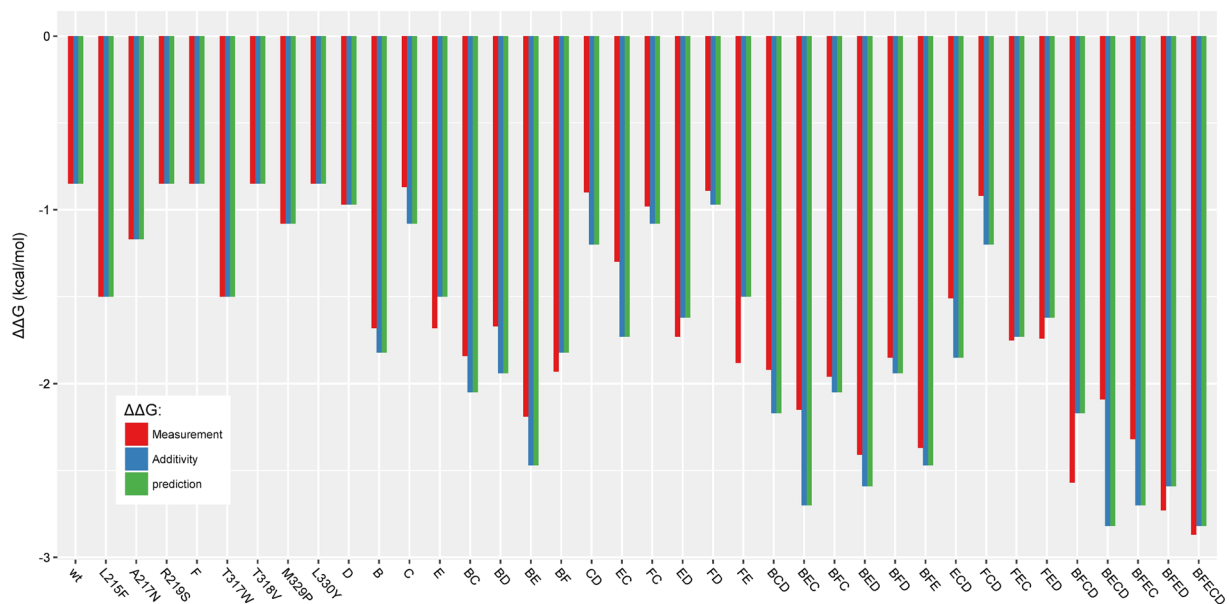
## Discussion

Previously, attempts have been made to study the effect of amino acid substitutions on the activity, function and stability of proteins whose structures have been resolved. This resulted in Quantitative Structure Function Relationship (QSFR) and Quantitative Structure Stability Relationship (QSSR) studies, respectively<sup>60–62</sup>. Particularly, the impact of mutations on the stability of proteins is of specific industrial interest and has been the subject of various studies. Although these structure dependent methods are effective in deriving the correlation between the mutation and its effect on the protein activity, they are limited by their requirement of the availability of the protein structure. The question of enhancing enzyme stereoselectivity was not addressed in these studies.



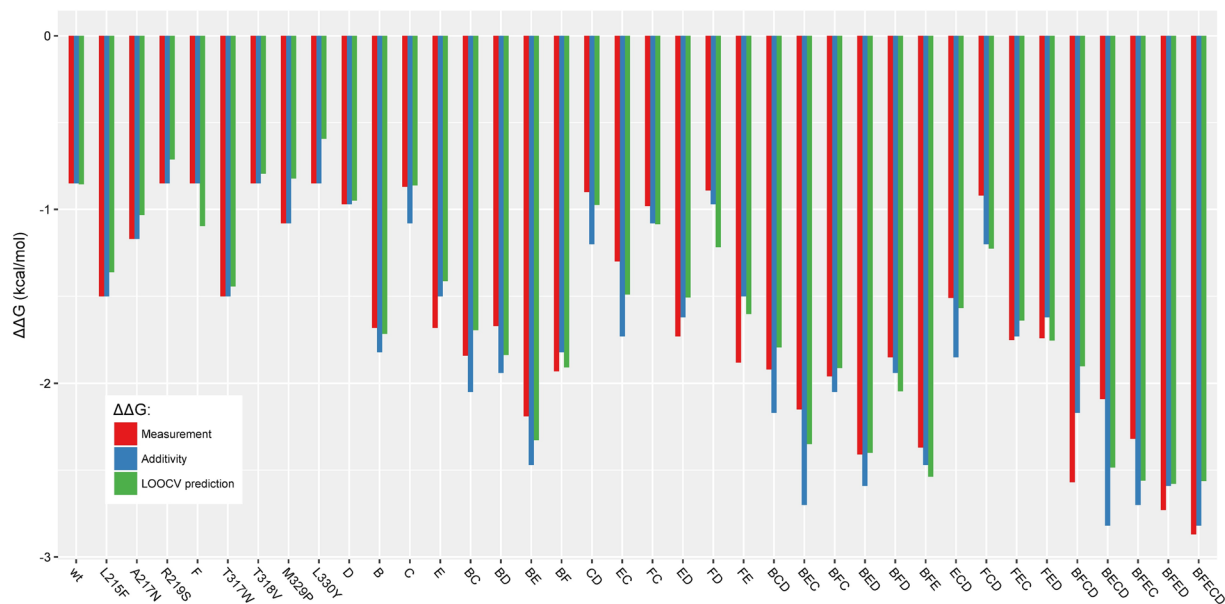


**Figure 7.** Prediction of  $\Delta\Delta G^*$  values by LOOCV method from model DSB\_FFT, based on the WT and the 37 single and multiple point mutants. Experimental values are in red, theoretical values for predictions based only on addition of mutations are in blue, and innov'SAR predictions are in green.

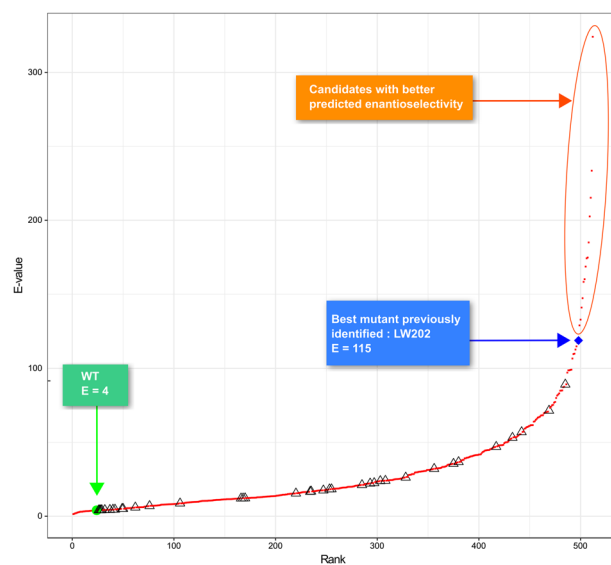


**Figure 8.** Prediction from model DSA\_noFFT, based on WT and 9 single point mutants and without FFT during the encoding phase. Experimental values are in red, theoretical values for predictions based only on addition of mutations are in blue, and innov'SAR predictions without FFT are in green.

Hence interest lies in deciphering the impact of mutations irrespective of the availability of structural information, purely based on physico-chemical and other molecular properties of the varying amino acids and statistical analysis thereof. The generation and accumulation of data about protein mutants and their properties as well as the increase of knowledge of protein structure such as enzymes and antibodies has enabled and supported the appearance of knowledge-based predictive algorithms for protein evolution. Quantitative structure-activity relationships (QSAR) have largely been applied to molecule modelling. QSAR methods have also been applied to model peptide or protein activity, but not focused on improving or reversing enantioselectivity<sup>63–65</sup>. It consists in using sets of descriptors derived from sequence information. One of these applications was termed as Protein Sequence Activity Relationship or ProSAR<sup>20</sup>. In this paper, the methodology relied on the binary encoding of the amino acid sequences of the wild type and a collection of few mutants whose activities are known. A statistical



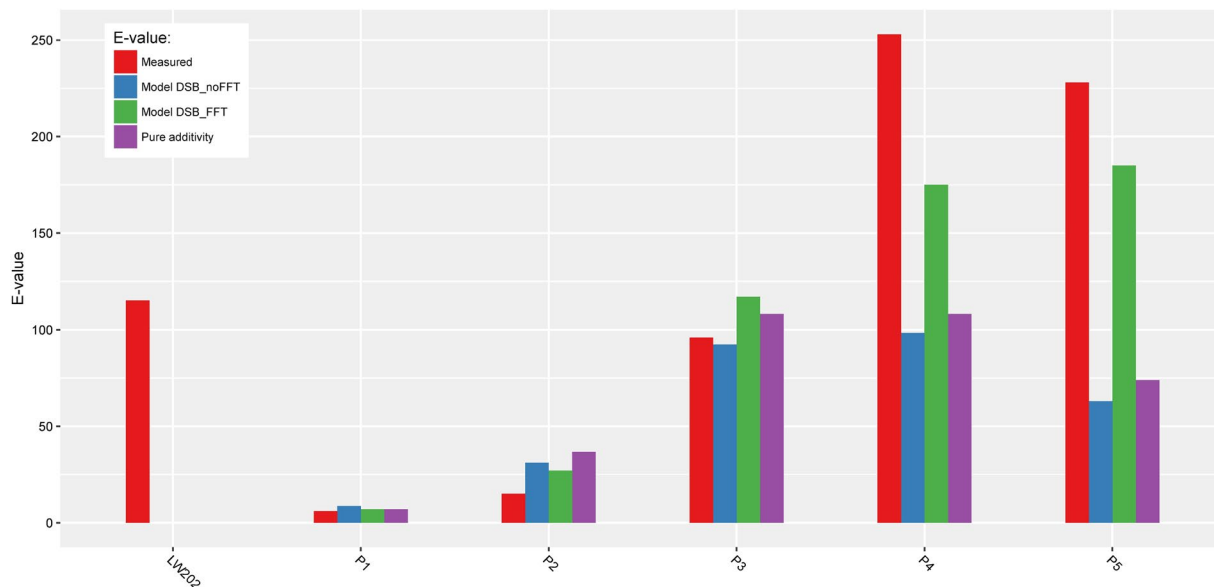
**Figure 9.** Prediction from model DSB\_noFFT, based on the WT and 37 single and multiple point mutants and without FFT during the encoding phase. Experimental values are in red, theoretical values for predictions based only on addition of mutations are in blue, and innov'SAR predictions without FFT are in green.



**Figure 10.** Ranking of the E-values for the 512 possible variants of ANEH with innov'SAR and the model DSB\_FFT. ( $\Delta$ ): E-value measured for WT and 37 single and multiple point mutants. ( $\bullet$ ): E-value predicted for all 512 possible variants.

Variant	Mutations	Predicted $\Delta\Delta G^{\ddagger}$	Predicted E-value	Experimental E-value
WT		-1.07	6	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_L330Y_C350V	-3.14	185	228

**Table 2.** Performance of new ANEH mutants as catalysts in the hydrolytic kinetic resolution of *rac-1* as predicted by innov'SAR.



**Figure 11.** E-values for the 5 mutants identified selected by innov'SAR and for the mutant LW202 originally evolved experimentally. In red: measured E-values, in blue: predicted E-values without FFT during the encoding (model DSB\_noFFT), in green: predicted E-values with FFT applied by innov'SAR (model DSB\_FFT), in purple: predicted E-value based only on the addition of mutations.

model is built to represent the relationship between the mutation and the activity<sup>20</sup>. Subsequent mutant libraries are generated by favouring those mutations that positively affect the activity. The ProSAR method has been demonstrated to show a 4000-fold improvement in the volumetric productivity of the enzyme halohydrin dehalogenase, while maintaining complete (*R*)-selectivity as already shown by the WT<sup>21</sup>. ProSAR was also used to enhance the activity of a transaminase, again with no trade-off in stereoselectivity; in this study, iterative saturation mutagenesis (ISM) at CAST-sites lining the binding pocket was also applied<sup>66</sup>. Both ProSAR and the structure dependent method QSFR, fall under the category of iterative mutant screening methods. The main assumption in iterative mutant library screening methods is that the effects of the mutations are additive in nature<sup>67–69</sup>. The additive nature of the fitness property is exploited mainly to avoid exhaustive and time-consuming search of the vast sequence space. However, in the case of ANEH it was shown that non-additive effects are predominant which makes the application of machine learning methods mainly based on additivity of mutations less efficient. Hence, these methods demonstrate difficulties to predict non-additive epistatic effects as well as long range interactions between amino acids which are non-adjacent to each other. Another method that combines structure activity analysis with machine learning algorithms is the SCHEMA based predictive method developed by the Arnold group<sup>70,71</sup>. Based on a train set of chimeras of the cytochrome P450 enzyme BM3, they developed a linear regression model that allowed the prediction of improved functional chimeras with  $r = 0.96$ .

This machine learning method has also been used to predict the properties of membrane bound proteins. Based on SCHEMA, chimeras of channel rhodopsins (ChR) were generated and characterized. The experimental values served as training set for a predictive machine learning application based on regression and Gauss models. Iterative cycles of prediction and empirical testing of ChR chimeras demonstrated that their properties could be efficiently predicted, resulting in the identification of improved variants. In contrast to innov'SAR, this hybrid approach however also requires structural knowledge of the protein<sup>72</sup>.

In the present report, we describe the first application of innov'SAR to guide the choice of mutations to be combined, and to find local or even global optima in the sequence space for enhanced enantioselectivity. A novel statistical model that links protein sequence to protein property was developed and successfully applied. The combination of descriptors and FFT resulted in appropriate predictions and the identification of improved epoxide hydrolase variants in terms of enhanced stereoselectivity. This is mainly due to the capability of the innov'SAR approach to predict the epistasis of interacting mutations. We could demonstrate that the prediction of epistatic effects of combinations of single point mutants is mainly due to the application of the FFT step. This becomes evident when new distinctly improved previously uncharacterized recombinants were correctly predicted, identified and substantiated experimentally. In this case the innov'SAR application comprising FFT was the only method that could predict any improved recombinants, while a pure additive model and innov'SAR without FFT proposed at best only the already known experimental LW202 mutant. This means that these models would have missed the local optima, which is a key criterion for the evaluation of machine learning algorithms.

The difference existing between the predictions generated when only the first step of innov'SAR (which converts amino acids into a series of descriptor dependent numbers) and the FFT (which converts the numbers into a spectrum with different frequencies) suggests that the later plays an important role in the accuracy of the predictions. In addition, we hypothesize that the frequency of a protein spectrum and each amino-acid position in the protein sequence is not linked to one position, but several amino-acid residues or positions at the same time.

The frequencies that have a better correlation with the experimental activity contribute better to the prediction of epistatic interactions. We are currently further deciphering the reasons for these observations.

In this work we propose an efficient combinatorial mutant library screening tool for the rational screening and improvement of epoxide hydrolase. While writing this paper, a study appeared describing genetic algorithms to improve an epoxide hydrolases in silico, but stereoselectivity was not treated<sup>73</sup>. Experimental proof of the predictive efficacy of these algorithms was also not provided. In a fundamentally different approach to in silico guidance when evolving the epoxide hydrolase ANEH for enhanced enantioselectivity, we have previously applied the ASRA-algorithm (Adaptive Substituent Reordering Algorithm)<sup>74</sup>. Accordingly, ASRA identifies the underlying regularity of the protein property landscape, in this case enantioselectivity. Consequently, it is not a QSAR-type approach, but a machine learning process very different from innov'SAR.

The present study is the first report that describes the application of innov'SAR in directed enzyme evolution. In this particular case we have limited the application to recombine learned mutations which are contained in the train set. We could identify and experimentally confirmed extremely fit mutants out of 512 possible mutational combinations by using a dataset made of only 7% of the mutants contained in the total fitness landscape. In a next step we are considering going further and to test to what extent our approach is able to predict the properties of non-learned mutations on learned or non-learned positions.

As a conclusion, innov'SAR approach based on sequence information and experimental data combined with Digital Signal Processing such as FFT has demonstrated its ability to capture and predict mutational epistasis. Predictions of new improved versions of the epoxide hydrolase enzyme for enantioselectivity have been confirmed experimentally. This machine learning approach takes into account the interactions between the amino-acids in a protein sequence and is very fast. It opens new opportunities in terms of protein engineering and screening.

As pointed out<sup>75</sup> and demonstrated in a series of recent publications, user-friendly and publicly accessible web-servers represent the future direction for demonstrating new findings or approaches. Actually, many practically useful web-servers have significantly increased their impacts on medical science, driving medicinal chemistry into an unprecedented revolution<sup>76</sup>, we shall make efforts in our future work to provide a web-server to display the findings that can be manipulated by users according to their needs.

## Materials and Methods

**Materials.** KOD Hot Start DNA Polymerase was obtained from Novagen. Restriction enzyme *Dpn* I was bought from NEB. The oligonucleotides were synthesized by Life Technologies. Plasmid preparation kit was ordered from Zymo Research, and PCR purification kit was bought from QIAGEN. DNA sequencing was conducted by GATC Biotech. All commercial chemicals were purchased from Sigma-Aldrich, Tokyo Chemical Industry (TCI) or Alfa Aesar.

**Methods.** *Evaluation of modelling performances by innov'SAR.* innov'SAR evaluates the generated models with the values of the root mean squared error (RMSE) and the coefficient of determination ( $R^2$ ) during the cross-validation stage and the full set stage. The formulas of these metrics are shown below:

$$R^2 = \frac{(\sum_{i=1}^S (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}))^2}{\sum_{i=1}^S (y_i - \bar{y})^2 \sum_{i=1}^S (\hat{y}_i - \bar{\hat{y}})^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^S (y_i - \hat{y}_i)^2}{S}}$$

where,  $y_i$  is the measured activity of the  $i^{\text{th}}$  sequence,  $\hat{y}_i$  is the predicted activity of the  $i^{\text{th}}$  sequence,  $\bar{y}$  is the average and  $S$  the number of sequences.

**PCR based methods for construction of ANEH mutants.** The ANEH, constructed in PET-22b<sup>77</sup>, was chosen as template for mutants constructing with over-lap PCR and megaprimer approach<sup>78</sup>. 50  $\mu$ L reaction mixtures typically contained 30  $\mu$ L water, 5  $\mu$ L KOD hot start polymerase buffer (10 $\times$ ), 3  $\mu$ L 25 mM MgSO<sub>4</sub>, 5  $\mu$ L 2 mM dNTPs, 2.5  $\mu$ L DMSO, 0.5  $\mu$ L (50~100 ng) template DNA, 100  $\mu$ M primers Mix (0.5  $\mu$ L each) and 0.5  $\mu$ L (short fragment PCR) or 1  $\mu$ L (megaprimer PCR) KOD hot start polymerase. The PCR conditions for short fragment: 95 °C 3 min, (95 °C 30 sec, 56 °C 30 sec, 68 °C 40 sec)  $\times$  30 cycles, 68 °C 120 sec. For mega-PCR: 95 °C 3 min, (95 °C 30 sec, 60 °C 30 sec, 68 °C 5 min 30 sec)  $\times$  28 cycles, 68 °C 10 min. The PCR products were analysed on agarose gel by electrophoresis and purified using a Qiagen PCR purification kit. 2  $\mu$ L NEB CutSmart™ Buffer and 2  $\mu$ L *Dpn* I were added in 50  $\mu$ L PCR reaction mixture and the digestion was carried out at 37 °C for 7 h. After *Dpn* I digestion, the PCR products 1.5  $\mu$ L were directly transformed into electro-competent *E. coli* BL21(DE3) to create the final library.

**Primer design and creation of ANEH mutants.** Primer design depend upon the particular amino-acid chosen, and in the case of P1 involves three sites mutation: (1) Amplification of the short fragments of WTANEH using mixed primers P1-A217N/R219S-F/P1-L249Y-R (Table S6); (2) Amplification of the whole plasmid WTANEH using the PCR products of step1 as megaprimers, leading to the final plasmids for mutant P1 generation.

For mutant P2: (1) Amplification of the short fragments of P1 using mixed primers P2-P1-A217N/S219R-F/P2-P1-T317W/M329P/L330Y-R and P2-P1-T317W/M329P/L330Y-F/P2-P1-C350V-R, respectively; (2) Over-lap PCR using the PCR products of step1 as template and mixed primers P2-P1-A217N/S219R-F/P2-P1-C350V-R;

(3) Amplification of the whole plasmid of P1 using the over-lap PCR product of step2 as megaprimers, leading to the final plasmids for mutant P2 generation.

For mutant P3: (1) Amplification of the short fragments of P1 using mixed primers P2'-P1-L215F/A217N/R219S-F/P3-P1-T317W/M329P-R; (2) Amplification of the whole plasmid of P1 using the PCR product of step1 as megaprimers, leading to the final plasmid for mutant P2' generation. (3) Amplification of the short fragments of P2' using mixed primers P3-P2'-T317W/T318V-F/P3-P2'-C350V-R; (4) Amplification of the whole plasmid of P2' using the PCR product of step1 as megaprimers, leading to the final plasmids for mutant P3 generation.

For mutants P4 and P5: (1) Amplification of the short fragments of P3 using mixed primers P4-P3-S219R-F/P4-P3-C350V-R and P5-P3-P329M-L330Y-F/P5-P3-R; (2) Amplification of the whole plasmid of P3 using the PCR product of step1 as megaprimers, leading to the final plasmids for mutants P4 and P5 generation. All the primers used are listed in Table S6. The PCR products were digested by *Dpn* I and transformed into electro-competent *E. coli* BL21 (DE3) to create the library.

**Protein expression.** *E. coli* BL21 (DE3) cells carrying the recombinant plasmid were cultivated in 5 mL LB medium containing carbenicillin (100 µg/mL) overnight at 37 °C. The overnight culture was inoculated into 100 mL of TB medium containing carbenicillin (100 µg/mL) and grown at 37 °C. The culture was induced by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) with a final concentration of 0.2 mM when OD<sub>600</sub> reached 0.6, and then allowed to grow for additional 12 h at 25 °C. After centrifugation at 6000 g for 15 min at 4 °C, the bacterial pellet was washed once with phosphate buffer (50 mM, pH 7.4), and resuspended in a phosphate buffer (50 mM, pH 7.4).

**Hydrolytic kinetic resolution of *rac-1*.** A 1 mL mixture of 50 mM *rac-1* (10 mM for mutant P2) and recombinant expressed whole cells of WT ANEH or mutants (WT ANEH: OD<sub>600</sub> = 1, P1: OD<sub>600</sub> = 0.02, P2: OD<sub>600</sub> = 25, P3: OD<sub>600</sub> = 0.02, P4: OD<sub>600</sub> = 0.3 and P5: OD<sub>600</sub> = 0.02) in PBS buffer (50 mM, pH 7.4) was stirred at 25 °C. Then 0.2 mL reaction product were extracted with ethyl ether (0.2 mL) at 10 min and 30 min (P2: 1 h, 3 h, 5 h and 7 h), respectively.

## Data Availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

## References

- Chen, K. & Arnold, F. H. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. USA* **90**, 5618–5622 (1993).
- Reetz, M. T., Zonta, A., Schimossek, K., Jaeger, K.-E. & Liebeton, K. Creation of Enantioselective Biocatalysts for Organic Chemistry by *In Vitro* Evolution. *Angew. Chem. Int. Ed. Engl.* **36**, 2830–2832 (1997).
- Lutz, S. & Iamurri, S. M. Protein Engineering: Past, Present, and Future. In *Protein Engineering* (eds Bornscheuer, U. T. & Höhnle, M.) **1685**, 1–12 (Springer New York, 2018).
- Zeymer, C. & Hilvert, D. Directed Evolution of Protein Catalysts. *Annu. Rev. Biochem.* **87**, 131–157 (2018).
- Directed Enzyme Evolution: Advances and Applications*. <https://doi.org/10.1007/978-3-319-50413-1> (Springer International Publishing, 2017).
- Hammer, S. C., Knight, A. M. & Arnold, F. H. Design and evolution of enzymes for non-natural chemistry. *Curr. Opin. Green Sustain. Chem.* **7**, 23–30 (2017).
- Reetz, M. T. *Directed Evolution of Selective Enzymes: Catalysts for Organic Chemistry and Biotechnology*, <https://doi.org/10.1002/9783527655465> (Wiley-VCH Verlag GmbH & Co. KGaA, 2016).
- Reetz, M. T., Bocola, M., Carballeira, J. D., Zha, D. & Vogel, A. Expanding the Range of Substrate Acceptance of Enzymes: Combinatorial Active-Site Saturation Test. *Angew. Chem. Int. Ed.* **44**, 4192–4196 (2005).
- Turner, N. J. Directed evolution drives the next generation of biocatalysts. *Nat. Chem. Biol.* **5**, 567–573 (2009).
- Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes. *Protein Sci.* **21**, 717–726 (2012).
- Denard, C. A., Ren, H. & Zhao, H. Improving and Repurposing Biocatalysts via directed Evolution. *Curr. Opin. Chem. Biol.* **25**, 55–64 (2015).
- Franceus, J., Verhaeghe, T. & Desmet, T. Correlated positions in protein evolution and engineering. *J. Ind. Microbiol. Biotechnol.* **44**, 687–695 (2017).
- Cheng, F., Zhu, L. & Schwaneberg, U. Directed evolution 2.0: improving and deciphering enzyme properties. *Chem. Commun.* **51**, 9760–9772 (2015).
- Muggleton, S., King, R. D. & Stenberg, M. J. E. Protein secondary structure prediction using logic-based machine learning. *Protein Eng. Des. Sel.* **5**, 647–657 (1992).
- Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130 (2001).
- Wold, S., Trygg, J., Berglund, A. & Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* **58**, 131–150 (2001).
- Lapins, M., Prusis, P., Gutcaits, A., Lundstedt, T. & Wikberg, J. E. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta* **1525**, 180–190 (2001).
- Damborský, J. & Brezovsky, J. Computational tools for designing and engineering biocatalysts. *Curr. Opin. Chem. Biol.* **13**, 26–34 (2009).
- Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned Protein Embeddings for Machine Learning. *Bioinformatics*, <https://doi.org/10.1093/bioinformatics/bty178> (2018).
- Fox, R. *et al.* Optimizing the search algorithm for protein engineering by directed evolution. *Protein Eng.* **16**, 589–597 (2003).
- Fox, R. J. *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **25**, 338–344 (2007).
- Choi, Y. H., Kim, J. H., Park, B. S. & Kim, B.-G. Solubilization and Iterative Saturation Mutagenesis of α1,3-fucosyltransferase from *Helicobacter pylori* to enhance its catalytic efficiency: Protein Engineering of Fucosyltransferase. *Biotechnol. Bioeng.* **113**, 1666–1675 (2016).
- Smith, S. W. *The scientist and engineer's guide to digital signal processing*. (California Technical Pub, 1997).
- Benson, D. C. Digital signal processing methods for biosequence comparison. *Nucleic Acids Res.* **18**, 3001–3006 (1990).
- Viari, A., Soldano, H. & Ollivier, E. A scale-independent signal processing method for sequence analysis. *Comput. Appl. Biosci.* **6**, 71–80 (1990).



26. Cosić, I. Macromolecular bioactivity: is it resonant interaction between macromolecules?—Theory and applications. *IEEE Trans. Biomed. Eng.* **41**, 1101–1114 (1994).
27. Cosić, I. & Pirogova, E. Bioactive peptide design using the Resonant Recognition Model. *Nonlinear Biomed. Phys.* **1**, 7 (2007).
28. McLachlan, A. D. Analysis of periodic patterns in amino acid sequences: collagen. *Biopolymers* **16**, 1271–1297 (1977).
29. Cornette, J. L. *et al.* Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **195**, 659–685 (1987).
30. Aggeli, A., Hamodrakas, S. J., Komitopoulou, K. & Konsolaki, M. Tandemly repeating peptide motifs and their secondary structure in *Ceratitis capitata* eggshell proteins Ccs36 and Ccs38. *Int. J. Biol. Macromol.* **13**, 307–315 (1991).
31. Taylor, W. R., Heringa, J., Baud, F. & Flores, T. P. A Fourier analysis of symmetry in protein structure. *Protein Eng. Des. Sel.* **15**, 79–89 (2002).
32. Hejase de Trad, C., Fang, Q. & Cosić, I. The resonant recognition model (RRM) predicts amino acid residues in highly conserved regions of the hormone prolactin (PRL). *Biophys. Chem.* **84**, 149–157 (2000).
33. Pan, Y.-X. *et al.* Predicting protein subcellular location using digital signal processing. *Acta Biochim. Biophys. Sin.* **37**, 88–96 (2005).
34. Mainardi, L. T., Pattini, L. & Cerutti, S. Application of the Ramanujan Fourier Transform for the analysis of secondary structure content in amino acid sequences. *Methods Inf. Med.* **46**, 126–129 (2007).
35. Marsella, L., Sirocco, F., Trovato, A., Seno, F. & Tosatto, S. C. E. REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics* **25**, i289–i295 (2009).
36. Walsh, I. *et al.* RAPHAE: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics* **28**, 3257–3264 (2012).
37. Hrabe, T. & Godzik, A. ConSole: using modularity of Contact maps to locate Solenoid domains in protein structures. *BMC Bioinformatics* **15**, 119 (2014).
38. Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K.-C. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. *J. Biomol. Struct. Dyn.* **34**, 1946–1961 (2016).
39. Kawashima, S., Ogata, H. & Kanehisa, M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **27**, 368–369 (1999).
40. Kawashima, S. *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202–205 (2008).
41. Veljković, V., Cosić, I., Dimitrijević, B. & Lalović, D. Is it possible to analyze DNA and protein sequences by the methods of digital signal processing? *IEEE Trans. Biomed. Eng.* **32**, 337–341 (1985).
42. Veljkovic, V. *et al.* Characterization of conserved properties of hemagglutinin of H5N1 and human influenza viruses: possible consequences for therapy and infection control. *BMC Struct. Biol.* **9**, 21 (2009).
43. Veljkovic, V. *et al.* Identification of hemagglutinin structural domain and polymorphisms which may modulate swine H1N1 interactions with human receptor. *BMC Struct. Biol.* **9**, 62 (2009).
44. Cosić, I. *The Resonant Recognition Model of Macromolecular Bioactivity*, <https://doi.org/10.1007/978-3-0348-7475-5> (Birkhäuser Basel, 1997).
45. Nwankwo, N. Digital Signal Processing Techniques: Calculating Biological Functionalities. *J. Proteomics Bioinform.* **04** (2012).
46. Cotic, I., Pavlovic, M. & Vojisavljevic, V. Prediction of «hot spots» in interleukin-2 based on informational spectrum characteristics of growth- regulating factors. *Comparison with experimental data. Biochimie* **71**, 333–342 (1989).
47. Fontaine, N. & Cadet, F. Method and electronic system for predicting at least one fitness value of a protein, related computer program product (2016).
48. Archelas, A. & Furstoss, R. Synthetic applications of epoxide hydrolases. *Curr. Opin. Chem. Biol.* **5**, 112–119 (2001).
49. Reetz, M. T., Wang, L.-W. & Bocola, M. Directed Evolution of Enantioselective Enzymes: Iterative Cycles of CASTing for Probing Protein-Sequence Space. *Angew. Chem. Int. Ed.* **45**, 1236–1241 (2006).
50. Reetz, M. T. & Sanchis, J. Constructing and Analyzing the Fitness Landscape of an Experimental Evolutionary Process. *ChemBioChem* **9**, 2260–2267 (2008).
51. Reetz, M. T. *et al.* Directed Evolution of an Enantioselective Epoxide Hydrolase: Uncovering the Source of Enantioselectivity at Each Evolutionary Stage. *J. Am. Chem. Soc.* **131**, 7334–7343 (2009).
52. Reetz, M. T. Recent Advances in Directed Evolution of Stereoselective Enzymes. In *Directed Enzyme Evolution: Advances and Applications* (ed. Alcalde, M.) 69–99, [https://doi.org/10.1007/978-3-319-50413-1\\_3](https://doi.org/10.1007/978-3-319-50413-1_3) (Springer International Publishing, 2017).
53. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **273**, 236–247 (2011).
54. Zhou, H. & Zhou, Y. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins* **54**, 315–322 (2004).
55. Rackovsky, S. & Scheraga, H. A. Differential geometry and polymer conformation. 4. *Conformational and nucleation properties of individual amino acids. Macromolecules* **15**, 1340–1346 (1982).
56. Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design: How Epistasis Impairs Predictability in Enzyme Evolution. *Protein Sci.* **25**, 1260–1272 (2016).
57. Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867 (2008).
58. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution: Epistasis in Protein Evolution. *Protein Sci.* **25**, 1204–1218 (2016).
59. Reetz, M. T. The Importance of Additive and Non-Additive Mutational Effects in Protein Engineering. *Angew. Chem. Int. Ed.* **52**, 2658–2666 (2013).
60. Fersht, A. R., Leatherbarrow, R. J. & Wells, T. N. Structure-activity relationships in engineered proteins: analysis of use of binding energy by linear free energy relationships. *Biochemistry* **26**, 6030–6038 (1987).
61. Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **8**, 243–256 (1994).
62. Damborský, J. Quantitative structure-function and structure-stability relationships of purposely modified proteins. *Protein Eng.* **11**, 21–30 (1998).
63. Hellberg, S., Sjöström, M., Skagerberg, B. & Wold, S. Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.* **30**, 1126–1135 (1987).
64. Hellberg, S. *et al.* Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships. *Int. J. Pept. Protein Res.* **37**, 414–424 (1991).
65. Norinder, U. Single and domain mode variable selection in 3D QSAR applications. *J. Chemom.* **10**, 95–105 (1996).
66. Savile, C. K. *et al.* Biocatalytic Asymmetric Synthesis of Chiral Amines from Ketones Applied to Sitagliptin Manufacture. *Science* **329**, 305–309 (2010).
67. Sandberg, W. S. & Terwilliger, T. C. Engineering multiple properties of a protein by combinatorial mutagenesis. *Proc. Natl. Acad. Sci. USA* **90**, 8367–8371 (1993).
68. Wells, J. A. Additivity of mutational effects in proteins. *Biochemistry* **29**, 8509–8517 (1990).
69. Verma, D., Grigoryan, G. & Bailey-Kellogg, C. Structure-based design of combinatorial mutagenesis libraries. *Protein Sci. Publ. Protein Soc.* **24**, 895–908 (2015).
70. Otey, C. R. *et al.* Structure-Guided Recombination Creates an Artificial Family of Cytochromes P450. *PLoS Biol.* **4**, e112 (2006).
71. Li, Y. *et al.* A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **25**, 1051–1056 (2007).
72. Bedbrook, C. N. *et al.* Structure-guided SCHEMA recombination generates diverse chimeric channelrhodopsins. *Proc. Natl. Acad. Sci.* **114**, E2624–E2633 (2017).

73. Barley, M. H., Turner, N. J. & Goodacre, R. Improved Descriptors for the Quantitative Structure–Activity Relationship Modeling of Peptides and Proteins. *J. Chem. Inf. Model.* **58**, 234–243 (2018).
74. Feng, X., Sanchis, J., Reetz, M. T. & Rabitz, H. Enhancing the Efficiency of Directed Evolution in Focused Enzyme Libraries by the Adaptive Substituent Reordering Algorithm. *Chem. - Eur. J.* **18**, 5646–5654 (2012).
75. Chou, K.-C. & Shen, H.-B. REVIEW: Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.* **01**, 63–92 (2009).
76. Chou, K.-C. An Unprecedented Revolution in Medicinal Chemistry Driven by the Progress of Biological Science. *Curr. Top. Med. Chem.* **17** (2017).
77. Reetz, M. T. & Zheng, H. Manipulating the Expression Rate and Enantioselectivity of an Epoxide Hydrolase by Using Directed Evolution. *ChemBioChem* **12**, 1529–1535 (2011).
78. Li, G., Zhang, H., Sun, Z., Liu, X. & Reetz, M. T. Multiparameter Optimization in Directed Evolution: Engineering Thermostability, Enantioselectivity, and Activity of an Epoxide Hydrolase. *ACS Catal.* **6**, 3679–3687 (2016).

### Author Contributions

F.C., R.P., J.S., B.O. and M.T.R. designed the study. F.C. and M.T.R. obtained the funding. G.L., J.S. and M.T.R. conducted the experiments and initial data acquisition. F.C., N.F., M.N., R.P. and B.O. applied statistical, mathematical, computational techniques to analyse the data and build the models. F.C., N.F., G.L., J.S., M.N., R.P., I.V. and M.T.R. analysed and interpreted the results. F.C., N.F., G.L., J.S., M.N., R.P., I.V. and M.T.R. wrote the manuscript.

### Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-35033-y>.

**Competing Interests:** F.C. and B.O. are co-founder of PEACCEL. N.F., M.N., R.P., I.V. are paid employees of PEACCEL. G.L., J.S. and M.T.R. declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018