Prediction of enzyme catalysis by computing reaction energy barriers via steered QM/MM Molecular Dynamics Simulations and Machine Learning

Daniel Platero-Rochart,[†] Tatyana Krivobokova,[‡] Michael Gastegger,[¶] Gilbert Reibnegger,[†] and Pedro A. Sánchez-Murcia[*,†]

[†]*Laboratory of Computer-Aided Molecular Design, Division of Medicinal Chemistry, Otto-Loewi Research Center, Medical University of Graz, Neue Stiftintalstr. 6/III, A-8010 Graz, Austria*

[‡]*Department of Statistics and Operations Research, University of Vienna, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria*

[¶]*Institute of Software Engineering and Theoretical Computer Science, Machine Learning Group, Technische Universität, 10587 Berlin, Germany*

E-mail: pedro.murcia@medunigraz.at

Abstract

The prediction of enzyme activity in a general extend is maybe one of the main challenges nowadays in catalysis. Computer-assisted methods have been proven to be able to simulate the reaction mechanism at the atomic level of detail. However, these methods tend to be expensive to be used in a large scale as it is needed in protein engineering campaigns. To alleviate this situation, machine learning methods can help

in the generation of predictive-decision models. Herein we train different regression algorithms for the prediction of the reaction energy barrier of the rate-limiting step of the hydrolysis of mono-(2-hydroxyethyl)terephthalic acid by the MHETase of *Ideonella sakaiensis*. As training data set we use steered QM/MM MD simulation snapshots and their corresponding pulling work values. We have explored three algorithms together with three chemical representations. As outcome, our trained models are able to predict pulling works along the steered QM/MM MD simulations with a mean absolute error below 3 kcal mol^{-1} and a score value above 0.90. More challenging is the prediction of the energy maximum with a single geometry. Whereas the use of the initial snapshot of the QM/MM MD trajectory as input geometry yields a very poor prediction of the reaction energy barrier, the use of an intermediate snapshot of the former trajectory brings the score value above 0.40 with a low mean absolute error (ca. 3 kcal mol^{-1}). Altogether, in this work we have faced some initial challenges of the final goal of getting an efficient workflow for the semi-automatic prediction of enzyme-catalyzed energy barriers and catalytic efficiencies.

Introduction

Inspired by nature, humankind has manipulated enzymes for their particular use for a long time. The impact of these biocatalysts in our lives is tremendous in the feed&food industry, bioremediation, pharmaceutical production or industrial biocatalysis.[1–4] The significant advances in molecular biology techniques[5] in the last decades have converted cloning and expression of enzymatic variants into a common task in laboratories around the world. Nowadays, properties like (thermo)stability[6–8] or substrate specificity[9] in enzymes can be modulated by engineering campaigns. However, experimental procedures like directed evolution (DE),[10,11] can deliver unpredictable outcomes and require extensive effort of sequencing and deconvolution. In this regards high hopes are placed in *in-silico* design due to their intrinsic lower cost compared to experimental campaigns. Most of the current computational strate-

gies for protein redesign rely on getting more stable protein scaffolds by computing the total protein energy after modification of the amino acid sequence (e.g., Funclib,[12] PROSS[13] or HotSpotWizard[14]). Another group of methods are based on the assumption that, since the catalytic activity can be governed by the reaction energy barrier of the rate-limiting step,[15,16] one enzyme variant will be catalytically more efficient than a second one when having a reduced energy barrier. As examples of reaction-barrier-based methods are, amongst others, cluster method,[17] metadynamics,[18] umbrella sampling,[19] path collective variable sampling[20] or empirical valence bond.[21] All these methods rely on the main assumption that in enzyme catalysis the observed rate acceleration is caused by transition state (TS) stabilization with respect to the substrate.[22] Within this group of methods, steered QM/MM molecular dynamics (sMD)[23,24] has been shown as a good alternative to simulate enzyme reactivity at the active site in a reasonable time scale. By applying harmonic forces on selected atoms, the system can be forced to explore a proposed reaction coordinate (RC). As outcome, sMD delivers pulling work values along the RC. As illustration, in the last years some of us have used sMD for the industrial design of novel thiolases,[25] to explain the reaction mechanism in glycosyltransferases,[26] or to understand the stereoselectivity observed in a ω-transaminase.[27]

Unfortunately, reaction-barrier-based design methods are limited by their computational cost. Calculating a single barrier using any of the above mentioned methods is already an expensive process. Moreover, barrier evaluations can involve a substantial trial and error component depending on the initial geometries/snapshots and the chosen level of theory. Alternative strategies to obtain free energy barriers while circumventing these problems are therefore nvaluable for successful enzyme-engineering campaigns. Machine learning (ML) methods have emerged as a promising tool in computational chemistry in general,[28–31] and in enzyme engineering[32] in particular. Examples include the prediction of Enzyme Commission (EC) numbers from enzyme sequence and substrate enzyme complex structures,[33] and the automated identification of chemical features promoting enzyme catalysis.[34] The deep learning approach DLKcat[35] is able to predict accurate catalytic constants k_{cat} by combining

graph neural networks (GNNs) and convolutional neural networks (CNNs) to represent substrates and proteins. DLKcat is trained on experimental kinetic data from the BRENDA[36] and SABIO-RK databases.[37] Finally, there are several examples where ML is used to predict the activation barriers of chemical reactions.[35]

The goal of this work is the use of ML models for the prediction of enzyme-catalyzed energy barriers using input data derived from chemical dynamics trajectories. In particular, we explore the possibility of using sMD simulations to generate training data for the ML models. Based on the obtained data set, we compare the performance of different combinations of regression algorithms and ML representations of the active site.[38] The former algorithms include kernel ridge regression (KRR),[39] support vector regression (SVR),[40] and ElasticNet.[41] On the descriptor side, the Coulomb matrix, atom-centered symmetry functions (ACSF)[42] and smooth overlap of atomic positions (SOAP) were used.

Computational Details

Molecular Dynamics (MD) simulations

The chain A of the crystallographic structure of the native *Ideonella sakaiensis* MHETase (1.8 Å resolution, PDB id 6QZ4[43]) was selected for our MD simulations. The protonation states of the tritatable residues of the protein were computed via H++ server[44] at pH 7.0. Residues, His91, His241, His293, His467, and His488 are single protonated at the epsilon-nitrogen (HIE), His528 single protonated at delta-nitrogen (HID), and His166 doubly protonated (HIP). Five disulfide bonds were defined: Cys51-Cys92, Cys224-Cys529, Cys303-Cys320, Cys340-Cys348, Cys577-Cys599. The protein residues were described with the AMBER force field parameters ff19SB.[45] The modified side of chain of Ser225 was modified manually to include the terepthalic acid (TPA) ester. The atoms of this modified residue Ser225(TPA) were defined as GAFF/AMBER atom types and the point charges of these molecules were derived using a RESP model (RHF/6-31G**) on a DFT-optimized geometry (B3LYP/6-

31+G*)[46–49] with antechamber. All QM optimizations were run with Gaussian16, v. C.01.[50] The system was embedded in a truncated octahedron of TIP3P water molecules[51] and the solvated system was neutralized using Na^+ and Cl^- ions by random substitution of water molecules. Before production, the system was energetically minimized in three steps, where all protons, the solvent molecules, and the entire system, were gradually relaxed, respectively. Then, the system was heated up using the Langevin thermostat (1 ps^{-1} collision frequency) from 100 to 300 K in 1 ns with a linear increase of the temperature in a NVT ensemble. For this step, all atoms of the solute were restrained by means of application of a harmonic force (40 kcal mol^{-1} A^{-2}). Finally, the restrains were gradually removed in six steps, where the last two steps were run under a NPT ensemble. The system was further simulated for 1 μs at 300 K with a time step of 2 fs. Long-range interactions were calculated using Particle Mesh Ewald summations[52] under Periodic Boundary Conditions and a cutoff of 10 Å was defined for non-bonded interactions. The SHAKE algorithm was applied to all water molecules. MD simulations were run in Amber20[53] and the MD trajectories were analyzed using *cpptraj* v5.1.0.[54]

Steered QM/MM MD (sMD) simulations

The hybrid quantum mechanics/molecular mechanics (QM/MM) simulations were performed with *sander* in Amber20.[53] The QM region encompassed the side chain of residues His528, Ser225(TPA), Asp492, each cut across the C_α/C_β bond, and the closest molecule of water (39 atoms and a net charge of -2). The chosen reaction coordinates (RC) were: the shortening of the distance between the attacking water oxygen and the carbonyl carbon (RC_1) and the enlarging of the bond between the Ser225(TPA) side chain oxygen and the carbonyl carbon of TPA (RC_2). Link atoms defined as *dummy* hydrogens were added after breaking covalent bonds in the limit between QM and MM regions. All QM/MM calculations were performed using Self-Consistent-charge Density-Functional Tight-Binding of third order (DFTB3).[55] The pulling work (W, kcal mol^{-1}) profiles of the deacylation step of Ser225(TPA) were

obtained via sMD. The harmonic constant applied was 600 kcal mol^{-1} Å^{-2} for both reaction coordinates in order to ensure the completion of the reaction. 2 ps simulations with a time step of 0.2 fs were carried out at constant temperature of 300 K and pressure of 1 bar using Langevin thermostat. The mean PMF (kcal mol^{-1}) value was calculated from the pulling work using the Jarzynski identity (Eq. 1):[23,24]

$$e^{-\beta_T \Delta G} = <e^{-\beta_T W_i}> \tag{1}$$

where $\beta_T = \frac{1}{k_B T}$.

Machine Learning methods

Kernel regression methods

We employed three regression methods: kernel ridge regression (KRR), support vector regression (SVR) and ElasticNet. In the case of KRR and SVR, the input data are transformed into a high-dimensional space and then the relation between the input features and the output is learned.[29,56] In all cases, we used the method implementation in *scikit-learn*.[57] All methods use linear regression to describe the relationship between the input data and the response. The fitting of the model to the data is done via the minimization of a loss function $\mathcal{L}$ of the form:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}_i) \right)^2, \tag{2}$$

where y_i is the true value of the response of sample i and $f(\mathbf{x}_i)$ is the one predicted by the model based on the feature vector $\mathbf{x}_i$. n is the total number of samples.

In the case of ridge regression, an additional $L2$ regularization term is included in the loss:

$$\mathcal{L}_{\mathrm{RR}} = \mathcal{L} + \lambda_2 \sum_j \omega_j^2, \tag{3}$$

where λ_2 is the L_2 regularization parameter and ω are the regression coefficients for the input

features.

The loss function used in ElasticNet regression introduces a L_1 penalty term in addition to L_2 regularization:

$$\mathcal{L}_{\text{EN}} = \mathcal{L} + \lambda_2 \sum_j \omega_j^2 + \lambda_1 \sum_j |\omega_j|, \tag{4}$$

with λ_1 as the corresponding regularization parameter.

On the other hand in SVR - more specifically in ϵ-SVR - two loss functions are used: $L2$ (Eq. 3) and the ϵ-*insensitive* loss function (Eq. 5 and 6):

$$\min \frac{1}{2}||\omega||^2 + C \sum_{i=1}^{n} (\xi_+ + \xi_-) \tag{5}$$

$$\text{subject to:} \begin{cases} y_i - (\omega x_i + b) \leq \epsilon + \xi_+ \\ (\omega x_i + b) - y_i \leq \epsilon + \xi_- \\ \xi_+, \xi_- \geq 0 \end{cases} \tag{6}$$

where C is the regularization parameter, and ξ_+ and ξ_- are the deviation to both sides of the ϵ-zone (they define the threshold values for fitting the error).

For mapping the data into a high-dimensional space, we used the Gaussian kernel $k_{Gaussian}(x, x')$, also known as *radial basis function* (RBF) :

$$k_{Gaussian}(x_i, x_j) = \exp\left(-\frac{d(x_i, x_j)^2}{2\sigma^2}\right) \tag{7}$$

where σ is the length-scale of the kernel and $d(x_i, x_j)$ refers to the euclidean distance.

We employed a grid search with 3-fold cross-validation to select the hyperparameters. The training data is randomly shuffled and split into 3 groups. One group is left out and the grid search is performed in the remaining groups. The combinations of the hyperparameters are used for training the first two groups and then evaluated on the left-out group. This process is repeated 3 times until each group is used one time as a validation set and 2 times

as a training set. After completing the grid search the hyperparameters with the best score are selected for fitting the model. To evaluate the goodness of the predictions with the different algorithms, we checked the following parameters: (i) score value of the regression - the interpretation of this value is similar to the linear regression R^2 value; (ii) mean absolute error (MAE) of the prediction; and (iii) root mean square error (RMSE). The Coulomb matrix representation was coded in python by us. The atom-centered symmetry functions (ACSF)[42] and the smooth overlap of atomic positions (SOAP),[38] were used as defined in *DScribe*.[58]

Partial least square (PLS) regression

Partial least squares (PLS) is a parametric method, where it is assumed that:

$$f(\mathbf{x_i}) = \mathbf{x_i}\beta \tag{8}$$

and the aim is to estimate the unknown parameter β. PLS has a long tradition in chemometrics and has been introduced first by Wold as an algorithm.[59] A formal definition of a PLS estimator of order d is given by:

$$\hat{\beta}_d = \arg \min_{\beta \in \mathcal{K}_d(X^TY, X^TX)} \|Y - X\beta\|^2 \tag{9}$$

where $Y = (y_1, \ldots, y_n)^T$, $X = (x_1, \ldots, x_n)^T$ and $\mathcal{K}_d(X^TY, X^TX)$ denote a Krylov space of order d, that is, $\mathcal{K}_d(X^TY, X^TX) = \text{span}\{X^TY, X^TXX^TY, \ldots, (X^TX)^{d-1}X^TY\}$. The order of the PLS estimator d is a regularisation parameter and can be chosen by cross-validation. PLS estimator is known to deliver a robust and parsimonious model that can be readily interpreted in chemical systems.[60]

Results and discussion

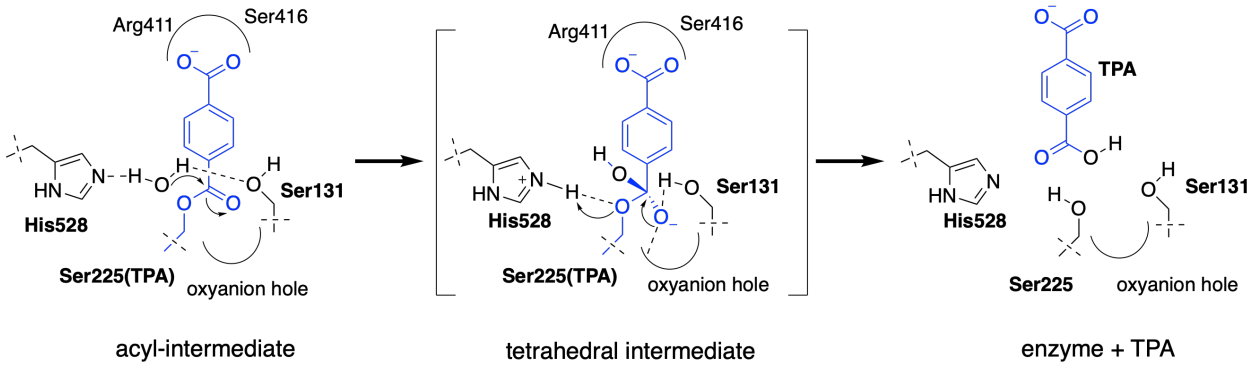
System of study

We selected as system of study the second step of the reaction of hydrolysis of MHET by the plastic-degrading *I. sakaeinsis* MHETase, where ethylen glycol and terepthalic acid (TPA) are released after a water molecule attacks the carbonyl carbon on the acylated Ser225(TPA) (Scheme 1). Several facts support this decision: (i) the reaction mechanism of the enzyme has been extensively investigated and it is well-established,[43,61] (ii) there are sufficient available experimental data for the catalytic activity of wild-type and single-point MHETase variants with MHET as substrate,[43,62,63] (iii) MHET does not imply problems with the crystalline degree of the initial plastic substrate; (iv) there are currently available some experimental structures of the enzyme in complex with small molecules like the substrate analog MHETA (PDB id. 6QGA),[62] and of course, (v) because of the scientific and industrial relevance of PETases and related enzymes.[64,65]

Based on the general activity of Ser hydrolases, the whole reaction mechanism of hydrolysis of MHET would consists of two main steps involving the acylation and deacylation of the catalytic Ser225. In a first step, the side chain of Ser225 is activated via proton abstraction by His528 and the resulting alkoxyde attacks to the ester carbon of MHET to yield the acyl-enzyme intermediate showed in Scheme 1. Then, this intermediate reacts in a second step with a water molecule to regenerate the side chain in Ser225, with the release of TPA as product of the reaction (deacylation step). According to previous computational results of McGeehan and co-workers,[43] the latter step of the reaction mechanism (resolution of the acyl-intermediate) is the rate-limiting event with an activation energy barrier of 19.8 kcal mol,$^{-1}$[43] a value ca. 5.9 kcal mol^{-1} larger than the one for the energy barrier of the first part of the reaction. Additionally, no tetrahedral intermediate were found for any of the two global reaction steps.

In Figure 1A we show the active site of the native MHETase after acylation of Ser225

by MHET with the TPA moiety (named as Ser225(TPA)). We represent as sticks the most relevant residues, the catalytic triad, and those involved in the hydrogen bond network keeping the architecture of the active site. As described previously, we observed that the carboxylate group of Ser225(TPA) stably interacts with both side chains of Arg411 and Ser416 along the MD simulation of the native enzyme. Importantly, the guanidinium group of Arg411 keeps the architecture of the active site by interacting with the substrate and also bridging the loop before α-helix H21 (via hydrogen bond with Ala494) with the helix H15, where Ser416 is located (PDBsum notation). Complementary, the carbonyl oxygen of the TPA moiety in the acylated Ser225(TPA) interacts with the oxyanion hole defined by Gly132 and Glu226. The aromatic ring on Ser225(TPA) is almost parallel to the planes defined by the side chains of Trp397 and Phe495. The nucleophile of the reaction, the attacking water molecule, is positioned by the side chain of His528, and occasionally, by the side chain of Ser131 above the carbonyl carbon on Ser225(TPA), the electrophile of the reaction. Indeed, we found out that the side chain of Ser131 interacts 29 % of the simulation time with the attacking water molecule. We also identified in our simulations that Phe415 populates mostly the open conformation of the dual occupancy found in the available experimental structures.[62] Finally, the distal residue Ser416, located on the α-helix H15, establishes a side chain-side chain interaction with Asp423 located one helix turn upstream.



Scheme 1: Accepted reaction mechanism for the second step of the hydrolysis of MHET catalyzed by *Is*-MHETase.

Conformational screening and selection of snapshots for simulation of enzyme reactivity

In order to simulate the reaction mechanism of the enzyme, we needed to define first which initial snapshots from the MD simulations would be used for running the sMD simulations. In principle, a reasonable approach would be the selection of those conformations that are closed to the Michaelis complex (aka. near-attack conformations). That is, conformations where the distances between nucleophile and electrophile are optimal for the reaction. On the one hand, the electrophile of the reaction (the carbonyl carbon of Ser225(TPA)) is tightly positioned and activated by its interaction with the amino groups of Glu226 and Gly132. On the other hand, the attacking water molecule is positioned and activated by the ϵ-nitrogen of His528. As we pointed out before, the attacking water molecule occasionally interacts with the side chain of Ser131. Thus, we defined three distances (Figure 1A) and we analyzed their distribution (Figure 1B) along the MD simulation: d_1, the distance between the attacking water oxygen and the carbonyl carbon of Ser225(TPA); d_2, the distance between the ϵ-nitrogen of His528 and the water oxygen; and d_3, the distance between the latter atom and the alcohol oxygen of Ser131. Indeed, the variant Ser131Gly, which has no possibility of interaction via its side chain with the attacking water molecule, has been shown to retain ca. one third of the experimental catalytic activity of the native variant.[43] In our MD analysis, we found out that the water molecule is well activated by His528, with a d_2 value close to 3.0 Å and that is also well positioned to attack the carbonyl center ($d_1 \sim 3.5$ Å). However, we found a broad distribution for d_3, what may indicate a less stable interaction pattern between these atoms, and maybe, an assistant but not a key role of Ser131 in the MHETase catalysis. Altogether, we considered d_1 and d_2 as inspiration for the definition of the reaction coordinate to study the enzymatic reaction and we used the three distances (also d_3) for the selection of the initial MD snapshots in these studies.

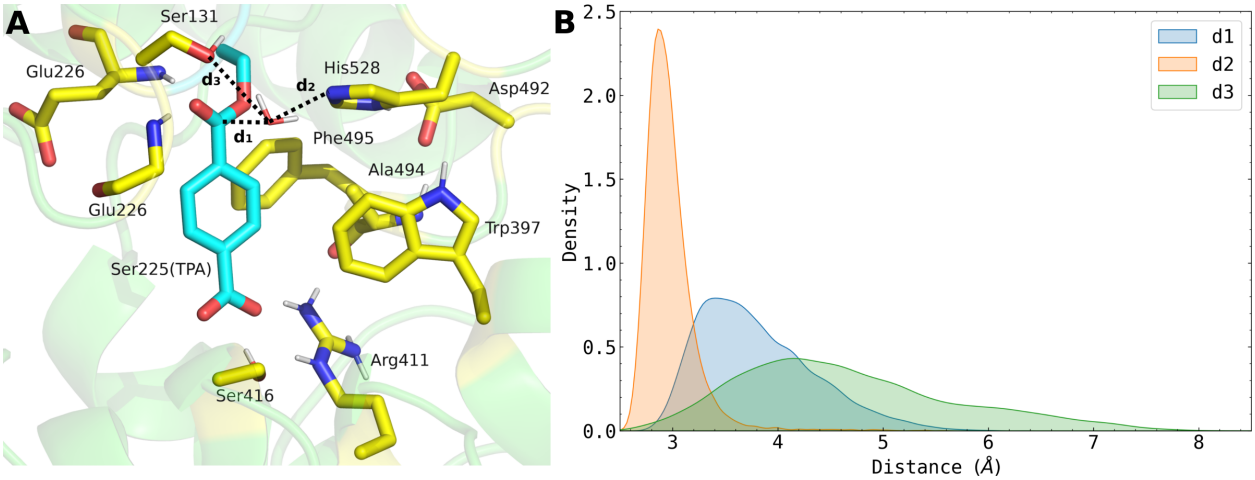


Figure 1: (**A**) Detail of the active site of *Is*-MHETase with the acylated Ser225(TPA) (C-atoms colored in blue). Relevant residues are highlighted as sticks (C-atoms in yellow). Distances $d_1 - d_3$ are shown as dotted lines. (**B**) Distribution for distances $d_1 - d_3$ (Å) along the MD simulation.

Calculation of energy barriers for the catalyzed reaction

We run sMD simulations for the catalyzed reaction highlighted in Scheme 1 by the wild-type MHETase. We defined two reaction coordinates (RC): the shortening of the distance between the attacking water oxygen and the carbonyl carbon (RC_1), and the enlarging of the bond between the Ser225(TPA) side chain oxygen and the carbonyl carbon of TPA (RC_2). This second coordinate was included to ensure that the reaction is completed. A total of 1,500 sMD trajectories were generated from MD snapshots and the mean value for the energy barrier was computed using the Jarzinsky equality (see Methods) from the maximum values of the pulling work.[23,24] The input MD snapshots were selected randomly within the low-value 60 % of the MD conformer distribution for d_1-d_3 values. This way, selecting these populations we removed some artefacts and fluctuations from the MD simulations but we kept a statistical sampling. After running ca. 240 sMD trajectories (see Figure S1), the energy barrier for the native MHETase enzyme converges. A final value of 18.22 kcal mol^{-1} was obtained, which is in good agreement with the free energy barrier ($\Delta G^{\ddagger}$ = 19.8 kcal mol^{-1}) computed by McGeehan[43] and coworkers using 2D umbrella sampling.

Prediction of pulling work values from sMD snapshots

Having validated that our sMD simulations reproduce the energy values obtained by 2D umbrella sampling, we investigated the performance of different ML methods for predicting the pulling work value obtained for each of the sMD snapshots. An important aspect of ML in chemistry is the definition of the representation of the chemical system. For example, explicitly including all the atoms present in a description of this system can be computationally demanding and, to some extent, unfeasible. Taking into account that MHETase belongs to the Ser hydrolases family, it seems reasonably to select the residues involved in the catalytic process: His528, Asp492, and Ser225(TPA). Hence, we explored three different feature representations: (i) a Coulomb matrix (CM) representation[66] based on the atoms of the QM region, including the side chains of the former residues from the β-carbon and a water molecule (data set 1, 33 atoms, Figure S2A) (ii) a CM representation of the atoms of the side chains of Ser225(TPA) and His528, without the β-carbon atoms, and a water molecule (data set 2, 27 atoms, Figure S2B), and (iii) the set of distances d_1, d_2 and d_3 (data set 3). For the model selection, we studied three regression methods: kernel ridge regression (KRR)[39], the supported vector regression (SVR)[40] and ElasticNet.

As training sets, a total of 100 sMD trajectories (10,000 frames) of the native MHETase were used as input geometries for both methods. The data were split into groups of 2,500 frames, and the analysis was divided into four runs with increasing input data sizes (2,500, 5,000, 7,500, and 10,000 frames). In all cases, 25 % and 75 % of the data were selected as validation and training sets, respectively. With respect to the different representations, the prediction of the pulling work for the input data using the CM representation (data sets 1 and 2) presents a very high score factor ($\sim$ 0.90) even for the smaller number of input points (2,500, see Table 1). In contrast, with data set 3, where distances d_1-d_3 are given as descriptors, the scoring value slightly decreases while larger mean absolute error (MAE) and root-mean square error (RMSE) values are obtained. In case of the regression methods, KRR and ElasticNet tend to show smaller MAE and RMSE values (ca. 2.5 kcal mol^{-1}) than

those observed for SVR (ca. 13.0 kcal mol^{-1}). Furthermore, only an slight decrease in the score value and a soft increase in the MAE values are observed when moving from data set 1 to data set 3.

Table 1: Training process to predict sMD-derived pulling work (kcal mol^{-1}) using KRR and SVR.

Size	**Data set**	**Kernel method**	**Score**	**MAE**	**RMSE**
		KRR	0.92	2.47	3.22
	1	SVR	0.89	13.29	16.35
		ElasticNet	0.92	2.46	3.23
		KRR	0.91	2.76	4.05
2,500	2	SVR	0.89	14.15	17.54
		ElasticNet	0.91	2.72	3.95
		KRR	0.86	2.94	4.36
	3	SVR	0.87	14.13	17.52
		ElasticNet	0.84	3.86	5.19
		KRR	0.92	2.46	3.33
	1	SVR	0.90	12.82	15.79
		ElastiNet	0.92	2.46	3.32
		KRR	0.92	2.51	3.51
5,000	2	SVR	0.92	13.89	17.16
		ElasticNet	0.92	2.53	3.52
		KRR	0.91	2.72	3.81
	3	SVR	0.90	13.94	17.23
		ElasticNet	0.86	3.56	4.60
		KRR	0.91	2.55	3.57
	1	SVR	0.89	12.94	15.97
		ElasticNet	0.91	2.53	3.56
		KRR	0.91	2.66	3.74
7,500	2	SVR	0.90	13.68	16.89
		ElasticNet	0.91	2.68	3.76
		KRR	0.89	2.98	4.11
	3	SVR	0.88	13.54	16.71
		ElasticNet	0.84	3.78	4.82
		KRR	0.90	2.53	3.54
	1	SVR	0.89	12.68	15.65
		Elasticnet	0.90	2.52	3.51
		KRR	0.91	2.56	3.58
10,000	2	SVR	0.91	13.43	16.58
		ElasticNet	0.91	2.61	3.63
		KRR	0.89	2.90	4.02
	3	SVR	0.88	13.35	16.49
		ElasticNet	0.85	3.61	4.66

It is worth noting that during the model selection phase, no improvement in the scores are achieved with the increment of the size of the training data. Therefore, we selected the models obtained with KRR and ElasticNet using 2,500 input data for further analysis. In Table 2 are included the results for the prediction of the sMD pulling work values along the sMD trajectory from a test set (40,000 points). The test data was divided into 4 test subsets of 10,000 structures each one (A-D, Table 2). Both MAE and RMSE values slightly increase with respect to the values obtained in the training phase (Table 2). However, both algorithms are able to predict well the sMD trajectory pulling work values. Nevertheless, although this error is equal or smaller than the oscillations observed during the sMD simulations, the MAE is still above chemical accuracy (> 1 kcal mol^{-1}).

Table 2: Prediction of sMD pulling work (kcal mol^{-1}) for unseen data.

ML method	**Score**	**MAE**	**RMSE**
	0.86	3.15	4.42
	0.89	2.93	3.94
KRR	0.88	3.00	4.11
	0.88	3.02	4.11
	0.87	3.07	4.32
	0.90	2.84	3.83
ElasticNet	0.88	2.94	4.02
	0.89	2.95	4.03

In Figure 2 we illustrate this prediction of sMD pulling work values considering 10 sMD trajectories (10,000 frames, KRR). The model predicts well values between 5 and 35 kcal mol^{-1} but underestimates those values out of this range, as can be seen in the superposition of predicted vs reference values (Figure 2A). Similar behavior was found for ElasticNet (see Figure S3). Analyzing the profile of several sMD trajectories, we found out some of these upper extreme values correspond to unrealistic situations where the pulling work increases after crossing the maximum that corresponds to the transition state of the reaction (see left panel in Figure 2A). In addition to that, the initial geometries of the active site are problematic for the regression method, since geometrically diverse structures need to be mapped to very similar pulling work values (close to zero). Thus, we removed the initial

and final snapshots out of 100 snapshots per trajectory (leaving frames from 10 to 90) and we repeated the training process (Figure 2B). The reduced data set leads to a very limited improvement of the relevant regions of the trajectory (see Table S1). These results lead us to conclude that obtaining better scoring values and lower errors is limited by the inherit variability of the sMD method (see Figure S4).

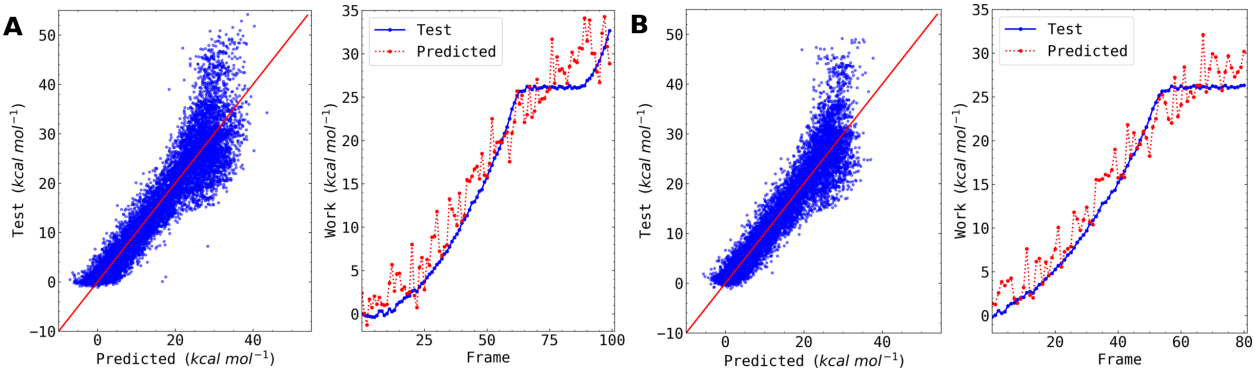


Figure 2: Reference vs predicted values for the entire trajectory using KRR.

Overall, we found that a combination of KRR and ElasticNet together with a Coulomb matrix representation of the active site of the enzyme, is able to predict the pulling work values along sMD trajectories with an error close to 3 kcal mol^{-1}. A problem of this approach is its reliance on geometries from sMD trajectories as an input to the ML models: in order to generate these structures, QM/MM simulations would need to be carried out before every prediction. The associated computational effort makes the above ML approach too expensive in many practical scenarios, especially in studies where several enzyme variants are explored. One cheapest scenario in terms of computational resources would be the use of classical MD snapshots as input data.

Prediction of energy barriers from MD simulation snapshots

Prompted with the idea of reducing the computational cost of our approach, we decided to assess if we could predict the maximum of the sMD pulling work not from sMD snapshots (QM/MM trajectories) but from classical MD simulation snapshots. A closest example is the

pioneering work published by Ochsenfeld and co-workers in 2019 for the detection of reactive conformations from MD simulations.[67] By means of ElasticNet regression,[41] the authors trained 100 initial points using 15 input geometric features. The initial points were obtained via QM/MM adiabatic mapping of the reaction mechanism. As outcome, the authors got a humble regression coefficient of 0.28 for the prediction of the energy barriers of the reaction from MD snapshots. Nevertheless, they were able to predict energy barriers with a MAE of 3.6 kcal mol^{-1} along the MD simulation. In our case, we took the initial MD snapshot prior sMD (0^{th} snapshot) together with the maximum pulling work value obtained for each of the initial MD snapshots. The selection of the maximum pulling work value was guided by the averaged energy computed via Jarzynski equality. We represented the chemical system as before (with a CM) and we used KRR and ElasticNet as regression methods (Tables 3 and Table S2). The analysis was divided in 15 runs with increasing size of the input data. The score coefficient is remarkably low in all cases and the prediction of the trained model is pretty poor, with energies after Jarzynski averaging overestimated in more than 6.00 kcal mol^{-1} for both methods (Table 4).

Table 3: Training process to predict maximum work values (kcal mol^{-1}) from MD snapshots.

ML method	**Size**	**Score**	**MAE**	**RMSE**
KRR	800	0.02	3.11	4.11
ElasticNet	800	0.06	3.02	4.02

Table 4: Prediction of the energy barrier (kcal mol^{-1}) from MD snapshots.

ML method	**MAE**	**RMSE**	**Score**	$\Delta G^{\ddagger}_{pred}$	$\Delta\Delta G$
KRR	3.11	15.71	0.08	4.94	6.21
ElasticNet	3.27	17.38	0.04	5.13	8.09

In order to elucidate if the chemical representation may be the reason of the poor predictive power of the sMD maximum by of our model, we decided to explore additional representations (Table 5). We selected atom-centered symmetry functions (ACSF)[42] and smooth overlap of atomic positions (SOAP).[38] ACSF represents the local environment of an atom

by means of two- and three-body symmetry functions (i.e., radial and angular functions). SOAP is computed via local expansion of Gaussian atomic densities (based on spherical and radial basis functions) and a smooth three-body correlation function. The outcome of our calculations are summarised on Table 5. Overall, there are no differences between all representations and no improvement is observed when compared to CM. Indeed, large errors are obtained when predicting the energy barrier of the reaction $\Delta G^{\ddagger}_{pred}$ (Figure 3).

Table 5: Training process to predict maximum work values (kcal mol^{-1}) from MD snapshots using different representations

ML method	Size	Representation	Score	MAE	RMSE
KRR	900	ACSF	0.11	2.90	3.87
	900	SOAP	0.10	2.91	3.90
ElasticNet	900	ACSF	0.09	2.94	3.92
	900	SOAP	0.10	2.90	3.90

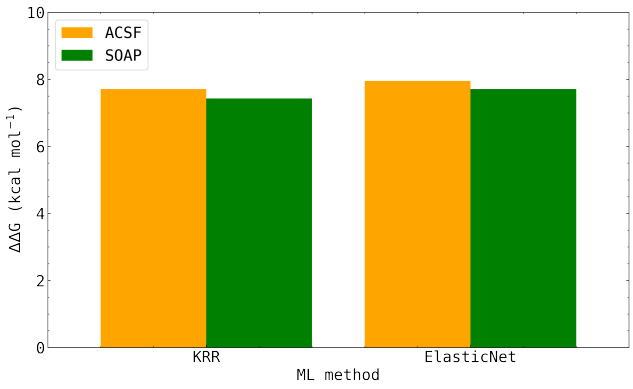


Figure 3: Error (kcal mol-1) obtained for the prediction of the energy barrier using a single MD snapshot as input data.

Having demonstrated that no improvement was achieved by exploring other chemical representations, we decided to check if we could still predict the pulling work maximum using a single snapshot but, in this case, from the sMD trajectory. Thus, we selected the 30^{th}, 40^{th}, 50^{th}, and 60^{th} frames (Table 6). Importantly, all these structures are located before the pulling work maximum. First, we observe that with the increase of the simulation time (from 30^{th} to 60^{th} frame) the fitting improves (Table 6). Second, for the 60^{th} frame, all score values

are above 0.30 and the MAEs below 3 kcal mol^{-1}. In this case, the SOAP representation yields the best accuracy (0.43 and 0.41 for ElasticNet and KRR, respectively) and lower error in the prediction of the energy barrier (Figure 4 and Table S3). Both observations suggest, that the structure of the initial MD snapshot (0^{th} frame) is too dissimilar from the maximum to encode sufficient information on how the reaction will proceed and which barrier height will be reached. Closer to the transition state, the snapshots recover this information, making it possible for the ML algorithm to learn a better mapping.

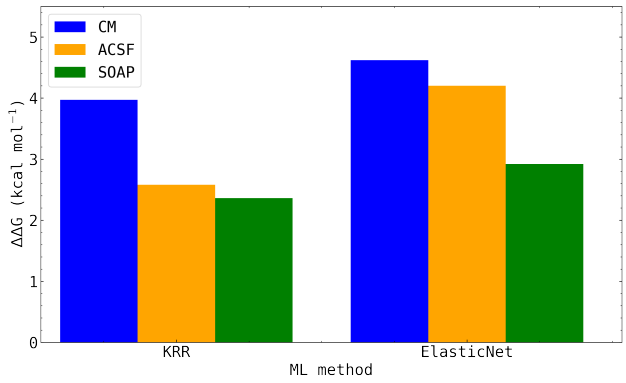


Figure 4: Difference of the energy calculated using the predicted pulling work with the energy obtained from the sMD calculation.

Partial least square (PLS) regression

Finally, in parallel to the non-parametric methods used so far, we used PLS for last predictions (Table S4). The score and error values obtained by PLS are similar to the ones obtained with KRR and ElasticNet. However, it must be stressed, that in the case of PLS, we could extract additional chemically-intuitive results. In particular, when using a CM matrix representations, the obtained β coefficients from the PLS analysis could be ascribed to distances between pair of atoms in the input structures (Table S5 and Figure S5). Whereas for the data set derived from the 0^{th} snapshots, no chemically understandable pair-interactions were observed, directly and/or indirectly related to the reaction, the analysis of the data set de-

rived from the 60^{th} frames shows some relevant contribution to the predicted pulling work values on the top of the list. As example, the two first β coefficients (6.94 and 6.61, respectively) refer to distances between the ester carbon of MHET and the two water hydrogen atoms. Nevertheless, the fact that significant features defined in the collective variables are not top listed within the β coefficients may also explain the low scoring value for the 60^{th} frames.

Table 6: Training process to predict maximum work values (kcal mol^{-1}) with one sMD snapshot.

		KRR			ElasticNet		
	Frame	**Score**	**MAE**	**RMSE**	**Score**	**MAE**	**RMSE**
CM	30^{th}	0.08	3.08	3.97	0.09	3.07	3.96
	40^{th}	0.04	3.18	4.05	0.02	3.21	4.10
	50^{th}	0.22	2.93	3.65	0.22	2.92	3.66
	60^{th}	0.38	2.58	3.25	0.34	2.64	3.36
ACSF	30^{th}	0.14	3.00	3.84	0.14	3.02	3.84
	40^{th}	0.11	3.05	3.90	0.11	3.04	3.90
	50^{th}	0.34	2.68	3.36	0.29	2.77	3.49
	60^{th}	0.42	2.46	3.16	0.36	2.61	3.32
SOAP	30^{th}	0.15	2.98	3.81	0.16	2.96	3.80
	40^{th}	0.15	2.96	3.82	0.14	2.98	3.84
	50^{th}	0.35	2.68	3.35	0.35	2.69	3.34
	60^{th}	0.41	2.49	3.17	0.43	2.46	3.14

Conclusions

Here we have explored the use of sMD libraries and ML algorithms for the prediction of enzyme-catalyzed energy barriers. As showcase, we have studied the second step of the hydrolysis of MHET by the plastic-degrading *I. sakaeinsis*-MHETase. We have used KRR, SVR and ElasticNet together with different chemical representations of the active site in the presence of the substrate, to predict sMD pulling work maxima for this reaction step. Our results indicate that KRR and ElasticNet are suitable algorithms for the prediction of pulling

work values when using sMD trajectories as training data and the atoms of the active site of the enzyme. However, more challenging is the prediction of the sMD pulling work maximum using a single snapshot per trajectory as training data. Whereas the use of a MD snapshot (0^{th} snapshot) - yields very poor predictions in our case, the use of a closer sMD snapshot to the reaction energy maximum (60^{th} snapshot), significantly improves the statistics of the prediction (MAE under 3 kcal mol^{-1}). To rationalize this finding, we run PLS analysis. We found out that the MD snapshot does not encode sufficient information to predict the pulling work values of our data set, as shown by the absence of relevant β coefficients for atomic pair-interactions related to the reaction. With this example we could confirm that, parametric methods like PLS can deliver chemically-interpetrable models, something that is sometimes challenging to obtain with non-parametric methods. Overall, we show in this work how the prediction of enzyme-based reaction energy barriers using ML and chemical-dynamics-derived all-atom models as training data is still state-of-art. Further investigation are in progress.

Acknowledgement

D.P.-R. thanks Otto-Loewi Research Center (Medical University of Graz) for funding support. MG works at the BASLEARN – TU Berlin/BASF Joint Lab for Machine Learning, co-financed by TU Berlin and BASF SE. The authors thank the Medical University of Graz (cluster MedBioNode) and the Spanish Supercomputing Network (RES) (project id. BCV-2022-3-0003 on Turgalium) for computation time.

References

(1) Kirk, O.; Borchert, T. V.; Fuglsang, C. C. Industrial enzyme applications. *Current Opinion in Biotechnology* **2002**, *13*, 345–351.

(2) Sheldon, R. A.; Woodley, J. M. Role of Biocatalysis in Sustainable Chemistry. *Chemical Reviews* **2018**, *118*, 801–838.

(3) Moroz, Y. S.; Dunston, T. T.; Makhlynets, O. V.; Moroz, O. V.; Wu, Y.; Yoon, J. H.; Olsen, A. B.; McLaughlin, J. M.; Mack, K. L.; Gosavi, P. M.; van Nuland, N. A. J.; Korendovych, I. V. New Tricks for Old Proteins: Single Mutations in a Nonenzymatic Protein Give Rise to Various Enzymatic Activities. *Journal of the American Chemical Society* **2015**, *137*, 14905–14911.

(4) Choi, J.-M.; Han, S.-S.; Kim, H.-S. Industrial applications of enzyme biocatalysis: Current status and future aspects. *Biotechnology Advances* **2015**, *33*, 1443–1454.

(5) Rosano, G. L.; Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology* **2014**, *5*.

(6) Modarres, H. P.; Mofrad, M. R.; Sanati-Nezhad, A. Protein thermostability engineering. *RSC Adv.* **2016**, *6*, 115252–115270.

(7) Huang, P.; Chu, S. K. S.; Frizzo, H. N.; Connolly, M. P.; Caster, R. W.; Siegel, J. B. Evaluating Protein Engineering Thermostability Prediction Tools Using an Independently Generated Dataset. *ACS Omega* **2020**, *5*, 6487–6493.

(8) Rigoldi, F.; Donini, S.; Redaelli, A.; Parisini, E.; Gautieri, A. Review: Engineering of thermostable enzymes for industrial applications. *APL Bioengineering* **2018**, *2*.

(9) St-Jacques, A. D.; Eyahpaise, M.- C.; Chica, R. A. Computational Design of Multisubstrate Enzyme Specificity. *ACS Catalysis* **2019**, *9*, 5480–5485.

(10) Brustad, E. M.; Arnold, F. H. Optimizing non-natural protein function with directed evolution. *Current Opinion in Chemical Biology* **2011**, *15*, 201–210, Biocatalysis and Biotransformation/Bioinorganic Chemistry.

(11) Hammer, S. C.; Knight, A. M.; Arnold, F. H. Design and evolution of enzymes for non-natural chemistry. *Current Opinion in Green and Sustainable Chemistry* **2017**, *7*, 23–30, New Synthetic Methods 2017.

(12) Khersonsky, O.; Lipsh, R.; Avizemer, Z.; Ashani, Y.; Goldsmith, M.; Leader, H.; Dym, O.; Rogotner, S.; Trudeau, D. L.; Prilusky, J.; Amengual-Rigo, P.; Guallar, V.; Tawfik, D. S.; Fleishman, S. J. Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Molecular Cell* **2018**, *72*, 178–186.e5.

(13) Goldenzweig, A. et al. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Molecular Cell* **2016**, *63*, 337–346.

(14) Sumbalova, L.; Stourac, J.; Martinek, T.; Bednar, D.; Damborsky, J. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Research* **2018**, *46*, W356–W362.

(15) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations. *Science* **2004**, *303*, 186–195.

(16) Vaissier Welborn, V.; Head-Gordon, T. Computational Design of Synthetic Enzymes. *Chemical Reviews* **2019**, *119*, 6613–6630.

(17) Himo, F. Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *Journal of the American Chemical Society* **2017**, *139*, 6780–6786.

(18) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(19) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187–199.

(20) Pérez de Alba Ortíz, A.; Vreede, J.; Ensing, B. In *Biomolecular Simulations: Methods and Protocols*; Bonomi, M., Camilloni, C., Eds.; Springer New York: New York, NY, 2019; pp 255–290.

(21) Warshel, A.; Weiss, R. M. An empirical valence bond approach for comparing reactions in solutions and in enzymes. *Journal of the American Chemical Society* **1980**, *102*, 6218–6226.

(22) Schowen, R. L. In *Transition States of Biochemical Processes*; Gandour, R. D., Schowen, R. L., Eds.; Springer US: Boston, MA, 1978; pp 77–114.

(23) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *The Journal of Chemical Physics* **2003**, *119*, 3559–3566.

(24) Park, S.; Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *The Journal of Chemical Physics* **2004**, *120*, 5946–5961.

(25) Torres-Salas, P. et al. Engineering Erg10 Thiolase from Saccharomyces cerevisiae as a Synthetic Toolkit for the Production of Branched-Chain Alcohols. *Biochemistry* **2018**, *57*, 1338–1348.

(26) del Arco, J.; Perona, A.; González, L.; Fernández-Lucas, J.; Gago, F.; Sánchez-Murcia, P. A. Reaction mechanism of nucleoside 2-deoxyribosyltransferases: free-energy landscape supports an oxocarbenium ion as the reaction intermediate. *Org. Biomol. Chem.* **2019**, *17*, 7891–7899.

(27) Gavin, D. P.; Reen, F. J.; Rocha-Martin, J.; Abreu-Castilla, I.; Woods, D. F.; Foley, A. M.; Sánchez-Murcia, P.; Schwarz, M.; O'Neill, M. A. R., P; F, O. Genome mining and characterisation of a novel transaminase with remote stereoselectivity. *Scientific reports* **2019**, *9*, 1–15.

(28) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; MÃ¼ller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chemical Reviews* **2021**, *121*, 9816–9872.

(29) Lewis-Atwell, T.; Townsend, P. A.; Grayson, M. N. Machine learning activation energies of chemical reactions. *WIREs Computational Molecular Science* **2022**, *12*, e1593.

(30) Lemm, D.; von Rudorff, G. F.; von Lilienfeld, O. A. Machine learning based energy-free structure predictions of molecules, transition states, and solids. *Nature Communications* **2021**, *12*, 4468.

(31) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best practices in machine learning for chemistry. *Nature Chemistry* **2021**, *13*, 505–508.

(32) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catalysis* **2020**, *10*, 1210–1223.

(33) Watanabe, N.; Murata, M.; Ogawa, T.; Vavricka, C. J.; Kondo, A.; Ogino, C.; Araki, M. Exploration and Evaluation of Machine Learning-Based Models for Predicting Enzymatic Reactions. *Journal of Chemical Information and Modeling* **2020**, *60*, 1833–1843.

(34) Bonk, B. M.; Weis, J. W.; Tidor, B. Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. *Journal of the American Chemical Society* **2019**, *141*, 4108–4118.

(35) Lewis-Atwell, T.; Townsend, P. A.; Grayson, M. N. Machine learning activation energies of chemical reactions. *WIREs Computational Molecular Science* **2022**, *12*, e1593.

(36) Schomburg, I.; Jeske, L.; Ulbrich, M.; Placzek, S.; Chang, A.; Schomburg, D. The BRENDA enzyme information system–From a database to an expert system. *Journal of Biotechnology* **2017**, *261*, 194–206, Bioinformatics Solutions for Big Data Analysis in Life Sciences presented by the German Network for Bioinformatics Infrastructure.

(37) Wittig, U.; Rey, M.; Weidemann, A.; Kania, R.; Müller, W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Research* **2017**, *46*, D656–D660.

(38) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(39) Cristianini, N.; Shawe-Taylor, J., et al. *An introduction to support vector machines and other kernel-based learning methods*; Cambridge university press, 2000.

(40) Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. *Advances in neural information processing systems* **1996**, *9*.

(41) Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **2005**, *67*, 301–320.

(42) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics* **2011**, *134*, 074106.

(43) Knott, B. C.; Erickson, E.; Allen, M. D.; Gado, J. E.; Graham, R.; Kearns, F. L.; Pardo, I.; Topuzlu, E.; Anderson, J. J.; Austin, H. P., et al. Characterization and engineering of a two-enzyme system for plastics depolymerization. *Proceedings of the National Academy of Sciences* **2020**, *117*, 25476–25485.

(44) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research* **2012**, *40*, W537–W541.

(45) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation* **2020**, *16*, 528–552.

(46) Becke, A. Density-functional thermochemistry. III The role of exact exchange J Chem Phys 98: 5648–5652. 1993.

(47) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

(48) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics* **1980**, *58*, 1200–1211.

(49) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *The Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

(50) Frisch, M. J. et al. Gaussian16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(51) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(52) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **1995**, *103*, 8577–8593.

(53) Case, D.; Belfon, K.; Ben-Shalom, I.; Brozell, S.; Cerutti, D.; Cheatham, T.; Cruzeiro, V.; Darden, T.; Duke, R.; Giambasu, G., et al. AMBER 2020: University of California. *San Francisco* **2020**,

(54) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation* **2013**, *9*, 3084–3095.

(55) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *Journal of Chemical Theory and Computation* **2011**, *7*, 931–948.

(56) Rupp, M. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry* **2015**, *115*, 1058–1073.

(57) Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; VanderPlas, J.; Joly, A.; Holt, B.; Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning. 2013; pp 108–122.

(58) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **2020**, *247*, 106949.

(59) Wold, H. *Nonlinear Estimation by Iterative Least Square Procedures*; 1968.

(60) Krivobokova, T.; Briones, R.; Hub, J. S.; Munk, A.; de Groot, B. L. Partial Least-Squares Functional Mode Analysis: Application to the Membrane Proteins AQP1, Aqy1, and CLC-ec1. **2012**,

(61) Berselli, A.; Ramos, M. J.; Menziani, M. C. Novel Pet-Degrading Enzymes: Structure-Function from a Computational Perspective. *ChemBioChem* **2021**, *22*, 2032–2050.

(62) Palm, G. J.; Reisky, L.; Böttcher, D.; Müller, H.; Michels, E. A.; Walczak, M. C.; Berndt, L.; Weiss, M. S.; Bornscheuer, U. T.; Weber, G. Structure of the plastic-degrading Ideonella sakaiensis MHETase bound to a substrate. *Nature Communications* **2019**, *10*, 1–10.

(63) Sagong, H.-Y.; Seo, H.; Kim, T.; Son, H. F.; Joo, S.; Lee, S. H.; Kim, S.; Woo, J.-S.; Hwang, S. Y.; Kim, K.-J. Decomposition of the PET Film by MHETase Using Exo-PETase Function. *ACS Catalysis* **2020**, *10*, 4805–4812.

(64) Kawai, F.; Kawabata, T.; Oda, M. Current knowledge on enzymatic PET degradation and its possible application to waste stream management and other fields. *Applied Microbiology and Biotechnology* **2019**, *103*, 4253–4268.

(65) Lu, H.; Diaz, D. J.; Czarnecki, N. J.; Zhu, C.; Kim, W.; Shroff, R.; Acosta, D. J.; Alexander, B. R.; Cole, H. O.; Zhang, Y.; Lynd, N. A.; Ellington, A. D.; Alper, H. S. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **2022**, *604*, 662–667.

(66) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(67) von der Esch, B.; Dietschreit, J. C. B.; Peters, L. D. M.; Ochsenfeld, C. Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5. *Journal of Chemical Theory and Computation* **2019**, *15*, 6660–6667.