Journal of Computer-Aided Molecular Design (2017) 31:1085–1096
https://doi.org/10.1007/s10822-017-0090-x

# Learning epistatic interactions from sequence-activity data to predict enantioselectivity

**Julian Zaugg[1] · Yosephine Gumulya[1] · Alpeshkumar K. Malde[1] · Mikael Bodén[1]**

Received: 10 August 2017 / Accepted: 4 December 2017 / Published online: 12 December 2017


**Abstract**
Enzymes with a high selectivity are desirable for improving economics of chemical synthesis of enantiopure compounds. To improve enzyme selectivity mutations are often introduced near the catalytic active site. In this compact environment epistatic interactions between residues, where contributions to selectivity are non-additive, play a significant role in determining the degree of selectivity. Using support vector machine regression models we map mutations to the experimentally characterised enantioselectivities for a set of 136 variants of the epoxide hydrolase from the fungus *Aspergillus niger* (*An*EH). We investigate whether the influence a mutation has on enzyme selectivity can be accurately predicted through linear models, and whether prediction accuracy can be improved using higher-order counterparts. Comparing linear and polynomial degree = 2 models, mean Pearson coefficients ($r$) from $50 \times 5$-fold cross-validation increase from 0.84 to 0.91 respectively. Equivalent models tested on interaction-minimised sequences achieve values of $r = 0.90$ and $r = 0.93$. As expected, testing on a simulated control data set with no interactions results in no significant improvements from higher-order models. Additional experimentally derived *An*EH mutants are tested with linear and polynomial degree = 2 models, with values increasing from $r = 0.51$ to $r = 0.87$ respectively. The study demonstrates that linear models perform well, however the representation of epistatic interactions in predictive models improves identification of selectivity-enhancing mutations. The improvement is attributed to higher-order kernel functions that represent epistatic interactions between residues.

**Keywords** Epoxide hydrolase · *Aspergillus niger* · Support vector machine · Non-additive · Fitness · Bioinformatics · Machine learning

## Introduction

Enzymes with high selectivity are desirable to the biochemical and pharmaceutical industry for their potential to increase yields of enantiopure chemical and drug products, improve efficiency of bio-transformations and lower environmental impacts through reduction of chemical waste. Improvements in enantioselectivity, where one optically pure enantiomer is preferentially produced from a racemic substrate, are sought, in part, to address regulatory requirements enforced by drug regulation agencies [1–3]. To improve enzyme selectivity the most effective mutations are often introduced within the active site region where direct interactions with a substrate can occur [4, 5]. Epistatic interactions between residues within the active site region also play a significant role in influencing enzyme selectivity [5]. An epistatic interaction exists between two or more residues when their combined contribution to enzyme fitness deviates from that expected by simply adding their individual contributions, i.e. *non-additive* vs *additive* [6]. Non-additive fitness contributions will complicate the exploration of the "fitness landscape" by making its topology more rugged [6, 7]. It has not been established whether modelling sequence-activity relationships under the assumption of additivity [8–11] is sufficient to accurately predict beneficial mutations and their contributions to enzyme selectivity in the presence of strong epistatic effects, or whether non-additive

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s10822-017-0090-x) contains supplementary material, which is available to authorized users.

✉ Julian Zaugg
julian.zaugg@uqconnect.edu.au

✉ Mikael Bodén
m.boden@uq.edu.au

[1] School of Chemistry and Molecular Biosciences, University of Queensland, St Lucia, QLD 4072, Australia

methods [10, 12–14] are required. In this study linear models and counterparts representing both pairwise and higher-order (three or more) residue interactions are constructed and evaluated on enzymes whose enantioselectivities have been experimentally characterised.

Previous modelling studies predicting the preferred enantiomer and the degree of enantioselectivity have primarily used quantitative structure-activity relationship (QSAR) or molecular dynamics methods [15–26]. Such methods often require high-resolution protein structures, however the rate at which such structures are produced lags significantly behind the rate that proteins are sequenced and their activities characterised. Methods that guide the choice of beneficial mutations from sequence data alone are therefore desirable [27].

Machine learning kernel methods [28] including Gaussian processes (GPs) and support vector machines (SVMs) have been used to describe the relationship between protein sequence and activity/function by representing pairwise interactions between residues based on residue–residue contact maps [13, 14, 29]. The assumption is that sequences with similar structures, as described by a structure-based *kernel function*, will have similar functions. GP regression and classification has been used to improve the thermostability, catalytic activity and ligand binding affinity of chimeric cytochrome P450 102A1-3 sequences [14], and expression and localisation of chimeric channelrhodopsins [29]. SVMs have been applied to classify structural viability of chimeric cytochrome P450 sequences [13].

Simple models using linear regression have also been successfully applied to predict protein functional status, thermostability and biological activity [8, 11, 30–33]. Linear models based on characterised sequences generated with SCHEMA-guided recombination [8, 11, 30, 31] have demonstrated good predictive ability with minimal sampling of the protein fitness landscape accessible by recombination, also referred to as the protein recombinational landscape [11]. The SCHEMA algorithm [34–36] can be used to design libraries of chimeric sequences by taking advantage of the structural similarity of recombined sequences. By referring to a representative crystal structure and an alignment of homologous parent template sequences, each potential mutant in a library is assigned a disruption score that reflects the number of residue–residue contacts that would be broken due to novel combinations of fragments from parent template sequences. Potential cut-points can be then identified that minimise the degree of structural disruption in a library, i.e. the boundaries of structural sub-units or blocks [37]. Proteins sampled from these optimised libraries are more likely to be folded and functional. Minimising the number of residue–residue contacts broken during recombination will tend to partition epistatic interactions into structural sub-units, promoting an additive fitness contribution from each fragment [8, 11, 38]. In contrast, when using a focused mutagenesis strategy such as saturation mutagenesis [39] this partitioning will not occur, thus increasing the likelihood of non-additive fitness contributions.

The combinatorial active site saturation test (CAST) [40] is an experimental strategy developed to focus the exploration of the fitness landscape, producing what is often coined "small, but smart" sequence libraries. In this approach a number of residues with side-chains within the binding pocket of the enzyme are selected and assigned to groups of 2–3 residues. These groups are then subjected to (iterative) saturation mutagenesis (ISM) [41]. Simultaneous mutation of groups of residues allows exploration of potentially beneficial epistatic interactions.

Gumulya et al. [42] applied iterative CASTing to eight residues—Leu215, Arg219, Phe244, Leu249, Thr317, Thr318, Leu349 and Cys350—lining the binding pocket of the epoxide hydrolase (EH) from *Aspergillus niger* (*An*EH) in order to improve the enantioselective preference for the (*S*) enantiomer of glycidyl phenyl ether. The study produced a set of mutants with a wide range of improved enantioselectivities. Strong cooperative epistatic interactions between residues were observed along a number of the explored evolutionary pathways, as such this data set is suited to the development and evaluation of higher-order models.

In this study we determine whether the modelling of epistatic interactions for the above described set of experimentally characterised *An*EH sequence variants [42] can improve the prediction of selectivity-enhancing mutations. Support vector regression (SVR) models are fitted with lower-order kernels and counterparts representing natural substitution rates and higher-order interactions between residues. Models are evaluated on a small set of *An*EH mutants from separate protein engineering studies [43–45]. In addition, models are evaluated on two sequence-activity data sets with minimised and removed epistatic interactions—the thermostability data for a set of chimeric bacterial cytochrome P450 sequences [8] and a simulated control *An*EH data set where each mutation contributes additively to fitness.

## Methods

### Experimental data: *An*EH sequences

A set of 145 *An*EH sequence variants (including wild type) and their respective enantioselectivities for (*S*)-glycidyl phenyl ether was obtained [42] (Supplementary material Table S1). The enantioselectivity measurements are reported as the enantiomeric ratio between the fast and slow reacting enantiomers—an $E$ value [46, 47] ranging from $E = 5$ (wild type) to $E = 158$. Generally enzymes with $E$ values $< 15$ are considered not practically useful, 15–30 are moderate

and > 30 are excellent [47]. Nine of the sequences were observed to have identical amino acid insertions to at least one other, but also differing $E$ values (Supplementary material Table S4). Seven pairs of sequences have $E$ value differences ranging from 1 to 4, while two pairs have differences of 10 and 22. The mean difference for all nine pairs is 5.33. $E$ values for the pairs of sequences are well distributed, ranging from 24 to 88 with a mean of 50.88. For each pair of duplicate sequences one has been removed and the average of their $E$ values assigned to the remaining sequence, leaving 136 unique sequences for the purpose of generating models.

$E$ values for these *An*EH variants have been calculated from the enantiomeric excess ($e.e.$) values for the substrate ($s$) and product ($p$) using Eq. (1) [48].

$$E = \frac{\ln \frac{e.e._p(1-e.e._s)}{(e.e._p+e.e._s)}}{\ln \frac{e.e._p(1+e.e._s)}{(e.e._p+e.e._s)}} \tag{1}$$

## Experimental data: CYP102A sequences

If a given set of residues have no epistatic interactions, no improvement in predictive accuracy would be expected of a model that represents such interactions compared to one that assumes residue independence. To represent this scenario as closely as possible with experimental data, an additional interaction-minimised data set of 241 chimeric bacterial cytochrome P450 sequences and their respective thermostabilities was obtained (Supplementary material Table S2). The thermostability measurements are reported as the temperature at which 50% of the protein is inactivated after 10 minutes ($T_{50}^{10}$). These sequences are generated from the SCHEMA-guided recombination of eight sequence blocks of the haem domains of cytochrome P450 BM3 (CYP102A1) from *Bacillus megaterium* and its homologues (CYP102A2-3) [8]. Of the approximately 500 residue–residue interactions in the original parent template structure, the average inter-block interactions broken in these chimeric sequences is fewer than 30 [8]. Although comparatively few residue–residue interactions are broken, it is expected that the influence of epistatic effects on the thermostability is largely reduced rather than completely non-existent. For a number of the sequences a deletion was observed at positions 230, 465 and 466. These positions are removed from all sequences in order to simplify analysis.

## Simulated data: additive *An*EH sequences

Using the 136 *An*EH sequences described above as a template, a control set of sequences are generated where epistatic effects have been removed, i.e. the fitness contribution of each residue is made to be additive (Supplementary material Table S3). To model an additive fitness landscape, an NK-model as described by Kauffman and Weinberger [49] is applied. $K$ is a coupling parameter that controls the degree of interactions between residues; by setting $K = 0$ the fitness contribution from each residue is treated as independent. The total fitness $y$ of an $N$ length sequence is defined as the average fitness contribution $f_i$ of its constituent amino acids (Eq. 2).

$$y = \frac{1}{N}\sum_{i}^{N} f_i \tag{2}$$

Fitness contributions for each amino acid per mutation site are drawn from an $8 \times 20$ look-up table (Supplementary material Table S3), where each entry is randomly sampled from a gamma distribution with a mean and variance of 0.35 [50]. By sampling from this distribution, residues will tend to be neutral (with low fitness contributions) and a few residues will tend to have a large impact on fitness [50]. For ease of comparison, fitness values are scaled to the same range as the $E$ values observed in the experimentally derived set of *An*EH sequences. A small amount of noise is then added to the calculated total fitness values for each sequence to represent possible experimental error. Error values are randomly sampled from a uniform distribution $U(-20, 20)$ that approximates the maximum range of $E$ value differences observed in the set of duplicate *An*EH sequences (Supplementary material Table S4).

## Support vector machines and kernel functions

SVMs [51–54] find the maximum margin hyperplane between a set of training sequence-activity data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ in a given input space $\mathcal{X}$, where $x$ is the sequence content and $y$ the observed activity. This is achieved through the use of a kernel function $K(x, x')$, which maps the set of input $\{x_i\}$ into a feature space $\mathcal{F}$ by calculating the similarity between pairs of inputs $x$ and $x'$. The separating hyperplane in $\mathcal{F}$ may be non-linear in $\mathcal{X}$. For SVR, linear regression is performed on $\{x_i\}$ points once they have been mapped to $\mathcal{F}$. SVMs have been used extensively in the field of chemoinformatics to identify potential lead compounds and ligand interaction partners [55]. Sequences (or ligands) are often encoded as numeric vectors representing a number of physicochemical properties [56]. The expectation is that sequences or individual residues with similar encodings will have similar activities and functions [57, 58]. We adapt a kernel function proposed by Sulimova et al.[59] to represent pairwise and higher-order interactions between residues. The kernel function itself is based on the pioneering work of Dayhoff and colleagues [60] that saw the introduction of a Markovian based model of protein evolution and the production of a number of amino acid instantaneous rate

matrices ($Q$), which have been the basis for the development of a number of models of evolution [61–63]. A conditional probability matrix $P(t)$ containing the probabilities of an amino acid $i$ changing into another amino acid $j$ after a given time $t \geq 0$ is derived directly from $Q$ [64] through

$$P(t) = \exp^{tQ} \tag{3}$$

$Q$ can be any rate matrix estimated through Dayhoff or Henikoff [65] techniques. For the evolutionary model based kernel function used in the present study the Le and Gascuel rate matrix is used [63], due to the incorporation of evolutionary rate variability and use of a larger and more diverse set of sequences in its construction. Given the above, an amino acid at the $i$th position in a given sequence is encoded as a feature vector

$$aa_i^t = \left( \sqrt{P(aa_k)^t}\, P(aa_i|aa_k)^t, \right. \\ \left. k = 1, \ldots, k = 20) \in \mathbb{R}^{20} \tag{4}$$

where $aa_k$ is the $k$th possible ancestral amino acid within a standard 20 amino acid alphabet, $P(aa_k)$ the probability of the ancestral amino acid, and $P(aa_i|aa_k)$ the conditional probability of observing the transition from the ancestral to the extant amino acid at time $t$. As such, each $N$ length sequence $x$ is fully encoded as the concatenation $\frown$ of its respective vectors

$$x = \left( \mathbb{R}^{20}_{aa^t_{i=1}} \frown \mathbb{R}^{20}_{aa^t_{i=2}} \frown \ldots \mathbb{R}^{20}_{aa^t_{i=N}} \right) \tag{5}$$

In its simplest form the function assumes linearity between the individual positional terms, i.e. $K(x, x') = x^T x'$. This representation treats residues as not interacting with other residues.

This linear implementation has been extended to represent pairwise and higher-order residue interactions, specifically as both a polynomial and Gaussian radial basis function (RBF)

$$\text{Polynomial} \qquad K(x, x') = (x^T x' + c)^d \tag{6}$$

$$\text{RBF} \qquad K(x, x') = \exp\left(-\gamma ||x - x'||^2\right) \tag{7}$$

where $d$ and $\gamma$ are additional kernel parameters and $c$ an arbitrary constant.

For a simple baseline comparison the Spectrum kernel [66] is also applied. A sequence is encoded as the count of each $k$-mer $l$ subsequence, whose characters are derived from an alphabet $\mathcal{A}$

$$\phi(x) = (\phi_l(x))_{l \in \mathcal{A}^k} \tag{8}$$

where $\phi(x)$ is a mapping of $x$ from an input space $\mathcal{X}$ into $\mathcal{F}$, $\phi_l(x)$ is the number of times $l$ occurs in $x$ and $k \in \{1, 2, 3, 4\}$. A $k$-mer size of 1 is simply the frequency of each amino acid within a sequence. In contrast, the use of $k$-mer sizes $\geq 2$ captures the co-occurrence of multiple consecutive residues, providing a simplified representation of residue–residue interactions.

## Evaluating SVR models

Once optimal hyperparameters are found (Supplementary material S5), $50 \times 5$-fold (80% training set, 20% test set) cross-validation (CV) is performed for all kernel functions. Pearson correlation coefficients ($r$) are recorded for each CV fold and the mean $r$ from the resulting 250 models is used to compare the kernel functions. In addition, the mean absolute error (MAE) is calculated for each CV fold. To test the statistical significance of the differences between models fitted with each kernel function, a two-sided unpaired Welch $t$ test at a 99% confidence interval is used. To compensate for bias from repeated CV, MAE and Fisher transformed $r$ values are generated from a single stratified tenfold CV. For stratified CV each fold has approximately the same mean target value and is representative of the full data set.

The predictive performance of SVR models when trained on sequence-activity data sets of varying size is evaluated using the following procedure [14]: (i) a subset of sequences are randomly sampled from the full data set, (ii) models are trained on this subset, (iii) the activity/fitness values for unsampled sequences are predicted, and (iv) the predictive ability of each model is evaluated by calculating its respective $r$ and MAE values. This procedure is repeated 1000 times for each sample size while increasing the size of the training sample within the range of 15–115 for the experimental and simulated *An*EH and CYP102A data sets. As the CYP102A data set includes a larger number of sequences, the procedure is extended to sample sizes from 115 to 215 and repeated 100 times to reduce computation time. For the experimental *An*EH data set, a single SVR model is constructed for each kernel function by training on the full set of 136 variants. The predictive ability of the resulting models is evaluated by predicting the $E$ values for an additional set of 16 mutants produced during previous protein engineering studies [43–45] (Supplementary material Table S6). The respective enantioselective preferences for these 16 mutants for (*S*)-glycidyl phenyl ether have also been characterised. For two of the mutants the reported $E$ values are calculated based on the relationship between the extent of conversion ($c$) and $e.e._p$ according to Eq. (9) [67].

$$E = \frac{\ln[1 - c(1 + e.e._p)]}{\ln[1 - c(1 - e.e._p)]} \tag{9}$$

$E$ values obtained through different methods will vary depending on a number of factors [46]. For consistency, reported values for $c$, $e.e._p$ and $e.e._s$ [43] are used to calculate the $E$ values using Eq. (1) for these two mutants. For nine of

these mutants amino acid content at one or more mutation sites is not seen in the training data at equivalent sites (Supplementary material Table S6).

### Implementation

An in-house application was developed in Java to construct and evaluate SVR models. The implementation is based on an adaption of the LIBSVM package, version 2.82 [68]. Code developed for this study is available on request to authors.

## Results

### Pairwise and higher-order models predict *E* values with improved accuracy for *An*EH variants

From the SVR 50 × 5-fold CV results for the experimentally derived *An*EH sequences (Table 1), the polynomial and RBF models all demonstrate similar predictive ability with a mean *r* of 0.91. The linear model produces a lower mean *r* of 0.84. Of the Spectrum kernel models, the 2- and 3-mer models perform best with a mean *r* of 0.89. The 1- and 4-mer models have a lower mean *r* of 0.83 and 0.84 respectively. Polynomial $d = 2$ and $d = 3$ models demonstrate the lowest mean MAEs, with values of 11.42 and 11.49 respectively. Other models have mean MAEs, from lowest to highest, of RBF: 12.60, Spectrum 2-mer: 12.96, Spectrum 3-mer: 13.14, Spectrum 4-mer: 14.31, Spectrum 1-mer: 15.92 and linear: 16.1. Comparing the mean MAE for each model type against the experimental error of the nine pairs of duplicate sequences (Supplementary material Table S4), mean MAEs are higher than the average experimental error (±11–16 vs ± 5.33). Mean *r* values and MAEs for *An*EH models and respective hyperparameters are summarised in Table 1.

Comparing the average predictions from 50 × 5-fold CV for models fitted with the linear and polynomial $d = 2$ functions (Fig. 1a), the polynomial $d = 2$ models have substantially lower error for sequences with higher *E* values,

**Table 1** 50 × 5-fold cross-validation results and hyperparameters for support vector regression (SVR) models. Models are trained and tested on experimentally derived sequence-activity data for wild type variants and mutants derived from *Aspergillus niger* epoxide hydrolase (*An*EH) and chimeric bacterial P450s (CYP102A1-3), and on simulated *An*EH sequences with additive fitnesses. The Pearson correlation coefficient (*r*) and mean absolute error (MAE) is calculated per fold and the average taken over 250 models. SVR models for the simulated *An*EH sequences have the same hyperparameters as the experimentally derived counterpart

| Data | Kernel function | Avg. Pearson ($\mu \pm \sigma$) | Avg. MAE ($\mu \pm \sigma$) | C | $\varepsilon$ | $\gamma$ |
|---|---|---|---|---|---|---|
| *An*EH | Linear | 0.84 ± 0.07 | 16.10 ± 3.04 | 1E + 01 | 1E − 03 | 10 |
| | Polynomial $d = 2$ | **0.91 ± 0.04** | **11.42 ± 2.19** | 1E + 01 | 1E − 03 | |
| | Polynomial $d = 3$ | **0.91 ± 0.03** | 11.49 ± 2.00 | 1E + 02 | 1E − 02 | |
| | RBF | **0.91 ± 0.03** | 12.60 ± 2.53 | 1E + 02 | 1E − 03 | |
| | Spectrum 1-mer | 0.84 ± 0.06 | 15.92 ± 2.79 | 1E − 01 | 1E − 01 | |
| | Spectrum 2-mer | 0.89 ± 0.04 | 12.96 ± 2.41 | 1E − 01 | 1E − 03 | |
| | Spectrum 3-mer | 0.89 ± 0.04 | 13.14 ± 2.54 | 1E − 01 | 1E − 02 | |
| | Spectrum 4-mer | 0.83 ± 0.06 | 14.31 ± 2.60 | 1E − 01 | 1E − 02 | |
| CYP102A | Linear | 0.90 ± 0.03 | 1.90 ± 0.21 | 1E + 03 | 1E − 05 | 0.001 |
| | Polynomial $d = 2$ | **0.93 ± 0.02** | **1.60 ± 0.17** | 1E − 03 | 1E − 02 | |
| | Polynomial $d = 3$ | 0.92 ± 0.02 | 1.70 ± 0.19 | 1E − 04 | 1E − 02 | |
| | RBF | 0.91 ± 0.02 | 1.79 ± 0.19 | 1E + 03 | 1E − 02 | |
| | Spectrum 1-mer | 0.89 ± 0.03 | 1.92 ± 0.20 | 1E − 02 | 1E − 03 | |
| | Spectrum 2-mer | 0.90 ± 0.03 | 1.89 ± 0.22 | 1E − 03 | 1E − 02 | |
| | Spectrum 3-mer | 0.90 ± 0.03 | 1.89 ± 0.21 | 1E − 03 | 1E − 05 | |
| | Spectrum 4-mer | 0.89 ± 0.03 | 1.89 ± 0.23 | 1E − 03 | 1E − 03 | |
| Simulated *An*EH | Linear | **0.82 ± 0.06** | 13.98 ± 2.58 | – | – | – |
| | Polynomial $d = 2$ | **0.82 ± 0.06** | **12.40 ± 2.05** | – | – | |
| | Polynomial $d = 3$ | 0.79 ± 0.07 | 13.59 ± 2.08 | – | – | |
| | RBF | 0.80 ± 0.07 | 13.99 ± 2.70 | – | – | |
| | Spectrum 1-mer | 0.62 ± 0.10 | 20.24 ± 3.19 | – | – | |
| | Spectrum 2-mer | 0.77 ± 0.07 | 16.61 ± 2.50 | – | – | |
| | Spectrum 3-mer | 0.78 ± 0.07 | 14.91 ± 2.46 | – | – | |
| | Spectrum 4-mer | 0.76 ± 0.08 | 14.81 ± 2.75 | – | – | |

The best mean Pearson correlation and MAE values for each data set are highlighted in bold

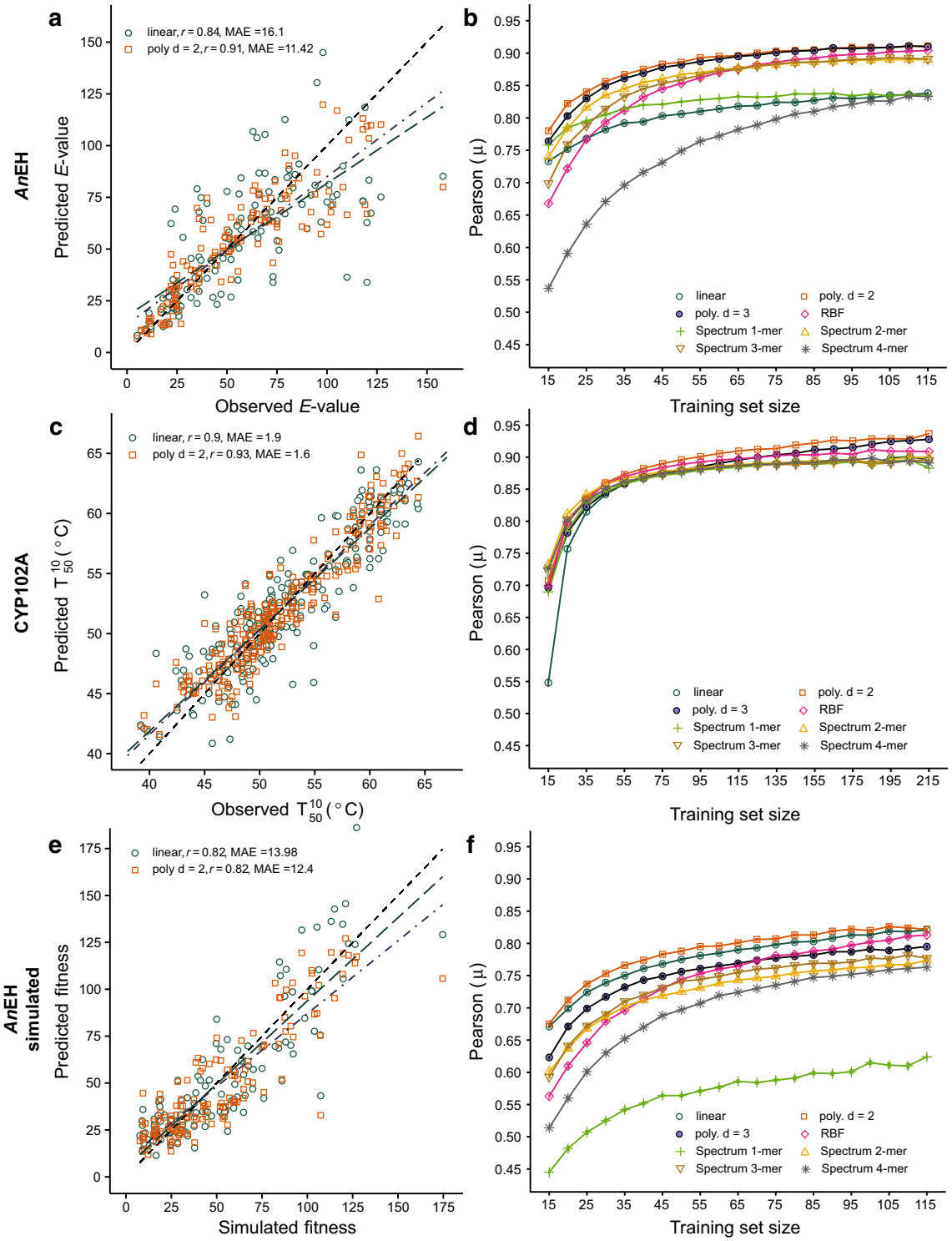


**Fig. 1** Observed vs average predicted (over 250 models) **a** *E* values for *An*EH variants, **c** thermostabilities for chimeric bacterial P450s (CYP102A1-3) and **e** simulated additive *E* values for *An*EH variants from 50 × 5-fold cross-validation of linear (long-dash) and polynomial $d = 2$ (dash-dot) models. Long-dash and dash-dot lines are linear models fitted to observed vs predicted values, the diagonal dashed line indicates perfect agreement between observed and predicted values. **b**, **d**, **f** The change in the mean Pearson correlation coefficient ($r$) as the number of sequences trained on is increased. Standard error bars have been included for each point

i.e. those with the strongest cooperative epistatic interactions. The mean MAE for sequences with $E$ values $\geq 100$ is 41.04 and 26.34 for linear and polynomial $d = 2$ models respectively (Supplementary material Table S7). In contrast the MAE is 12.56 and 9.24 for sequences whose $E$ values are $< 100$. Gradually increasing the number of sequences trained on from 15 to 115, the polynomial models have higher mean $r$ values across all training set sizes compared to other models (Fig. 1b). RBF, Spectrum 2- and 3-mer models have slightly lower average $r$ values, while linear and Spectrum 1- and 4-mer models display markedly lower $r$ values compared to the other model types across most sample sizes. On average when trained on approximately 100 sequences, linear and Spectrum 1- and 4-mer models have the same predictive power as the polynomial models trained on approximately 20 sequences ($r \approx 0.8$). Similar results are observed for mean MAE values—the polynomial models have a lower mean MAE across all sample sizes compared to all other models (Supplementary material Fig. S8). On average polynomial models require approximately 40 sequences to achieve a mean MAE of 10, whereas linear and Spectrum 1- and 4-mer models require approximately 60.

Comparing the distributions of $r$ values from the stratified tenfold CV (Fig. 2a), functions that result in significant improvements in model predictive ability compared to linear models include the polynomial $d = 2$ ($p$ value $\leq 0.01$), $d = 3$ ($p$ value $\leq 0.01$), RBF ($p$ value $\leq 0.01$) and Spectrum 3-mer ($p$ value $\leq 0.05$) functions. MAE distributions produced by polynomial $d = 2$ ($p$ value $\leq 0.01$), $d = 3$ ($p$ value $\leq 0.01$) and RBF ($p$ value $\leq 0.05$) models are significantly improved compared to linear models. Significance values for the comparison of all functions from stratified tenfold CV are provided as Supplementary material Fig. S9 (a and b).

For the test set of 16 *An*EH mutants (Fig. 3) the polynomial $d = 2$ and $d = 3$ models have the highest correlation between observed and predicted $E$ values, with $r$ values

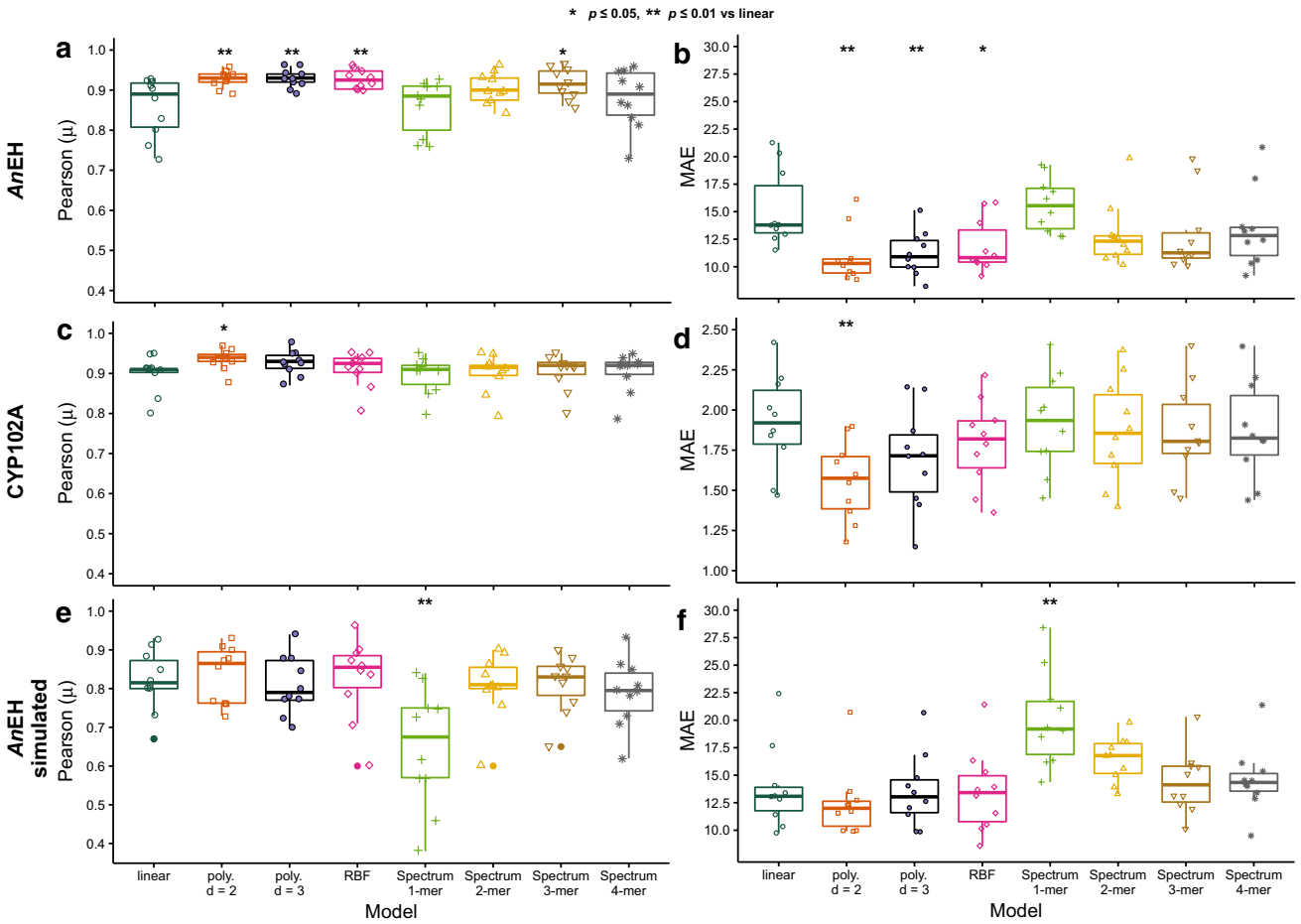


**Fig. 2** Distributions of **a**, **c**, **e** Pearson correlation coefficients ($r$) and **b**, **d**, **f** mean average errors (MAE) from stratified tenfold CV for the various model types. Models are trained and tested on either **a**, **b** *An*EH variants and $E$ values, **c**, **d** CYP102A1-3 chimeras and thermostability values or **e**, **f** simulated *An*EH variants with additive fitnesses. Significance $p$ values are calculated using a two-sided unpaired Welch $t$ test with a confidence interval of 99%, comparing $r$ and MAE distributions for all model types against those from linear models

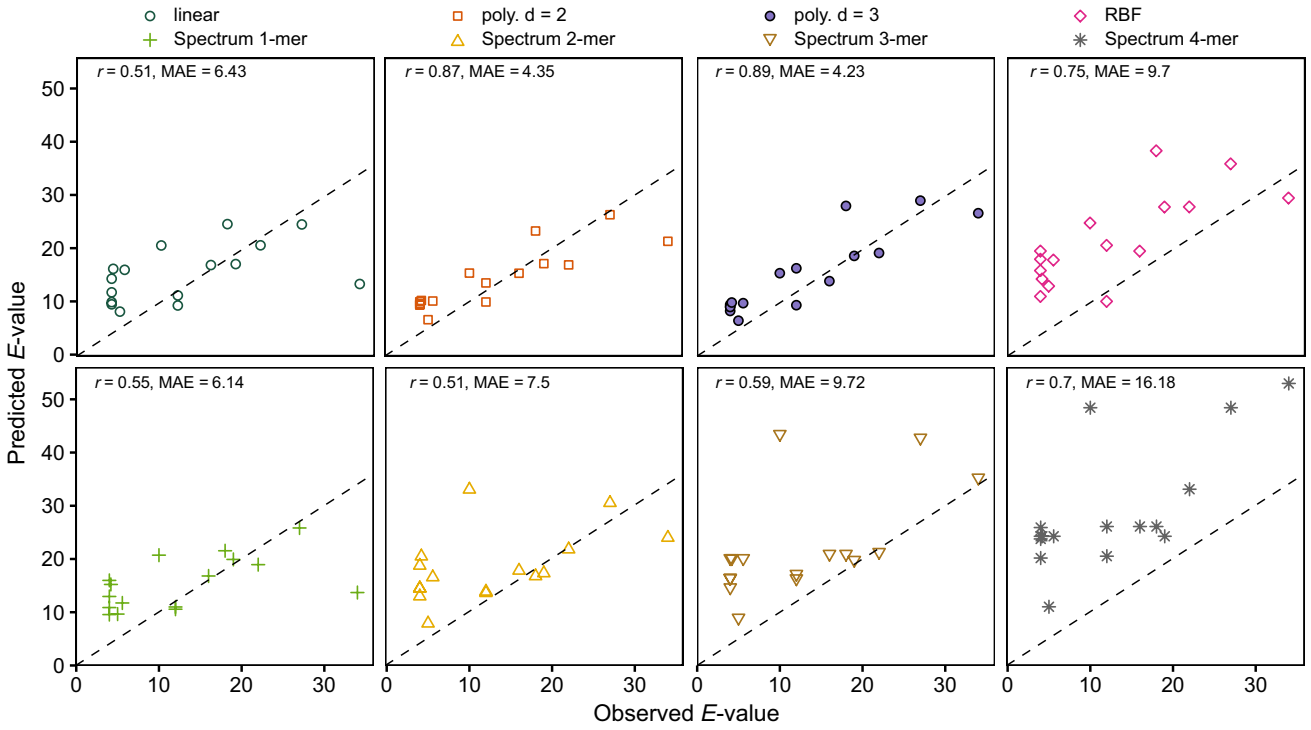


**Fig. 3** *E* value predictions for 16 *An*EH mutants from SVR models trained on the full data set of 136 *An*EH variants

of 0.87 and 0.89 respectively. RBF and Spectrum 4-mer models also display relatively high *r* values of 0.75 and 0.7 respectively. Other models have lower predictive power with *r* values of Spectrum 3-mer: 0.59, Spectrum 1-mer: 0.55, Spectrum 2-mer: 0.51 and linear: 0.51. Polynomial $d = 2$ and $d = 3$ models also have the lowest error with MAEs of 4.35 and 4.23 respectively. Other models have MAEs, from lowest to highest, of Spectrum 1-mer: 6.14, linear: 6.43, Spectrum 2-mer: 7.5, RBF: 9.7, Spectrum 3-mer: 9.72 and Spectrum 4-mer: 16.18.

## Minimisation of epistatic interactions in CYP102A variants results in similar accuracy across all models

For the CYP102A data set, the polynomial $d = 2$, $d = 3$ and RBF models have approximately the same predictive ability with mean *r* values from 50 × 5-fold CV of 0.93, 0.92 and 0.91 respectively. The linear and Spectrum models also all have *r* values of approximately 0.90. Polynomial $d = 2$ and $d = 3$ models have the lowest mean MAEs with values of 1.60 and 1.70. All other models have mean MAEs of approximately 1.9. Mean *r* and MAE values for each CYP102A model and respective hyperparameters are summarised in Table 1. Average predictions from 50 × 5-fold CV for linear and polynomial $d = 2$ models (Fig. 1c) shows both models are similar in accuracy for sequences across the full range of thermostabilities. Gradually increasing the number of training sequences from 15 to 115 shows the polynomial and RBF models marginally outperform other models in terms of the mean *r* value across most sample sizes (Fig. 1d). At higher sample sizes (> 115), mean *r* values for polynomial and RBF models continue to improve up to approximately 0.92 while *r* values for linear and Spectrum models level off at approximately 0.87.

From the stratified tenfold CV, only the polynomial $d = 2$ models demonstrate significant improvement in *r* ($p$ value $\leq 0.05$) and MAE ($p$ value $\leq 0.01$) values compared to those models fitted with a linear function (Fig. 2c, d). Significance values for the comparison of all functions from stratified tenfold CV are provided as Supplementary material Fig. S9 (c and d).

## A lack of epistatic interactions results in no gain in accuracy from pairwise and higher-order functions in simulated *An*EH sequences

For the simulated *An*EH sequences, with the exception of the Spectrum 1-mer function, models produce a mean *r* from 50 × 5 CV of approximately 0.8. The *r* values from highest to lowest being polynomial $d = 2$ and linear: 0.82, RBF: 0.8, polynomial $d = 3$: 0.79, Spectrum 3-mer: 0.78, Spectrum 2-mer: 0.77 and Spectrum 4-mer: 0.76. Models fitted with a Spectrum 1-mer function have a substantially lower mean *r* value of 0.62. The mean MAEs produced by most models

are similar and are polynomial $d = 2$: 12.4, polynomial $d = 3$: 13.59, linear: 13.98, RBF: 13.99, Spectrum 4-mer: 14.81 and Spectrum 3-mer: 14.91. Spectrum 1- and 2-mer models have higher mean MAEs of 20.24 and 16.61 respectively. Mean $r$ values and MAEs for the simulated *An*EH models and respective hyperparameters are summarised in Table 1. Comparing the average predictions from $50 \times 5$-fold CV for models fitted with the linear and polynomial $d = 2$ functions (Fig. 1e) shows both models produce similar predictions and error for the full range of simulated $E$ values.

Gradually increasing the size of the training data from 15 to 115, linear and polynomial $d = 2$ models have approximately equal $r$ values at all samples sizes (Fig.1f). Polynomial $d = 3$ and RBF models (at sample sizes $> 65$) have marginally lower mean $r$ values. Spectrum 2- and 3-mer models have slightly lower mean $r$ values compared to polynomial $d = 3$ models at all sample sizes. Spectrum 4-mer models on average require $> 75$ sequences to have $r$ values approximately equal to Spectrum 2- and 3-mer models ($r \approx 0.73$ ). The Spectrum 1-mer models have substantially lower mean $r$ values compared to all other models, only achieving a maximum $r$ value of approximately 0.6 at sample sizes of $> 90$. Differences in the mean $r$ values between linear, polynomial $d = 2$ and $d = 3$ models, and between Spectrum 2- and 3-mer models, are largely reduced when removing the error randomly assigned to the simulated fitness values (Supplementary material Fig. S10).

The distributions of $r$ and MAE values from stratified tenfold CV (Fig. 2e, f) show that models fitted with any of the kernel functions, with the exception of Spectrum 1-mer, are not significantly different from those fitted with a linear function. Models fitted with the Spectrum 1-mer function have significantly lower $r$ ($p$ value $\leq 0.01$) and higher MAE ($p$ value $\leq 0.01$) values compared to linear models. Significance values for the comparison of all functions from stratified tenfold CV are provided as Supplementary material Fig. S9 (e and f).

## Discussion and conclusions

There are a number of significant challenges that are faced in the engineering of new and useful biocatalysts [69]. One challenge is the presence of epistatic interactions which, although potentially beneficial to the activity of an enzyme, are difficult to study experimentally. Computational methods can be applied to capture and model the complex relationship between residues and enzyme activity [27]. The use of structural data, though very informative, will assume that the crystal structure is representative of reaction conditions. By developing predictive models from experimental data it is possible to implicitly capture the factors that contribute to the activity of an enzyme. These models can guide exploration of the fitness landscape to those areas more likely to yield proteins with useful properties [8–11, 14, 23, 29–31, 70, 71]. As more cost-effective assaying and sequencing technologies are developed, the need for methods that can learn from characterised sequences and guide protein design will increase.

In this study we demonstrate that SVR models representing pairwise and higher-order residue interactions, i.e. with polynomial and RBF kernel functions, predict enantioselectivity-enhancing mutations for a set of experimentally characterised *An*EH variants with significantly improved accuracy compared to models simply using amino acid frequencies or linear representations. Evaluating models on a control set of simulated *An*EH sequences with additive fitnesses and an additional set of *An*EH mutants with experimentally characterised $E$ values supports these observations. Models representing residue interactions also explain more of the variation in enantioselectivity measurements, able to learn from smaller sequence-activity data sets. For the experimental *An*EH sequences it is interesting to note that models fitted with the Spectrum 1-mer function, representing sequences simply as their respective amino acid frequencies, perform largely equivalently to those models fitted with a linear function. When fitted with Spectrum kernel functions with $k$-mer sizes of 2 and 3, models also display comparatively high predictive ability, likely due to the simplified representation of residue–residue interactions. The lower predictive ability resulting from the use of larger $k$-mer sizes is likely due to the generation of extremely sparse sequence encodings, i.e. most large $k$-mers will not appear in the training set of sequences and receive values of zero, or only appear once.

A major concern for predictive models is whether they are overfitting the data. One indicator of overfitting is when model error is lower than experimental error. The focus of the present study is the modelling of enantioselectivity $E$ values for *An*EH variants. The error in $E$ value measurements for biocatalysis is rarely reported in the literature, partly due to the difficulty in comparing values from different calculation methods and reaction conditions, e.g. pH and temperature. Where experimental error has been reported, values range from less than $\pm$ 5 [72–76] to (significantly) higher [72, 73, 77]. We estimate the true experimental error by referring to the *An*EH sequences with multiple $E$ value measurements (Supplementary material Table S4), whose average error is $\pm$ 5.33. The errors of the predictions (calculated as MAEs) are generally greater than the errors gauged from the experimental data ($\pm$11–16), meaning that overfitting does not explain the differences in prediction accuracy between the functions used to fit the models.

The study also demonstrates that if a library design strategy is used that partitions epistatic interactions into structural sub-units, such as SCHEMA-guided recombination,

models based on amino acid frequencies or assumptions of additivity will have predictive accuracies largely equivalent to pairwise and higher-order counterparts. However, some additional predictive power can be gained from pairwise and higher-order models when they are constructed on a greater number of sequences. These observations are exemplified by the prediction of thermostabilities for the set of chimeric bacterial P450s. The results from this study highlight the sensitivity of different engineering strategies to epistatic interactions. The choice of strategy should therefore be considered carefully given its implications for the predictability of enzyme activity in computational studies.

**Acknowledgements** The authors thank G. Foley (School of Chemistry and Molecular Biosciences, University of Queensland) for reviewing the manuscript. Funding to support this study was provided by the Australian Research Council Discovery Project scheme (DP160100865) and by the Australian Government Research Training Program (RTP).

## References

1. Agranat I, Caner H, Caldwell J (2002) Putting chirality to work: the strategy of chiral switches. Nat Rev Drug Discov 1(10):753–768
2. Agranat I, Wainschtein SR, Zusman EZ (2012) The predicated demise of racemic new molecular entities is an exaggeration. Nat Rev Drug Discov 11(12):972–973
3. Branch SK, Agranat I (2014) "New drug" designations for new therapeutic entities: new active substance, new chemical entity, new biological entity, new molecular entity. J Med Chem 57(21):8729–8765
4. Morley KL, Kazlauskas RJ (2005) Improving enzyme properties: when are closer mutations better? Trends Biotechnol 23(5):231–237
5. Miton CM, Tokuriki N (2016) How mutational epistasis impairs predictability in protein evolution and design. Protein Sci. 25(7):1260–1272
6. Starr TN, Thornton JW (2016) Epistasis in protein evolution. Protein Sci 25(7):1204–1218
7. Kondrashov DA, Kondrashov FA (2015) Topological features of rugged fitness landscapes in sequence space. Trends Genet 31(1):24–33
8. Li Y, Drummond DA, Sawayama AM, Snow CD, Bloom JD, Arnold FH (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. Nat Biotechnol 25(9):1051–1056
9. Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK, Chung LM, Ching C, Tam S, Muley S, Grate J, Gruber J, Whitman JC, Sheldon RA, Huisman GW (2007) Improving catalytic function by ProSAR-driven enzyme evolution. Nat Biotechnol 25(3):338–344
10. Liao J, Warmuth MK, Govindarajan S, Ness JE, Wang RP, Gustafsson C, Minshull J (2007) Engineering proteinase K using machine learning and synthetic genes. BMC Biotechnol 7(1):16
11. Romero PA, Arnold FH (2012) Random field model reveals structure of the protein recombinational landscape. PLoS Comput Biol 8(10):e1002,713
12. Fox R (2005) Directed molecular evolution by machine learning and the influence of nonlinear interactions. J Theor Biol 234(2):187–199
13. Buske FA, Their R, Gillam EMJ, Bodén M (2009) In silico characterization of protein chimeras: Relating sequence and function within the same fold. Proteins 77(1):111–120
14. Romero PA, Krause A, Arnold FH (2013) Navigating the protein fitness landscape with Gaussian processes. Proc Natl Acad Sci (USA) 110(3):E193–201
15. Funar-Timofei S, Suzuki T, Paier JA, Steinreiber A, Faber K, Fabian WMF (2003) Quantitative structure-activity relationships for the enantioselectivity of oxirane ring-opening catalyzed by epoxide hydrolases. J Chem Inf Comput Sci 43(3):934–940
16. Caetano S, Aires-de Sousa J, Daszykowski M, Heyden YV (2005) Prediction of enantioselectivity using chirality codes and classification and regression trees. Anal Chim Acta 544(1–2):315–326
17. Gu J, Liu J, Yu H (2011) Quantitative prediction of enantioselectivity of *Candida antarctica* lipase B by combining docking simulations and quantitative structure–activity relationship (QSAR) analysis. J Mol Catal B 72(3–4):238–247
18. Hartman JH, Cothren SD, Park SH, Yun CH, Darsey JA, Miller GP (2013) Predicting CYP2C19 catalytic parameters for enantioselective oxidations using artificial neural networks and a chirality code. Bioorg Med Chem 21(13):3749–3759
19. Tomić S, Kojić-Prodić B (2002) A quantitative model for predicting enzyme enantioselectivity: application to *Burkholderia cepacia* lipase and 3-(aryloxy)-1,2-propanediol derivatives. J Mol Graph Model 21(3):241–252
20. Wijma HJ, Marrink SJ, Janssen DB (2014) Computationally efficient and accurate enantioselectivity modeling by clusters of molecular dynamics simulations. J Chem Inf Model 54(7):2079–2092
21. Wijma HJ, Floor RJ, Bjelic S, Marrink SJ, Baker D, Janssen DB (2015) Enantioselective enzymes by computational design and in silico screening. Angew Chem Int Ed 54(12):3726–3730
22. Braiuca P, Lorena K, Ferrario V, Ebert C, Gardossi L (2009) A three-dimensional quanititative structure-activity relationship (3D-QSAR) model for predicting the enantioselectivity of *Candida antarctica* Lipase B. Adv Synth Catal 351(9):1293–1302
23. Feng X, Sanchis J, Reetz MT, Rabitz H (2012) Enhancing the efficiency of directed evolution in focused enzyme libraries by the adaptive substituent reordering algorithm. Chem Eur J 18(18):5646–5654
24. Liang J, Mundorff E, Voladri R, Jenne S, Gilson L, Conway A, Krebber A, Wong J, Huisman G, Truesdell S, Lalonde J (2010) Highly enantioselective reduction of a small heterocyclic ketone: biocatalytic reduction of tetrahydrothiophene-3-one to the corresponding (R)-alcohol. Org Process Res Dev 14(1):188–192
25. Chaput L, Sanejouand YH, Balloumi A, Tran V, Graber M (2012) Contribution of both catalytic constant and Michaelis constant to CALB enantioselectivity: Use of FEP calculations for prediction studies. J Mol Catal B 76:29–36
26. Noey EL, Tibrewal N, Jiménez-Osés G, Osuna S, Park J, Bond CM, Cascio D, Liang J, Zhang X, Huisman GW, Tang Y, Houk KN (2015) Origins of stereoselectivity in evolved ketoreductases. Proc Natl Acad Sci (USA) 112(51):E7065–72
27. Minshull J, Ness JE, Gustafsson C, Govindarajan S (2005) Predicting enzyme function from protein sequence. Curr Opin Chem Biol 9(2):202–209
28. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge
29. Bedbrook CN, Yang KK, Rice AJ, Gradinaru V, Arnold FH (2017) Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. PLoS Comput Biol 13(10):e1005,786
30. Romero P, Stone E, Lamb C, Chantranupong L, Krause A, Miklos A, Hughes R, Fechtel B, Ellington A, Arnold FH (2012) SCHEMA-designed variants of human Arginase I and II reveal

sequence elements important to stability and catalysis. ACS Synth Biol 1(6):221–228
31. Smith MA, Rentmeister A, Snow CD, Wu T, Farrow MF, Mingardon F, Arnold FH (2012) A diverse set of family 48 bacterial glycoside hydrolase cellulases created by structure-guided recombination. FEBS J 279(24):4453–4465
32. Pissurlenkar RRS, Malde AK, Khedkar SA, Coutinho EC (2007) Encoding type and position in peptide QSAR: application to peptides binding to class I MHC molecule HLA-A*0201. Mol Inform 26(2):189–203
33. Verma J, Khedkar VM, Prabhu AS, Khedkar SA, Malde AK, Coutinho EC (2008) A comprehensive analysis of the thermodynamic events involved in ligand–receptor binding using CoRIA and its variants. J Comput Aided Mol Des 22(2):91–104
34. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH (2002) Protein building blocks preserved by recombination. Nat Struct Biol 9(7):553–558
35. Silberg JJ, Endelman JB, Arnold FH (2004) SCHEMA-guided protein recombination. Meth Enzymol 388:35–42
36. Zaugg J, Gumulya Y, Gillam EMJ, Bodén M (2014) Computational tools for directed evolution: a comparison of prospective and retrospective strategies. Methods Mol Biol 1179:315–333
37. Endelman JB, Silberg JJ, Wang ZG, Arnold FH (2004) Site-directed protein recombination as a shortest-path problem. Protein Eng Des Sel 17:589–594
38. Heinzelman P, Snow CD, Wu I, Nguyen C, Villalobos A, Govindarajan S, Minshull J, Arnold FH (2009) A family of thermostable fungal cellulases created by structure-guided recombination. Proc Natl Acad Sci (USA) 106(14):5610–5615
39. Packer MS, Liu DR (2015) Methods for the directed evolution of proteins. Nat Rev Genet 16(7):379–394
40. Reetz MT, Bocola M, Carballeira JD, Zha D, Vogel A (2005) Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. Angew Chem Int Ed 44(27):4192–4196
41. Reetz MT, Carballeira JD (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. Nat Protoc 2(4):891–903
42. Gumulya Y, Sanchis J, Reetz MT (2012) Many pathways in laboratory evolution can lead to improved enzymes: how to escape from local minima. Chembiochem 13(7):1060–1066
43. Reetz MT, Wang LW, Bocola M (2006) Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space. Angew Chem 118(8):1258–1263
44. Reetz MT, Sanchis J (2008) Constructing and analyzing the fitness landscape of an experimental evolutionary process. Chembiochem 9(14):2260–2267
45. Wang LW (2006) Directed evolution of the *Aspergillus niger* Epoxide Hydrolase. PhD thesis, Ruhr-Universität Bochum, Bochum
46. Straathof AJJ, Jongejan JA (1997) The enantiomeric ratio: origin, determination and prediction. Enzyme Microb Technol 21(8):559–571
47. Faber K (2011) Biotransformations In Organic Chemistry, 6th edn. Springer, Berlin
48. Rakels JL, Straathof AJ, Heijnen JJ (1993) A simple method to determine the enantiomeric ratio in enantioselective biocatalysis. Enzyme Microb Technol 15(12):1051–1056
49. Kauffman SA, Weinberger ED (1989) The NK model of rugged fitness landscapes and its application to maturation of the immune response. J Theor Biol 141(2):211–245
50. Fox R, Roy A, Govindarajan S, Minshull J, Gustafsson C, Jones JT, Emig R (2003) Optimizing the search algorithm for protein engineering by directed evolution. Protein Eng 16(8):589–597
51. Vapnik VN, Vapnik V (1998) Statistical learning theory, vol 1. Wiley, New York
52. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
53. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222
54. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G (2008) Support vector machines and kernels for computational biology. PLoS Comput Biol 4(10):e1000,173
55. van Westen GJP, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender A (2011) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. Med Chem Commun 2(1):16–30
56. Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28:374
57. Saraf MC, Horswill AR, Benkovic SJ, Maranas CD (2004) FamClash: a method for ranking the activity of engineered enzymes. Proc Natl Acad Sci (USA) 101(12):4142–4147
58. Pantazes RJ, Saraf MC, Maranas CD (2007) Optimal protein library design using recombination or point mutations based on sequence-based scoring functions. Protein Eng Des Sel 20(8):361–373
59. Sulimova V, Mottl V, Kulikowski C, Muchnik I (2008) Probabilistic evolutionary model for substitution matrices of PAM and BLOSUM families. DIMACS Tech Report
60. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. Atlas Protein Seq Struct 5:345–358
61. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8(3):275–282
62. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18(5):691–699
63. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25(7):1307–1320
64. Liò P, Goldman N (1998) Models of molecular evolution and phylogeny. Genome Res 8(12):1233–1244
65. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10,915–10,919
66. Leslie CS, Eskin E, Noble WS (2002) The spectrum kernel: a string kernel for svm protein classification. In: Pacific symposium on biocomputing, Hawaii, USA, vol 7, pp 566–575
67. Chen CS, Fujimoto Y, Girdaukas G, Sih CJ (1982) Quantitative analyses of biochemical kinetic resolutions of enantiomers. J Am Chem Soc 104(25):7294–7299
68. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27–27
69. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K (2012) Engineering the third wave of biocatalysis. Nature 485(7397):185–194
70. Ness JE, Cox T, Govindarajan S, Gustafsson C, Gross RA, Minshull J (2005) Empirical biocatalyst engineering: escaping the tyranny of high-throughput screening. ACS Symp Ser 900:37–50
71. van den Berg BA, Reinders MJT, van der Laan JM, Roubos JA, de Ridder D (2014) Protein redesign by learning from data. Protein Eng Des Sel 27(9):281–288
72. Dai DZ, Xia LM (2006) Resolution of (R, S)-2-octanol by *Penicillium expansum* PED-03 lipase immobilized on modified ultrastable-Y molecular sieve in microaqueous media. Process Biochem 41(6):1455–1460
73. Berglund P, Holmquist M, Hult K, Högberg HE (1995) Alcohols as enantioselective inhibitors in a lipase catalysed esterification of a chiral acyl donor. Biotechnol Lett 17(1):55–60
74. Machado SS, Wandel U, Jongejan JA, Straathof AJ, Duine JA (1999) Characterization of the enantioselective properties of the quinohemoprotein alcohol dehydrogenase of *Acetobacter pasteurianus* LMG 1635. 1. different enantiomeric ratios of whole cells

and purified enzyme in the kinetic resolution of racemic glycidol. Biosci Biotechnol Biochem 63(1):10–20
75. Horsman GP, Liu AMF, Henke E, Bornscheuer UT, Kazlauskas RJ (2003) Mutations in distant residues moderately increase the enantioselectivity of *Pseudomonas fluorescens* esterase towards methyl 3-bromo-2-methylpropanoate and ethyl 3-phenylbutyrate. Chem Eur J 9(9):1933–1939
76. Sun Z, Wikmark Y, Bäckvall JE, Reetz MT (2016) New concepts for increasing the efficiency in directed evolution of stereoselective enzymes. Chem Eur J 22(15):5046–5054
77. Léonard V, Fransson L, Lamare S, Hult K, Graber M (2007) A water molecule in the stereospecificity pocket of *Candida antarctica* lipase B enhances enantioselectivity towards pentan-2-ol. Chembiochem 8(6):662–667