Katharina Leitner, BSc

Point cloud-based analysis of the effects of amino acid variations on enzyme activity

MASTER'S THESIS

to achieve the university degree of

Master of Science

Master's degree programme:
Biochemistry and Molecular Biomedicine

submitted to

Graz University of Technology

Supervisor

Karl Gruber, Univ.-Prof. Mag. Dr.rer.nat.
Institute of Molecular Biosciences, University of Graz

Graz, September 2024

Abstract

In this thesis, an attempt is made to link computationally determined active site properties of enzymes to their respective activities. Measured activity data of two different enzymes and variants thereof are used to learn about the relationship between these and the properties of their active site cavities.
These enzymes are, on the one hand, a decarboxylase (Phenolic Acid Decarboxylase, PAD) and some of its single-, double-, and triple-variants, and on the other hand, a tautomerase (4-Oxalocrotonate Tautomerase) and nearly all of its possible single variants. The decarboxylase activity data contained measurements of the decarboxylation of three different substrates, while the tautomerase data consisted of tautomerization data, but also of activity data of a promiscuous Michael addition using two different Michael donors. Additionally, since one of the Michael additions produces a chiral product, stereoselectivity data on this reaction is also investigated.
Using various machine learning approaches, such as clustering, dimensionality reduction, and regression, the goal was to predict activities of unknown variants of these two enzymes using only computationally determined active site properties. Despite the different approaches that were evaluated, no reasonable predictions could be made on enzyme activity with the workflow applied here. It seems that active site properties of rigid structures alone are not sufficient to deduce an enzyme's activity.

Kurzfassung

In dieser Arbeit wird versucht, berechnete Eigenschaften der aktiven Zentren von Enzymen mit deren jeweiligen Aktivitäten zu verbinden. Experimentell bestimmte Aktivitätsdaten verschiedener Enzyme und Varianten derer werden verwendet, um Zusammenhänge zwischen ihnen und Eigenschaften der jeweiligen aktiven Zentren zu lernen.
Diese Enzyme sind einerseits, eine Decarboxylase (Phenolsäuredecarboxylase, PAD) und ein paar ihrer Varianten mit ein, zwei und drei Aminosäurenaustauschen und, auf der anderen Seite eine Tautomerase (4-Oxalocrotonat Tautomerase) und fast alle ihrer möglichen Einfachvarianten. Die Aktivitätsdaten der Decarboxylase enthalten Messungen der Decarboxylierung von drei verschiedenen Substraten, die Daten der Tautomerase bestehen aus Daten zur Tautomerisierung, aber auch Messungen der promiskuitiven Michael-Addition mit zwei verschiedenen Michael-Donoren. Außerdem produziert eine der beiden Michael Additionen ein chirales Produkt, daher werden auch entsprechende Stereoselektivitätsdaten untersucht.
Mithilfe einiger Machine Learning Techniken, wie Clustering, Dimensionsreduktion und Regression, sollte es ermöglicht werden, nur aufgrund der berechneten Eigenschaften des aktiven Zentrums, die Aktivität unbekannter Varianten dieser zwei Enzyme vorherzusagen. Obwohl unterschiedliche Herangehensweisen ausprobiert wurden, konnten mit den hier verwendeten Methoden keine sinnvollen Vorhersagen zur Enzymaktivität getroffen werden. Es wirkt, als wären die Eigenschaften der aktiven Zentren von starren Strukturen alleine nicht ausreichend, um die Aktivität eines Enzyms vorherzusagen.

Acknowledgements

I want to thank everybody who contributed to the successful completion of my thesis, especially Prof. Karl Gruber and Gregor Wirnsberger, BSc MSc for their supervision and their help with any questions and problems that I had over the course of the last year. I also want to thank Dr. Christian Gruber and the team of Innophore for the possibility to use their resources for this thesis, especially Dr. Michael Hetmann for his support concerning all steps performed on the Catalophore platform, Dr. Markus Fleck, Tobias Schopper, and PD Dr. Andreas Krassnigg for their work on the embeddings and their inputs on cavity procreation in the Catalophore platform, and also Dr. Marco Cespugli and Dr. Ursula Kahler for their thorough rational analysis of conservation and activity data of the tautomerase.
Moreover, I also want to thank my family and friends who have supported and motivated me throughout my studies.

Contents

1 Introduction

The rising importance of biocatalysis in chemical industry has drawn attention to enzyme engineering. The high stereospecifity and selectivity of enzymes makes them a very advantageous catalyst for the production of various fine and bulk chemicals. However, naturally occurring enzymes have limited applications and need to be improved in terms of stability, activity, and turnover rate.
For this reason, different enzyme engineering techniques have been developed, but still every single (supposedly) improved enzyme variant has to be expressed and tested, which makes the process very labor- and time demanding. Therefore, being able to identify few possible candidates based on computationally determined properties would reduce the overall workload and required resources in terms of time, staff and equipment (Braun et al., 2023).

1.1 PAD – Phenolic Acid Decarboxylase

Phenolic acids are plant secondary metabolites that link lignin to hemicellulose and cellulose in plant cell walls and can be sourced from various crops. This natural availability of the starting material has motivated many research groups to test and improve applications of Phenolic Acid Decarboxylases (PADs) (Morley et al., 2013; Sheng et al., 2015).
These enzymes catalyze the transformation of *para*-phenolic acids to their vinyl phenol derivatives, which is of interest to the polymer and food industry, as it yields important precursors for polymers, and different flavor and fragrance compounds (Frank et al., 2012; Sheng et al., 2015).

1.1.1 Structure and Activity

PAD-type enzymes share a high sequence and structure similarity, and while these enzymes exist in many different organisms, and a number of PADs have been characterized, one of the best-studied enzymes is PAD from *Bacillus subtilis* (*Bs*PAD). The proteins form homodimers, each protomer showing a lipocalin fold with a hydrophobic core that contains a number of conserved residues involved in the catalytic mechanism. These conserved residues include two tyrosines that are suspected to hold the phenolic hydroxy-group in place, while an arginine secures the phenolate anion (Figure 1.1, Frank et al., 2012; Myrtollari et al., 2024; Sheng et al., 2015).
Interestingly, not only various PADs show such a conserved active site architecture, but also other enzymes, like 4-hydroxycinnamoyl-CoA hydratase/lyase (HCHL) or vanillyl-alcohol oxidase (VAO) that transform *para*-hydroxylated aromatic substrates have a similar active site.

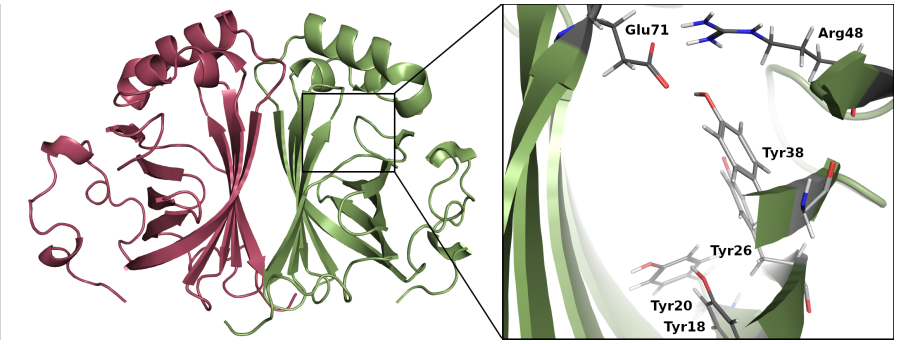

Figure 1.1: **AlphaFold structure of the N31 wildtype dimer** with a zoom into the active site of the enzyme with important catalytic residues colored in grey. In this orientation, at the bottom of the active site, the tyrosines (Tyr18 and Tyr20) that stabilize the hydroxy group can be seen, while the arginine (Arg48) and the glutamate (Glu71) that mediate the decarboxylation are situated at the top. Two other tyrosines (Tyr26 and Tyr 38) with minor catalytic roles can be seen at the right part of the cavity

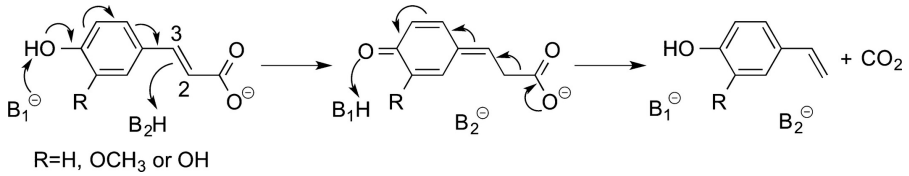

Figure 1.2: **Reaction mechanism of PAD-catalyzed decarboxylation** (Sheng et al., 2015)

It is therefore speculated, that the *p*-hydroxy group, which is hydrogen-bonded to the two active-site tyrosine residues, is crucial for substrate recognition and conversion. Calculations showed that the energetically most favorable reaction mechanism starts with deprotonation of this hydroxy group, which leads to formation of a quinone methide intermediate. Electron flow from the carboxylate group then leads to C-C-bond cleavage and release of CO_2 (Sheng et al., 2015; see Figure 1.2).

Moreover, in contrast to several other enzymatic decarboxylation reactions, the non-oxidative decarboxylation of phenolic acids performed by PAD-enzymes is independent of metals and cofactors (Frank et al., 2012; Sheng et al., 2015).

The enzyme whose activity data were analyzed here, namely PAD N31 (N31), is a (presumed) ancestor of the PAD enzymes. The enzyme's sequence has recently been reconstructed via ancestral sequence reconstruction. N31 shows a high thermostability and a high activity at elevated temperatures. (Myrtollari et al., 2024). The activity data received contained activity measurements of different variants of N31 on three phenolic acids performed by Dr. Kristin Bauer (Bauer, 2024): Caffeic Acid (CAc), Ferulic Acid (FAc), and Sinapic Acid (SAc). Their structures are depicted in Figure 1.3. With the aim to make room for the the larger substituents on the aromatic ring of the substrates, esp. of FAc and SAc, amino

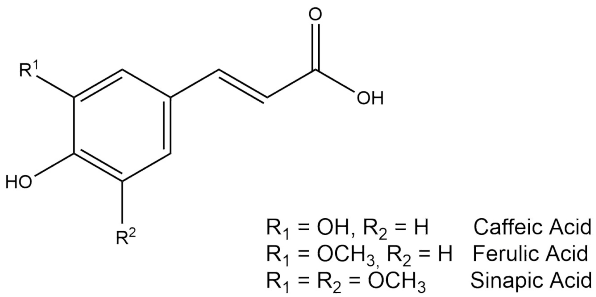

Figure 1.3: **Structures of phenolic acids used as substrates for PAD** created using ChemDraw Professional 16.0

acids located in the active site cavity had been chosen to be exchanged by small and apolar amino acids: In more detail, through overlap extension PCR and Gibson assembly, genes that code for variants of N31 with amino acid exchanges (AAEs) in positions 28, 79, and 92 had been cloned into pET28a vectors. The amino acids that had been available for these amino acid exchanges are alanine, leucine, isoleucine, serine, threonine, valine, or glycine. Note, that isoleucine is the native amino acid in positions 28 and 92, while in position 79 it is leucine. The resulting plasmid library had been transformed into *Escherichia coli* BL21(DE3) cells and after a pre-screen, 88 colonies had been chosen for overnight cultivation, followed by an activity assay.
The overnight cultures (ONCs; 400 µL lysogeny broth, 40 µg/mL kanamycin, inoculated with one of the chosen colonies, or a wt- or empty vector control; two sterile controls were not inoculated; incubated at 37 °C and 320 rpm in a 96-deep well plate) had been used to inoculate the main cultures. These had been prepared in a 96-deep well plate using 6 µL ONC and 594 µL terrific broth auto induction medium (TB-AIM; 12 g/L tryptone, 24 g/L yeast extract, 4 mL/L glycerol, 2.31 g/L KH_2PO_4, 12.54 g/L K_2HPO_4, 0.5 g glucose, 2 g/L α-lactose, 5 g/L glycerol) containing 40 µg/µL kanamycin and incubated for 4 hours at 37 °C, followed by 18 hours at 28 °C. After incubation, cells had been harvested by centrifugation (4 °C, 4000 rpm, 20 min) and pellets resuspended in 300 µL Bugbuster reagent (Millipore) After a short incubation (20 min, 900 rpm), the suspension had been centrifuged again (4 °C, 4000 rpm, 20 min). The activity had been determined by measuring the absorbance of the diluted supernatant (FAc and CAc: 10 µg/mL; SAc: 5500 µg/mL) in 50 mM phosphate buffer (pH 6) containing substrate at a concentration of 1.5 mM at 30 °C and at 335 nm for CAc and SAc, and 333 nm for FAc over 20 minutes. Additionally, the DNA of picked colonies had been sequenced via Sanger Sequencing in order to be able to link the activities to the respective variants.

1.2 4-OT – 4-Oxalocrotonate Tautomerase

4-Oxalocrotonate Tautomerase (4-OT) is a member of the tautomerase superfamily, which consists of several structurally homologous proteins that show a characteristic β-α-β-fold and a conserved proline (Pro1) residue at the N-terminus. Members of this family usually form

hexamers to mediate different reactions, such as tautomerization, dehalogenation, hydration, and decarboxylation reactions. Despite the different activities, Pro1 seems to be a critical residue in all cases, usually as a general acid or a general base, depending on its protonation state (Burks et al., 2010; Poddar et al., 2015; Poelarends et al., 2008; Poelarends & Whitman, 2004; van der Meer et al., 2016; Whitman, 2002; Zandvoort et al., 2011).
Additionally, some of the family members also show promiscuous activities that require the N-terminal proline as well. 4-OT naturally catalyzes a tautomerization reaction that is part of a pathway of bacterial organisms to process simple hydrocarbons as sources of carbon and energy. Intermediates of this pathway can be channeled into the Krebs cycle (Burks et al., 2010; Poelarends et al., 2008; van der Meer et al., 2016; Whitman, 2002; Zandvoort et al., 2011). Apart from this reaction, the enzyme promiscuously catalyzes carbon-carbon-bond formations, such as aldol reactions or Michael-type additions. The latter is an important reaction in the pharmaceutic industry, given that Michael additions of nitroalkenes to acetaldehydes yield chiral γ-nitroaldehydes, which are important precursors in the synthesis of γ-aminobutyric acids. Since this reaction rarely occurs in nature, understanding and optimizing catalysis by 4-OT is of major importance for the pharmaceutic industry (Poddar et al., 2015; van der Meer et al., 2016).

1.2.1 Structure

4-OT forms a homohexamer, each monomer consisting of 62 amino acids. Two monomers interact to form a dimer with a four-stranded β-sheet and antiparallel α-helices. Thus, two active sites are formed, located at both ends of the β-sheet. β-hairpin-loops at the C-termini of each monomer interact with the edges of the neighboring dimer's β-sheet, leading to formation of a hexamer. The total assembly then contains six active sites and residues from three chains are involved in each one (Figure 1.4, Whitman, 2002).

1.2.2 Activity

Tautomerization

In the tautomerization reaction, where 4-OT catalyzes the keto-enol tautomerization of 2-hydroxymuconate to 2-oxo-3-hexenedioate, Pro1 abstracts a proton from the substrate's 2-hydroxyl-group and delivers it to the C-5 atom. To mediate this proton transfer, Pro1 has to be unprotonated, which is achieved by its very low pK_a of about 6.4, caused by its location in the hydrophobic active site (Burks et al., 2010; Poddar et al., 2015; Poelarends et al., 2008; Poelarends & Whitman, 2004; van der Meer et al., 2016; Zandvoort et al., 2011). The hydrophobicity of the active site is maintained by a phenylalanine residue (Phe50), that forms one wall of the active site, preventing water molecules from entering. This residue also provides a platform for the C-terminal β-hairpin to protrude outward, thus stabilizing the hexamer and locking several residues in place. The importance of this residue is highlighted by the observation that the presence of an aromatic residue at this position is conserved in homologous enzymes. In cases where no aromatic amino acid is found at this position, usually it is replaced by a hydrophobic side chain. Furthermore, two arginine residues are involved

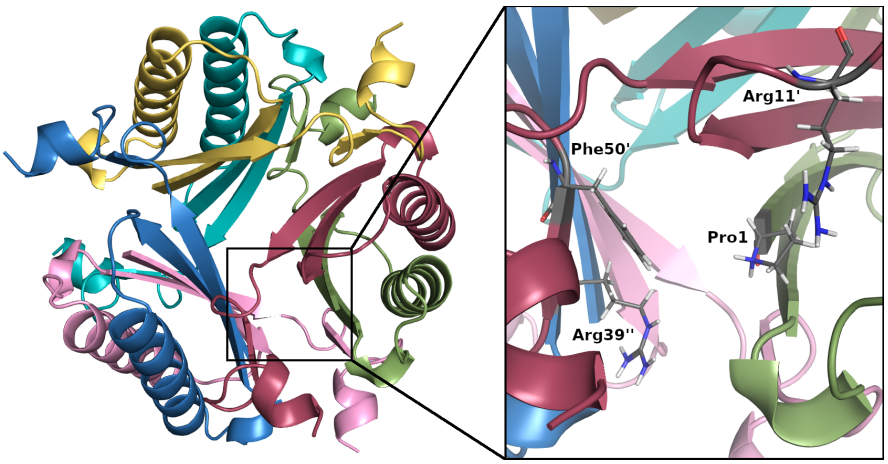

Figure 1.4: **AlphaFold structure of the 4-OT wildtype hexamer** with a zoom into the active site of the enzyme with important catalytic residues colored in grey. The catalytic Pro1 can be seen at the N-terminus of the green chain, two arginine (Arg11' and Arg39'') residues from two different chains (blue and red) can also be seen at opposite ends of the cavity. Additionally, the phenylalanine (Phe50') residue responsible for the low pK_a of the active site is provided by the red chain.

in the reaction mechanism by forming H-bonds: One at position 11 (Arg11) seemingly holds the substrate in place and draws electrons towards the C-6 atom, facilitating protonation at the C-5 atom. The second of the two arginines (Arg39) acts as a general acid catalyst and is needed for the structural integrity of the active site as well as lowering of the pK_a of the active site (Whitman, 2002; Zandvoort et al., 2011).

Michael Addition

Although less is known about this promiscuous reaction, Pro1 is the key catalytic residue also for the Michael addition. However, the proline doesn't act as a base in this case, but as a nucleophile that attacks the carbonyl carbon of the substrate acetaldehyde to form an iminium ion, followed by deprotonation, which yields an enamine intermediate. This activated Michael donor is thus able to perform a C-C-bond formation with the nitroalkene (Poddar et al., 2015; van der Meer et al., 2016; Zandvoort et al., 2011).

The activity data that has been analyzed here originates from an attempt to produce entire mutability landscapes of 4-OT from *Pseudomonas putida* mt-2 for protein expression, activities of tautomerization of phenylenolpyruvate to phenylpyruvate, and Michael additions of acetaldehyde or butanal to *trans*-β-nitrostyrene, and stereoselectivity of the latter Michael addition (van der Meer et al., 2016). Reaction schemes of both, the tautomerization and Michael addition are shown in Figures 1.5 and 1.6, respectively. The activity of each variant was defined by the area of the product peak in a gas chromatography (GC) separation of the reaction mixture. The ratio between each variant's activity and the wild-type activity was then provided for data analysis. In the case of the stereoselectivity screening, areas of the

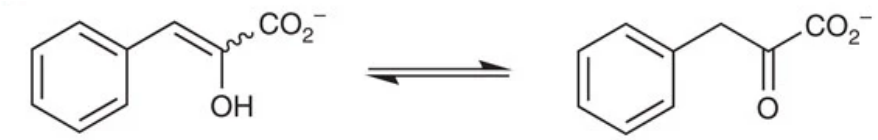

Figure 1.5: **Reaction scheme of the tautomerization of phenylenolpyruvate to phenylpyruvate**, modified from van der Meer et al., 2016

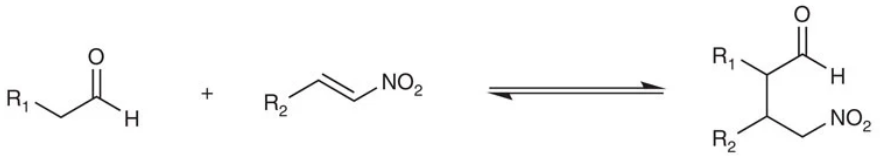

Figure 1.6: **Reaction scheme of the Michaelase activity of 4-OT**; R_1=H or Et, R_2=Ph, modified from van der Meer et al., 2016

peaks of both enantiomers were provided.

1.3 Point Clouds

Active site cavities can be represented by so-called Point Clouds (or Catalophores), that is, a representation of said cavity in three-dimensional space, representing not only the three-dimensional shape, but also several physico-chemical properties, such as hydrophobicity, electrostatics, or accessibility within this space (Hetmann et al., 2023). The points inside the cavity are detected employing the Ligsite algorithm, before all properties are calculated for each point. To detect a cavity, this algorithm starts by placing the protein inside a grid and assigning each grid point a *Ligsite score* of 0. Then, all points where a solvent molecule placed there would overlap with a protein atom, are labelled as solvent-inaccessible and assigned a value of -1. Cavity points are then identified by scanning the x-, y- and z-axes, as well as the four diagonals for *protein-solvent-protein-events* (PSP events). Such a PSP-event takes place, if an area of solvent accessible grid points is surrounded by solvent-inaccessible grid points on both sides along the respective scanning direction. Per PSP-event, the score of each grid point in a solvent-accessible area is augmented by 1. The maximum score is therefore 7 - if the grid point is surrounded by protein points on all axes and diagonals. Lastly, all points of one cavity are grouped together by starting with one grid point whose Ligsite score is at least as high as a chosen threshold, and all its direct neighbors corresponding to the same criterion are added to the cluster, etc. (Hendlich et al., 1997).
Once the cavities are created, all physicochemical properties of each point are calculated.

1.4 Machine Learning Algorithms

Through machine learning, it is possible to process and find a pattern in great amounts of data. The right algorithm might even find patterns that would not be detected otherwise.

Still, without high-quality input datasets, the best learner will not be able to derive relevant information from the data. The two main categories of machine learning methods used in this work are supervised and unsupervised learning. These can be described simply as using labelled or unlabelled input data, respectively ("What is machine learning (ML)?", n.d.).

1.4.1 Clustering

The goal of clustering is to divide data into different groups (*clusters*) in such a way that data points in the same group have a very high similarity, while keeping the similarity of data points assigned to different clusters at a minimum (Gao et al., 2023; Jain, 2010; Xu & Tian, 2015). Because the aim of clustering is to find a natural structure within the data, the input data does not contain any labels and thus clustering is considered an unsupervised learning algorithm. To find this underlying structure, it is always a requirement to measure (dis)similarities between data points, and the different approaches to similarity lead to a number of different algorithms available (Jain, 2010).

K-Means Clustering

The K-Means clustering algorithm is a very simple one, which assumes clusters to be spherical (*center-based partitioning clustering*). It starts with a fixed number (k) of cluster centers and assigns each data point to the closest cluster center. In the next step, the mean of each cluster becomes its new center point and data points are again assigned to their closest cluster centers. These two steps are iterated until the positions of the cluster centers do not change substantially between two iterations or a maximum iteration threshold is reached. The relatively simple algorithm makes the results of the cluster assignments easy to interpret. However, the algorithm is sensitive to its initialisation method, to outliers, and cannot identify clusters with a non-spherical shape and is very likely to converge in a local minimum (Gao et al., 2023; Jain, 2010; Xu & Tian, 2015).

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

In contrast to K-Means clustering, DBSCAN is a density-based clustering algorithm; it will therefore try to assign closely neighboring points to the same cluster, while points far away from dense regions will remain unclustered. This makes the algorithm insensitive to outliers and able to find clusters with a non-spherical shape. DBSCAN might have problems detecting clusters that have different densities and will become computationally expensive with large, high-dimensional datasets (Gao et al., 2023; Jain, 2010; Xu & Tian, 2015).

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

A third category of clustering mechanisms is hierarchical clustering, which builds a feature tree that divides the data points into clusters, that become smaller at each level of the tree. In the first step, subclusters are established based on a part of the dataset. Then, the rest of the data are fitted into this preliminary tree-network. All leaf nodes have a maximum size,

and once any cluster becomes too big, cluster assignments are recalculated to keep within the maximum size range. Also, BIRCH clustering is robust to outliers, but it is not very suitable for non-spherical clusters and is sensitive to the order in which the data is presented. On the other hand, this algorithm is fast and can be used in combination with other clustering algorithms, if desired (Gao et al., 2023).

1.4.2 Dimensionality Reduction

When the number of features in a dataset is relatively large compared to the number of training patterns, a multitude of model parameters have to be fitted, which leads to a high risk of overfitting a trained model on the training data. To overcome this problem, it makes sense to reduce the number of features per data point. However, one must be careful not to lose important information contained in the omitted data. Thus, dimensionality reduction should yield a low dimensional representation of the dataset but preserve as much of the overall data structure as possible (Amid & Warmuth, 2019; Guyon et al., 2002). Apart from being an important pre-processing step for machine learning, dimensionality reduction also facilitates visualization of the data (McInnes et al., 2020).

Principal Component Analysis (PCA)

PCA is a very simple linear technique that aims to produce a low-dimensional representation of the data that retains as much of the variance in the data as possible. It produces a set of uncorrelated variables (Principal Components, PCs) through linear combinations. These PCs are numbered according to the amount of variance they retain - that is, the most variance is retained by the first principal component (PC1), the second most by the second principal component (PC2), etc. Furthermore, PCs must be orthogonal to each other in order to ensure that the information contained in PC1 cannot be correlated to that in PC2 (Van Der Maaten et al., 2009; "What is principal component analysis (PCA)?", n.d.).

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)

UMAP attempts to reduce data dimensionality by approximating a manifold, on which the data is assumed to be uniformly distributed, and constructing a simplicial set representation of this approximated manifold. Simply put, this representation is an algebraic model, which retains the fundamental topological features and structure of the manifold.
Therefore, it is assumed that a manifold exists, on which the data is distributed uniformly and that this manifold is locally connected. The primary goal of UMAP is to preserve the structure of this manifold. Other than PCA, UMAP is a non-linear dimensionality reduction technique and the resulting dimensions do not have a meaning (Amid & Warmuth, 2019; McInnes et al., 2020).

t-distributed Stochastic Neighbor Embedding (t-SNE)

Also t-SNE is a non-linear dimensionality reduction technique, but in contrast to the other two algorithms, higher importance is assigned to preserving local structure instead of global structure. The dimensionality reduction of the input data is based on calculation of pairwise distances between points. t-SNE is able to simplify the structure of data lying on different manifolds. However, this algorithm is computationally expensive and random, so different runs will return different results (Amid & Warmuth, 2019; Schubert & Gertz, 2017; "scikit-learn User Guide", n.d.; van der Maaten & Hinton, 2008).

1.4.3 Feature Selection

Similar to dimensionality reduction, also through feature selection, the number of variables per data point can be reduced in order to improve the accuracy of the result by preventing e.g., overfitting ("scikit-learn User Guide", n.d.). This can be achieved by applying one of the following methods.

Recursive Feature Elimination (RFE)

In RFE, an estimator is run first over the whole dataset, and then, the estimator is fitted again several times, omitting one feature of the dataset in every run. The new dataset that shows the least difference from the original dataset is then the new starting dataset, as the missing feature of this dataset seems to have the least impact on the algorithm's performance. This step is repeated until a desired number of remaining features is left in the dataset, or until the model's performance doesn't improve anymore upon deletion of features (Guyon et al., 2002).

Permutation Feature Importance (PFI)

PFI follows a similar principle, but instead of deleting features, a feature's values are randomly permuted before recalculating the model. Thus, if the result does not change significantly, it shows the little impact of this feature on the result. On the other hand, the permutation of a feature with a high importance will result in a high error rate. In this way, features can be filtered based on their impact on the model and used for further analyses (Breiman, 2001).

1.4.4 Regression Analysis

During a training phase, regression algorithms aim to find a relationship between input data to use their findings later, in order to predict values of unknown samples ("What is machine learning (ML)?", n.d.). Different learners exist, that take up information of the training data in different ways and apply this acquired "knowledge" to the test dataset.

K-Nearest Neighbors (KNN) Regression

This very straightforward estimator assigns a value to a point in the test dataset based on its k closest neighbors in the training set, where k is a user-defined parameter, and the assigned value is the mean of the neighbors' labels. The simplicity of this regressor makes it easy to implement, as not many parameters exist that need tuning ("scikit-learn User Guide", n.d.).

Ridge Regression

Ridge Regression calculates a linear regression equation for the dataset, but adds a penalty-term that reduces the value of coefficients. This penalty-term affects large coefficients stronger than small ones. The former are often related to problems like overfitting of the estimator on the training set or multicollinearity of the input data, which are therefore mitigated by the regressor (Murel & Kavlakoglu, 2023).

Random Forest

Random Forest is a so-called ensemble method, which combines multiple decision trees in order to improve the overall accuracy of the model compared to one single estimator and to minimize overfitting. Random Forest averages the decisions of multiple decision trees to its final prediction - taking advantage of the fact that, as the number of estimators grows, the prediction error converges to a minimum (Bentéjac et al., 2021; Rokach, 2016; Wu et al., 2021). When a tree is generated, the first split criterion (at the root node) is calculated to achieve the best partition in the dataset. Depending on its concordance with the splitting criterion (true/false), each data point is then passed along a branch to the next node, where again, a new partition takes place. This procedure is repeated, until a stopping criterion (e.g., less than minimum samples for a split, or the maximum depth of a tree) is reached (Rokach, 2016). An example for such a decision tree is shown in figure 1.7. To increase variation between the estimators, each tree in the ensemble is trained on a random subset of the training data, and at each split, only a random part of the input features are available for selection of the best-split criterion. (Bentéjac et al., 2021; Wu et al., 2021).

Gradient Boosting

Also Gradient Boosting combines several weak learners in order to get a strong one. In this case however, each of these takes the errors of its preceding weak learner as input and corrects them. In order to prevent overfitting, a regularization term is applied at each gradient descent step (Bentéjac et al., 2021; "What is boosting?", n.d.).

1.5 About this work

Here, active site properties are calculated for variants of two different enzymes, namely a phenolic acid decarboxylase (PAD) and a tautomerase (4-OT). With these calculated data, together with activity data of the respective enzymes, provided by different groups, the goal

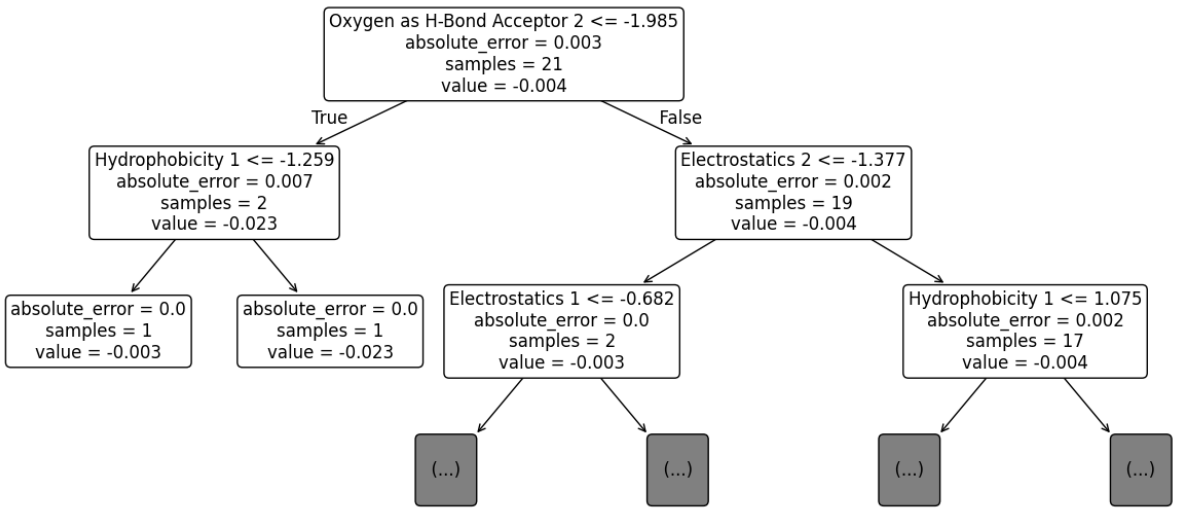

Figure 1.7: **An example decision tree of the PAD regression analysis.** At each non-leaf node, the first line contains the split criterion; the remaining lines of all nodes state the absolute error of all samples at this node, the number of samples passing through that node, and the value of the node.

of this thesis is to find properties that characterize an enzyme with high or low activity using different machine-learning approaches. These activities include the decarboxylase activity of PAD against three different phenolic acids and activity of 4-OT for a tautomerization reaction and two different Michael additions, as well as stereoselectivity data for one of the Michael additions.

2 Methods

2.1 Wetlab Data

Datasets were converted from the original format to uniform files. To do so, different workflows were necessary, depending on the original format.

2.1.1 PAD

PAD data were received as absorbance values of the reaction mixtures measured at 335 nm at different timestamps over a total of 20 minutes (18.85 min). Using this data, the rate of absorbance decrease for every measurement (slope) and the respective statistical estimators of all measurements of each variant were calculated. That is, not only the respective triplicate measurements, but also measurements of variants that occurred twice in the library were averaged.
Finally, one file per substrate was prepared that contained the respective slopes for each variant.

2.1.2 4-OT

Tautomerization and Michael Additions

The 4-OT data were presented as relative values of the area of the GC-product peak with respect to the corresponding product peak of the wildtype (wt) reaction. In the following analyses, the $\log_{10}$ of these values will be referred to as the *score* of the variant.

Stereoselectivity data

In this case, the areas of both peaks in the chromatogram - corresponding to the two enantiomers - were delivered. Therefore, the $\log_{10}$ of the enantiomeric ratio of the two products (*2R3S:2S3R*) was calculated for each variant and will be referred to as the *score* of the variant throughout this work.

2.2 Point Cloud Generation

The limited number of PAD-variants relevant for these analyses allowed for all structures to be predicted using AlphaFold 2 Multimer with the parameters listed in Table 4.1 (Evans et al., 2022; Jumper et al., 2021; Mirdita et al., 2022). Then, the relaxed structures of ranks

1-3 of all variants were uploaded to the Catalophore platform ("uni-graz.catalophore.com", 2024).
In contrast, the structures of the over 1000 variants of 4-OT were obtained by uploading the AlphaFold-structure of the wildtype enzyme to *Darwins Playground* of the Catalophore Platform and generating the variants via the function *request mutant*. The option to minimize structures was switched to *yes* for this step.
Additionally, for the embeddings, shaped cavities were prepared, in order to be able to select relevant cavities on the platform. Therefore, the nitrostyrene ligand of the crystal structure (PDB code: 5CLO; RCSB.org, Berman et al., 2000) was transformed into chain B of the AlphaFold-structure of the wildtype enzyme. Based on the rotation matrix of each of the other chains, the ligand was transferred into all five remaining active site cavities to obtain the protein structure with a ligand in all six active sites. This structure was then uploaded to the Catalophore Platform as a scaffold for Homology Modelling of all variants.

After the structures had been added to their respective collections, the Cavity Procreation was started with default settings (apart from adaptions to the minimum and maximum cavity size for 4-OT and in a second run of a cavity procreation of PAD, Table 4.3). The results of the first cavity procreation of PAD were used as input for ligand docking and cavity matching, while the results of the following procreation experiments were downloaded directly after a rough filtering step, through choosing *Aligning residues (sidechain contacts)*, $>$ and entering "ALA28, ALA79, ALA92" in the search box. This step was first found coincidentally and then intended to remove all cavities that were found, apart from the true active site cavities.

2.3 Ligand Docking and Cavity Matching

The structures of the three ligands (CAc, FAc, SAc; PDB codes: DHC, FER and SXX, respectively) were obtained from the PDB, and used for docking experiments in the Catalophore platform. The same cavities were also used as input for a cavity matching experiment.

2.4 Download of Point Cloud Data

Due to the platform lacking options to download all cavities at once, all data was downloaded employing web scraping methods ("Selenium", 2024). In the case of the first round of of PAD cavity procreation, whose resulting cavities were used for ligand docking, the dockings were downloaded and relevant information about the single Point Clouds was extracted from these files. Differently, the cavities created in the second round of cavity procreation of PAD and that of 4-OT were downloaded without docking. These files could still be used to read the Point Cloud information and add it to the structure file.
In all cases, a PDB file was created that contained information about the protein structure and the Point Clouds.

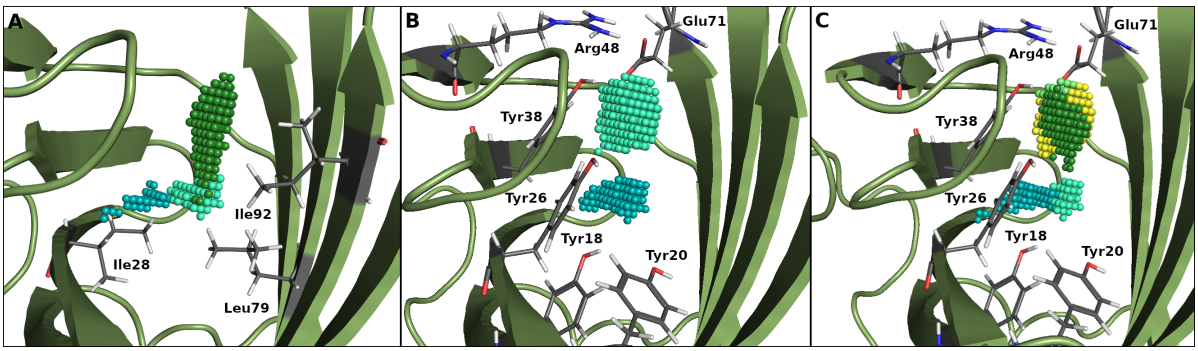

Figure 2.1: **Point Cloud splits of the wildtype cavity** In A, points in the vicinity of the positions 28, 79, and 92; i.e. those subjected to AAEs are shown in different colors. B: The cavity was divided into two parts; an upper part, which would supposedly surround the carboxyl group of the ligand, and a lower part, that would contain the phenol-ring. Lastly, in C, points were selected that are in the vicinity of either of the six residues involved in the mechanism (Tyr18, Tyr20, Tyr26, Tyr38, Arg48, and Glu71)

2.5 Point Cloud Splits

In order to better filter certain areas of the Point Clouds that are influenced by different amino acids, the Point Clouds were split into sections:

2.5.1 PAD

At first, it was attempted to consider those points of the Point Clouds that are in the vicinity of each ligand atom. Two other ways the Catalophores were divided, were by choosing points in the vicinity of either the residues that are involved in the reaction mechanism (Tyr18, Tyr20, Tyr26, Tyr38, Arg48, and Glu71; catRes), or those that were subjected to amino acid exchanges (Positions 28, 79, and 92). Lastly, points that are found in the part of the cavity that is expected to contain the ligands were divided into two groups at either end of the cavity part (catHalves). Thus one of these groups is expected to surround the phenol part of the ligand, while the other one should surround the carboxylic acid part. Figure 2.1 shows these splits of the wt-cavity. To narrow the datasets further prior to the regression analysis, different modes to select specific cavity properties were attempted. At first, three properties that should be beneficial to the reaction according to chemical reasoning were chosen, namely *Hydrogen-bond donor point-cloud*, *Oxygen as H-bond acceptor point-cloud*, and *Electrostatics point-cloud*. Lastly, the properties were clustered (using the K-Means algorithm), and for every cluster, the property with the highest total covariance was chosen as representative for that cluster and used as input for regression analysis.

2.5.2 4-OT

In the case of the tautomerase, only groups of points in the vicinity of some manually chosen amino acids (Pro1, Ile2, Gln4, His6, Leu8, Gly10, Arg11, Ala33, Ser37, Arg39, Met45, Phe50, Gly53, Gly54, Ala57, Arg61) were considered for further calculations. These amino acids were the ones found to have a significant impact on any of the tested enzyme activities (van der Meer et al., 2016). Moreover, the points in the vicinity of the respective important residues for each of the reactions were considered in another instance of regression analysis. In another attempt to find a relationship between Point Clouds and activity, point cloud properties were clustered, and a representative of each cluster was chosen for analysis, as described in section 2.5.1.

2.6 Embeddings

The shaped cavities created from the homology models of the 4-OT variants were filtered by their ligand coverage and cavities with a value $> 55\%$ were selected for the embeddings. These cavities resulted to be the ones whose ligand originates from the nitrostyrene in the crystal structure. Using Innophore GmbH's proprietary deep learning model, cavities were compressed into embeddings featuring 64 descriptors, i.e., the embedding length. These embeddings were subjected to further analysis as follows.

2.7 Analyses

Point Cloud data of all variants were extracted and parts of the catalophores were chosen acoording to the different Point Cloud splits. Property values of all points of one split were averaged and used as input for various analyses aiming to find a connection between cavity properties and enzyme activity.
Before each analysis attempt, Point Cloud data were converted to standard normal distributed values.

2.7.1 Clustering

In a first attempt to find active site properties that indicate a high enzyme activity, the data were used as input for different clustering algorithms from the scikit-learn (sklearn, version 1.5.0; Pedregosa et al., 2011) library. In all cases, the assigned clusters of all data points were plotted against their respective activity score.

2.7.2 Dimensionality Reduction

Two dimensionality reduction algorithms were run to reduce the large number of features per data point (especially in the case of PAD data). The results of PCA and UMAP analyses were again plotted, and the data points were colored according to their respective activity scores. Also, RFE and PFI analyses were performed on the dataset in order to find Point

Cloud properties with a high impact on the various enzyme activities. As for clustering, also all of the dimensionality reduction algorithms were taken from the sklearn library (Pedregosa et al., 2011).

2.7.3 Regression Analysis

Finally, different regression algorithms were tested for their ability to find a connection between physicochemical active site properties and activity of a given variant. RandomForestRegressor, KNeigborsRegressor, GradientBoostingRegressor and Ridge Regression from the sklearn library (Pedregosa et al., 2011) were applied to process the Point Cloud Data, first with the default values, but also with custom parameters. Employing RepeatedKFold, the input data was split into multiple different training and test datasets and results were evaluated by calculating the average and standard deviation of the mean pearson R values of the single runs. To optimize parameters for RandomForestRegressor and GradientBoostingRegressor, a GridSearchCrossValidation (GSCV, also from the sklearn library; Pedregosa et al., 2011) was set up. The parameters that have been optimized using GSCV, are listed in section 4.8

2.8 Rational Analysis and Sequence Alignment

Also Marco Cespugli from Innophore tried to analyze the data received for 4-OT. Together with Ursula Kahler (also from Innophore), he tried to combine findings about the conservation of each position to experimental results by performing a sequence alignment and assessing the impact of each AAE in each position of the protein.

3 Results and Discussion

3.1 PAD

Results obtained with the PAD dataset have to be interpreted with care, as the enzyme had not been purified, but cell-free extracts had been used for the measurements, absorbance data had not been normalized in order to eliminate differences in expression between the different variants. Therefore, absolute values were used, assuming constant enzyme expression across all variants. Additionally, the slope measured for the sterile control lies within the best third, half, and two thirds of the data, respectively in the case of CAc, FAc, and SAc data and some data of the decarboxylation of SAc showed positive average slopes.
In conclusion, these issues limit the quality of conclusions drawn from the data, but analyses were attempted anyway, to test the limits of the methods.

3.1.1 Point Cloud Generation

The PAD dataset contained measurements of enzyme variants with amino acid exchanges in three positions. Since only seven amino acids were available fore these exchanges, this gives a total of $7^3 = 343$ different variants. All three ranks of each AlphaFold prediction were used to generate Point Clouds and for all of the 1029 structures, cavities could be detected. In total, 2656 cavities were found; that is approx. 2.6 cavities per variant. However, for some of the structures (e.g., the wildtype structure), the active site cavity was missing. Later, the reason for this was found to be a too small cutoff for the minimum cavity size. In an attempt to filter the cavities to obtain a collection of only active site cavities, it was possible to retain only one active site cavity per input structure (instead of two, which would correspond to one cavity per monomer). Interestingly, this collection still contained very few non-active site cavities, which were removed in a later step. In a second attempt of Cavity Procreation, the minimum cavity size was set to a lower value (15 $Å^3$) and in this case, 6817 cavities (approx. 6.6 cavities per input structure) were detected by the algorithm. This time, 1175 structures remained after the filtering step and also here, some "wrong" cavities remained in the dataset and only one active site cavity was retained per structure.
Apart from Cavity Matching, which was only performed with the first, smaller set of cavities, all results described here, refer to the second, more complete set.

3.1.2 Ligand Docking and Cavity Matching

Ligand Docking

Ligand docking quality varied greatly between different structures. A comparison between the ligand binding, as described in section 1.1.1, and dockings into smaller and larger Point Cloud cavities showed that the ligand could be placed very well into small cavities, while multiple, very different binding modes were found in bigger cavities. These trends were noted across all three ligands, that is, enlargement of the active site cavity in order to fit a bigger ligand seems to make it more difficult to find a suitable binding mode - at least computationally (see Figure 3.1).

Cavity Matching

Cavity Matching was attempted and resulted in 289444 cavity-matches. However, the amount of data produced made it challenging to extract valuable information from the results. In the end, the embeddings produced at Innophore (see chapter 3.1.3) were also able to summarize cavity similarity and could be directly used as input for analyses like t-SNE and regression. Therefore, these results were preferred over the Cavity Matching.

3.1.3 Analyses

Dimensionality Reduction

The cavity dataset contains a high number of features compared to the rather small number of variants, and some of these features show high collinearity. Therefore, dimensionality reduction was the first method that was attempted in searching a correlation between the measured slopes and cavity properties. A PCA was able to produce principal components, where the first one usually showed greater variance, while a sharp decrease could be observed over PCs 2 to 5. The first two PCs of the dataset containing information from the Point Clouds around catalytically important residues are mapped in Figure 3.2, with the points colored according to the slopes of the decarboxylation reaction of FAc. UMAP dimensionality reduction is based on a different algorithm, and therefore, also this method was attempted, to see if it is possible to find something that was missed in PCA. However, also this approach did not produce any corrrelating results.

Clustering

In the next step, it was tested whether clustering algorithms would find a way to divide the dataset into groups of variants with similar activities. Because the structure of the data was not known, three different clustering algorithms were chosen that would search for clusters in the datasets based on different criteria: A center-based (K-Means), a density-based (DBSCAN), and a hierarchical (BIRCH) clustering algorithm. K-Means and BIRCH were able to find clusters, while DBSCAN did not divide the dataset at all. The algorithm assigned a value of -1 to all data points, which indicates noisy data ("scikit-learn User Guide",

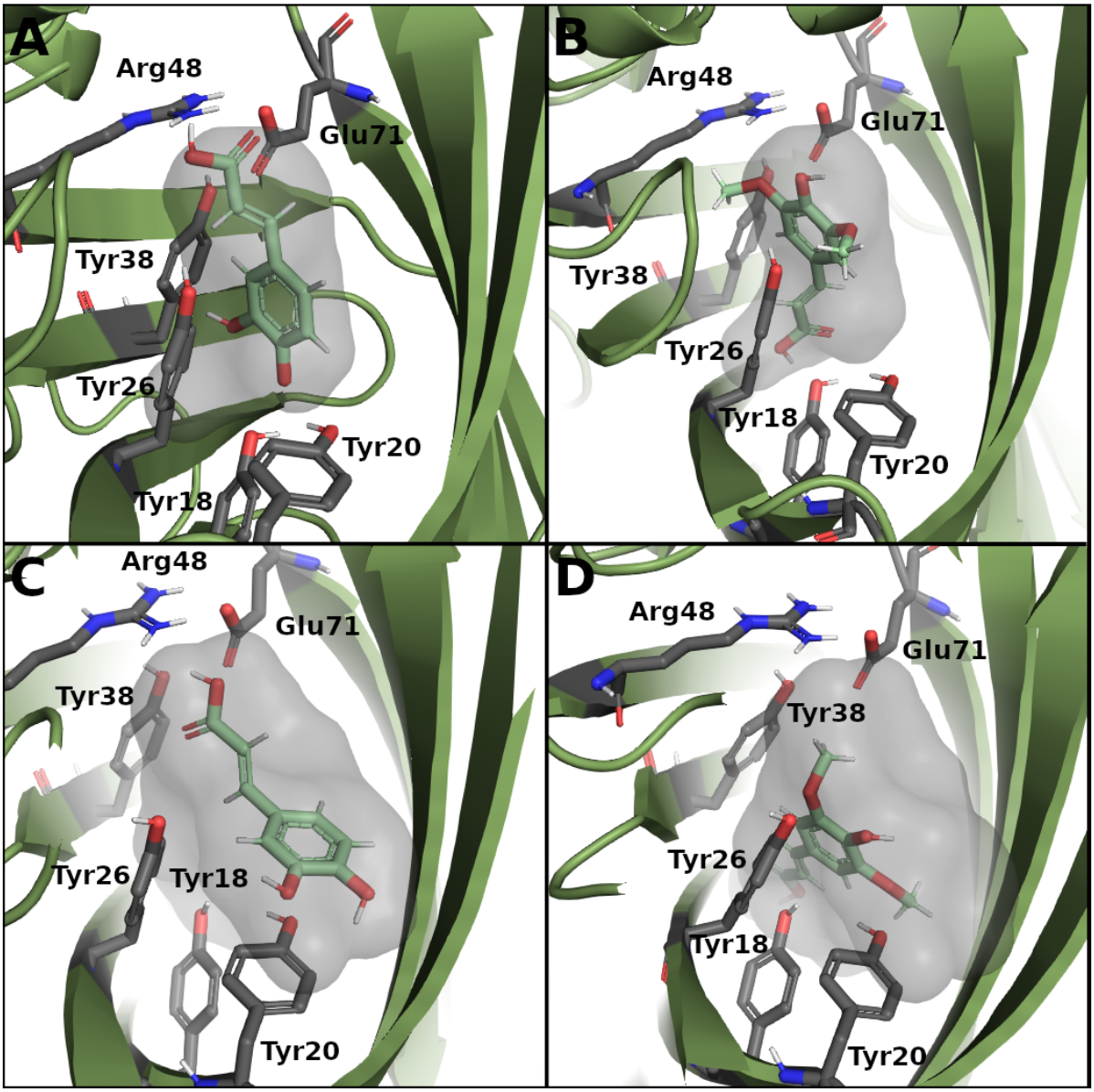

Figure 3.1: **Dockings of CAc and SAc into the wt and the biggest (GGG) cavity.** A: Docking of CAc into the wt-cavity is in accordance with the calculated binding mode: Tyr18 and Tyr20 hold the *para*-hydroxy-group of the substrate in place, Arg48 is found in close proximity to the substrate-carboxylate group (Sheng et al., 2015). B: Here, SAc was docked into the wt-cavity. The ligand is rotated approx. 180 degrees with respect to CAc in A. C: In the biggest cavity of the dataset, CAc was docked roughly in the right orientation, but the *para*-hydroxy-group is rather distant from the two tyrosine residues. In D, SAc was docked into the cavity of the GGG-variant. However, the bigger cavity doesn't seem to lead to a more correct binding of the ligand, as the two substituents on the phenol-ring occupy the space between the arginine- and the two tyrosine residues.

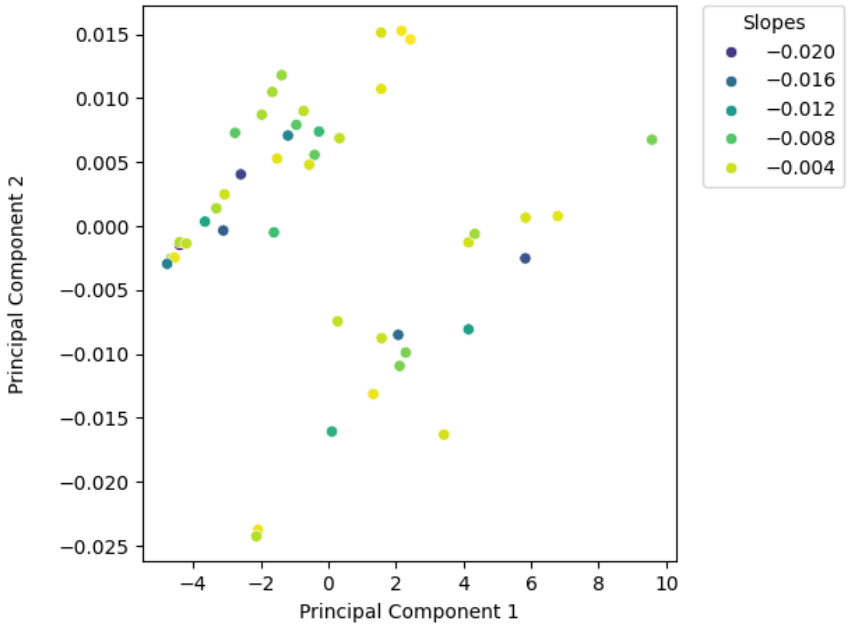

Figure 3.2: **First two principal components of Point Clouds of catalytically important residues.** The points are colored according to the slopes of FAc-decarboxylation. Although some of the points form groups, no correlation between the position of the points and their slopes can be made.

n.d.). Since the algorithm takes only cavity data as an input and not the measured slopes, this might indicate a problem when using Point Cloud data or insufficient differences in the data. However, the clusters found by the other two algorithms (K-Means and BIRCH) did not show any correlation with the measured slopes (Figure 3.3).

Feature Selection

Another possibility for identifying important features of a dataset is through feature selection. Here again, two different approaches were tested. PFI returned inconsistent results, probably because all of the properties were assigned low importance values with high standard deviations over the cross-validation runs. Predictions did not show any changes upon permutation of any of the features. An explanation for this problem might be that, due to the high covariance of the data, as one feature is changed, another similar feature is used as a predictor by the algorithm. On the other hand, collinearity should have a smaller impact on RFE, as here, the least important feature will be deleted from the dataset. Thus, features of a group of collinear data will be eliminated until only one feature of that group remains in the dataset. Yet, also, the results obtained by RFE were not consistent over multiple rounds, and no optimal number of features to use for prediction could be calculated with high accuracy.

Regression

Finally, regression analysis was attempted using two different regressors, namely Ridge and Random Forest regression. Neither of them could accurately predict the activities of the different variants. The best result was obtained with Random Forest Regression using Point Cloud properties around the residues subjected to AAE as input data. In Figure 3.4, the predicted slopes of one of these runs are mapped against the measured (true) slopes of SAc decarboxylation. The overall mean pearson R value of 10 runs is 0.401 ± 0.244. In another attempt to improve predictions of these regressors, some input features were chosen manually, based on different criteria (see chapter 2.5.1). Nonetheless, neither of these approaches improved predictions significantly. Overall, mean pearson R values that were obtained with PAD data showed great variations between single runs - resulting in high standard deviations. This might be due to the issues with the measurement data discussed in Chapter 3.1, or because the chosen method of activity prediction is not applicable for this case. Detailed results are listed in Tables 4.9 to 4.11

Classification

Assuming that the main problem encountered while trying to predict the slope of the reaction catalyzed by a given enzyme variant is the missing normalization in the data, a normalization-independent analysis is the prediction of the substrate, towards which a variant has the highest activity. However, for most of the variants (40%), the steepest slope was measured for SAc, resulting in biased input data and therefore also these predictions did not produce meaningful results. Figure 3.5 shows the predicted classification vs. the actually measured

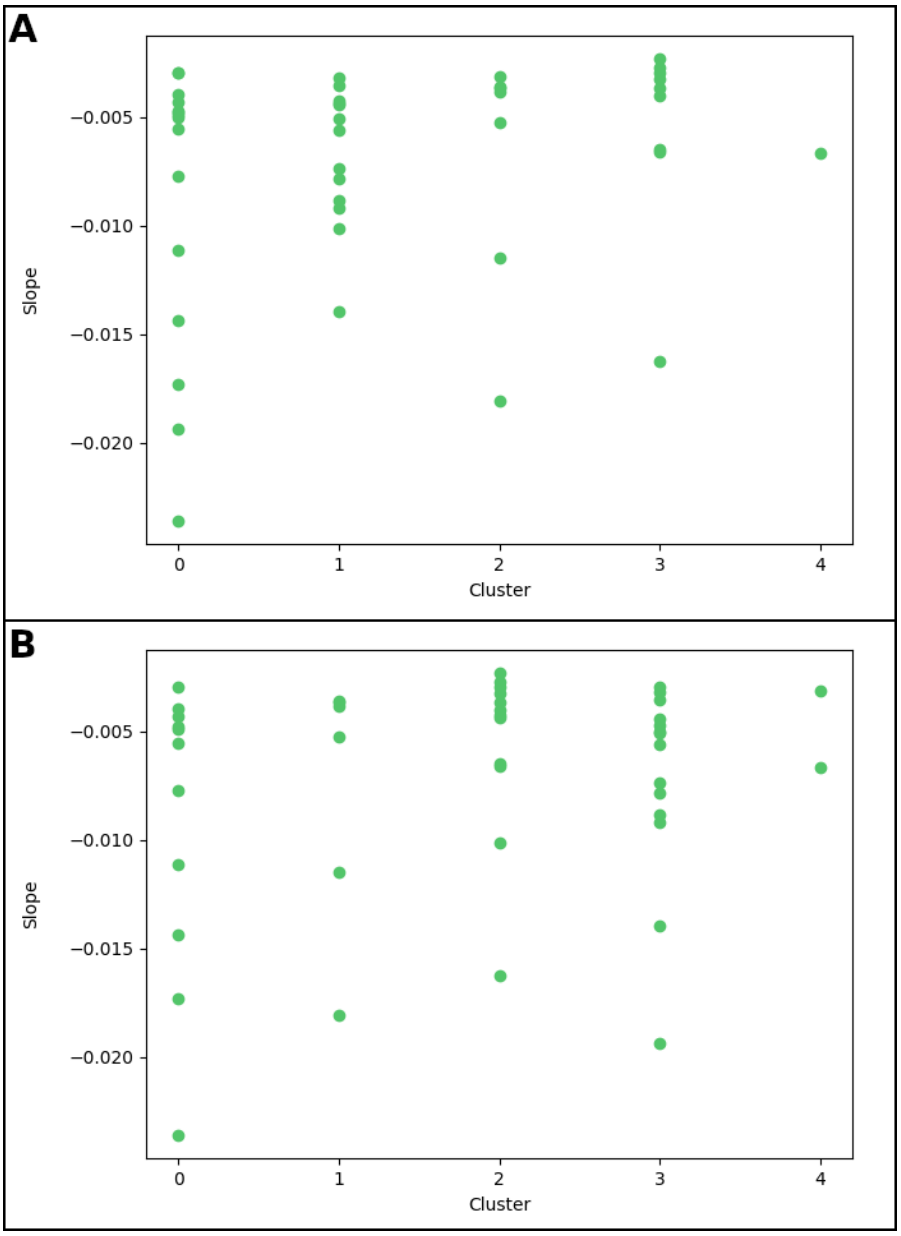

Figure 3.3: **Clustering results of point cloud properties around catalytically important residues.** Neither BIRCH clustering (A), nor K-Means (B) could produce clusters with similar measured slopes.

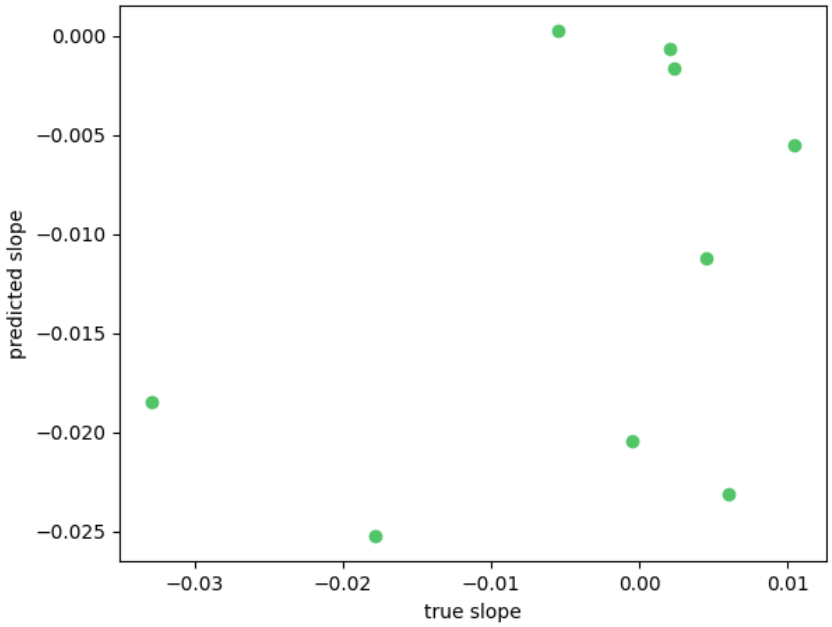

Figure 3.4: **Plot produced in one of the Random Forest regression runs.** The input data used here were the properties of the point cloud sections close to the amino acids in positions 28, 79, and 92. The slopes of SAc decarboxylation were used as y-values.

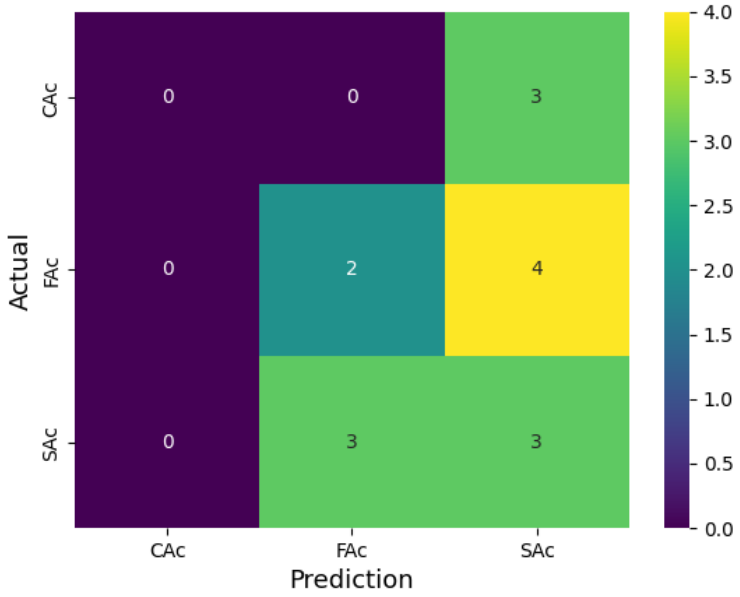

Figure 3.5: **Confusion matrix of the prediction of which ligand's decarboxylation is predicted to have the steepest slope** based on the Point Cloud properties close to catalytically important residues.

fastest decarboxylation. This plot shows the strong bias of the classifier towards SAc, while 50 % of the variants that should have been classified as SAc were misclassified. This might be due to the cavities being too similar altogether, or, as stated above, due to the usage of a model not capable of handling the presented data. All results can be seen in Tables 4.6 and 4.7.

Embeddings

Lastly, from the embeddings, one cavity could be detected that was very different from the others: Markus Fleck from Innophore calculated Matching Scores of the cavity with respect to the Human Cavitome, as a baseline, and to all PAD cavities. Through comparison of these two, he could identify this cavity as an exot, that is, more different than would be expected for that dataset. It was later identified as one of the few "wrong" cavities still present in the dataset used for the embeddings. However, this shows that in general, it is possible to detect cavities, that differ from the others. The question, whether cavities of similar variants, that differ by 1 - 3 AAEs are different enough to be detected by this method remains to be answered in the second dataset.

3.2 4-OT

3.2.1 Darwin's Playground

An AlphaFold model of the wildtype structure was uploaded to the catalophore platform and structures of all variants were produced. Due to this process crashing and being restarted, 1277 structures were produced, instead of 1220, which were planned (AAEs that resulted in the wt, like I2I, for example, were also produced). Among these structures, each variant occurred at least once, and some of them multiple times, but they were filtered in a later step.

3.2.2 Point Cloud Generation

Cavities were detected for all of the variants, with 11094 total cavities (approx. 8.7 cavities per variant) and filtering did not work in this case, so all of these cavities were downloaded and filtered later. Interestingly, some of the cavities extended so far that multiple cavities merged together to form a bigger one. In this way, cavities were found that contained up to six active sites, but most of the cavities contained only one or two active sites.

3.2.3 Analyses

Dimensionality Reduction

This dataset contains data of more variants than the PAD data, and therefore, the large number of values per datapoint is a lower risk for overfitting a model with too many parameters, but many features show a high collinearity. Hence, it was again attempted to reduce the dimensionality of the data. Also here, the resulting PCs and dimensions were not correlated to the measured reaction rates, but greater differences between the values of the PCs were observed instead of the steep decrease in variation seen in the first dataset. Figure 3.6 shows the first two dimensions resulting from UMAP dimensionality reduction with the points colored according to the tautomerization scores. Most of the data points are found in a big cluster of points, but again, no correlation could be found between the scores of the single points and their position.

Clustering

Due to a lack of knowledge of the structure of the data, the data were clustered in order to find groups of variants with similar scores. However, also in the case of 4-OT data, no correlation could be found between cluster assignment and scores of the variants. As already observed in the PAD dataset, the DBSCAN algorithm assigned a value of -1 to all points in the dataset, indicating noisy data. The other two algorithms (K-Means and BIRCH) were able to assign every point to a cluster, but comparing the cluster assignments to the slopes showed no correlation (Figure 3.7).

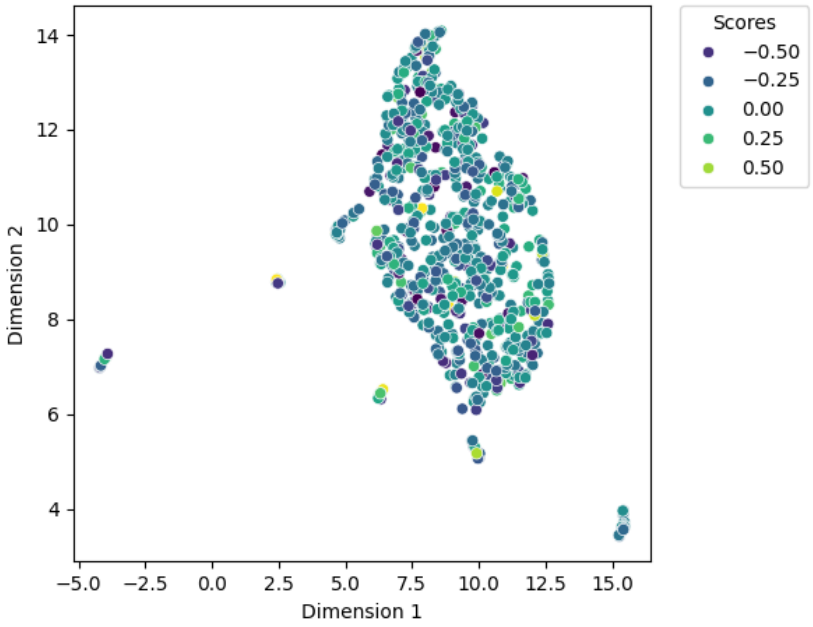

Figure 3.6: **First two dimensions of UMAP dimensionality reduction of 4-OT data.** The points are colored according to the scores of the tautomerization. Most of the points are found in one big group, but no correlation between the position of the points and their scores can be made.

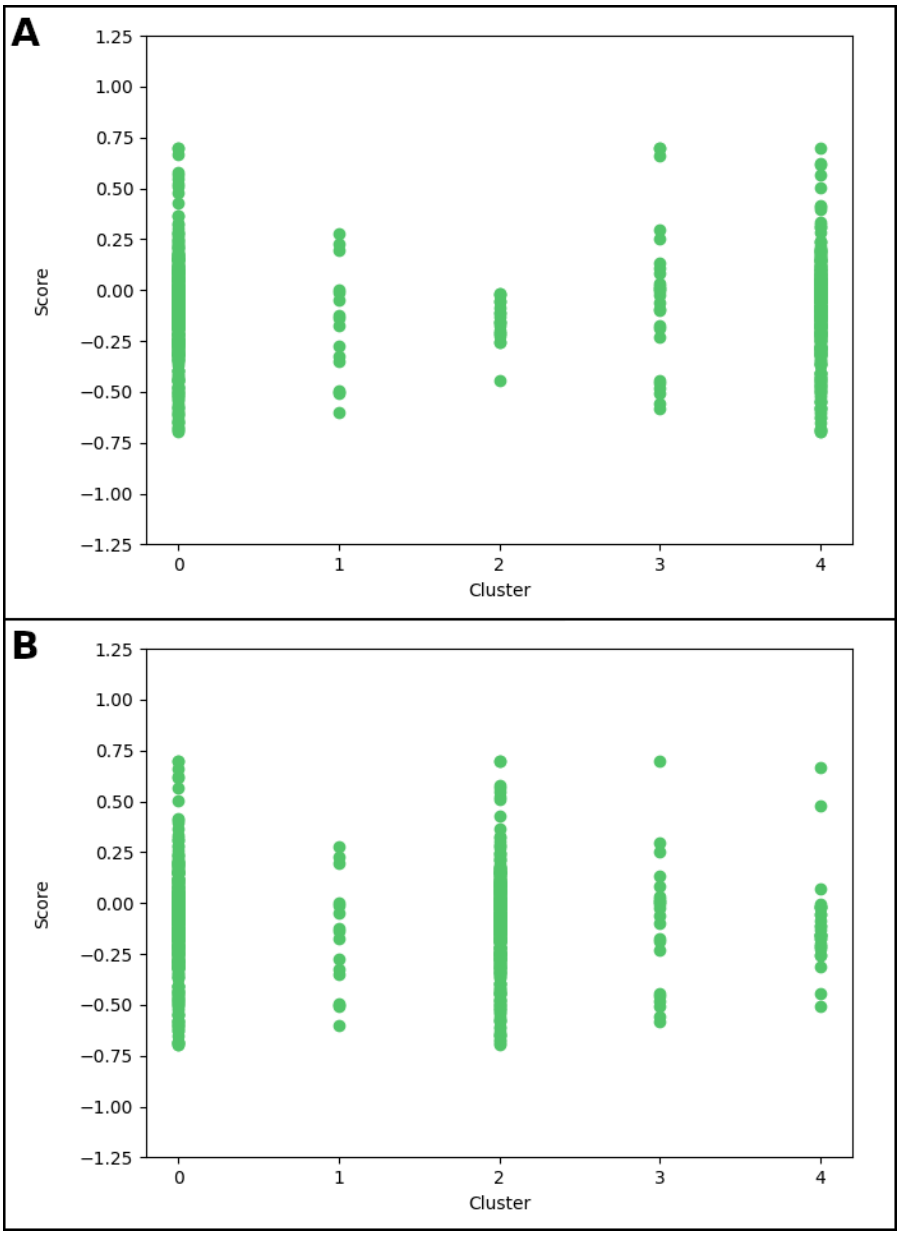

Figure 3.7: **Clustering results of the 4-OT dataset.** Also in the case of the 4-OT dataset neither BIRCH (A), nor K-Means (B) clustering produces clusters that can be correlated to the scores of the tautomerization reaction.

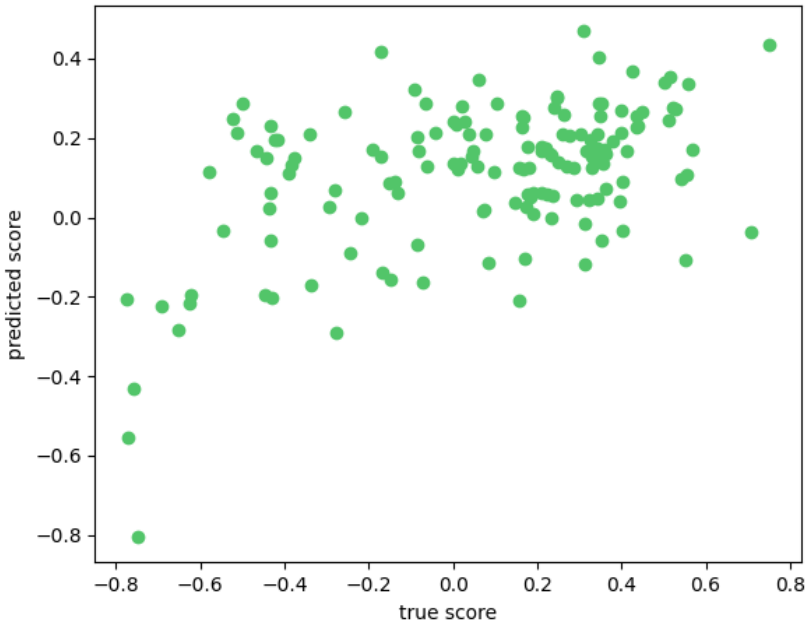

Figure 3.8: **Gradient Boosting Prediction of the stereoselectivity data of 4-OT.**

Feature Selection

Although the results of the PFI analysis were less inconsistent than those obtained with the PAD dataset, differences between the single runs were still high. Again, this could be due to the high collinearity in the dataset, and RFE might be better able to handle this kind of data. However, also RFE-chosen features were not consistent across cross-validation runs and no definite optimal number of features could be identified. Both these algorithms revealed few residues that would consistently be selected. However, the algorithms are also based on a regressor, which needs to be optimized in order to return good results, and before a feature selection would be able to return significant and consistent results.

Regression

All four algorithms (Ridge, KNN, GradientBoosting, and Random Forest) were able to predict scores for each data point, albeit with low accuracy. Maximum mean Pearson R values of parameter-optimized algorithms remained low, at values below 0.6. Different types of input data were tried, but neither of them could produce a significantly different result. Figure 3.8 shows the resulting predicted scores of one of the Gradient Boosting runs against the respective measured scores. Although some of the points are lying on the expected ideal line, there are still many points with a predicted score far from their measured score. The mean Pearson R-value of all Gradient Boosting runs of this set was 0.531. Generally, the results obtained with this dataset show less variation between the single runs than those of the PAD dataset (see Tables 4.12 to 4.14).

3.2.4 Embeddings

Structure Preparation and Homology Models

The nitrostyrene ligand that was transformed from the 4-OT crystal structure into one of the active site cavities of the AlphaFold Model of the 4-OT wildtype and further into all other active-site cavities of that structure was found to be positioned perfectly in only one of the cavities. Unsurprisingly, this cavity is the one that had been aligned to the cavity containing the ligand in the crystal structure. Though its five copies had been placed inside the active site cavity, the difference was apparent upon more thorough inspection of the structures. This also applies to the homology models that had been build based on this structure.

Point Cloud Generation

For 1220 homology models (also here, AAEs, like, for example, I2I, were still included in the dataset), 7258 cavities were created (5.9 cavities per variant). This was the only case in this work, where there were variants, for which no cavity was found. These cavities were filtered and only shaped cavities with a ligand coverage $>55\%$ were used for the embeddings. Although all the shaped cavities that were found were positioned reasonably inside the enzyme active site, only one cavity per structure showed such a high ligand coverage. Again, this is very likely due to the manual placement of the other five ligands.

t-SNE

A t-SNE dimensionality reduction was performed with the resulting embeddings, but also the result of this analysis did not correlate with the scores of the reactions. The first two dimensions of the t-SNE on the embeddings data are mapped against the scores measured for the tautomerization reaction in Figure 3.9.

Regression

Lastly, the data from the embeddings was also used as input for different regression algorithms, which, similar to previous experiments, were not able to predict meaningful scores. Mean Pearson r values of the prediction ranged between 0.17 and 0.45, with standard deviations between 0.039 and 0.106. Also here, the best results were obtained with the Random Forest regressor on the stereoselectivity dataset (mean Pearson R $= 0.447 \pm 0.049$). The result of one of these runs is shown in Figure 3.10. Using the embeddings dataset as input yielded results with mean Pearson R values with less variation between different results and estimators compared to the data derived directly from the Point Clouds (see Table 4.15).

3.2.5 Rational Analysis and Sequence Alignment

Marco Cespugli's inspection of the activity data and sequence alignment revealed a few interesting plausible hypotheses about the role of several amino acids. Figure 3.11 shows the position of these amino acids in the active site of 4-OT.

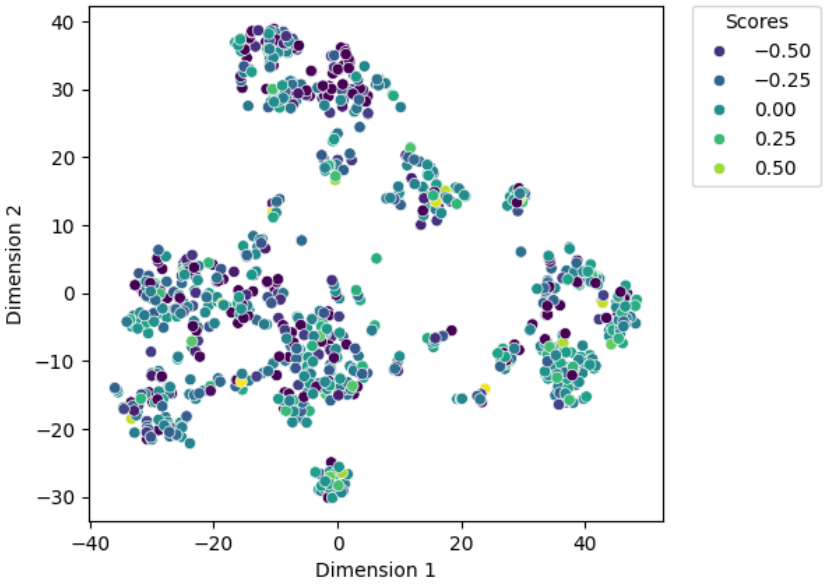

Figure 3.9: **t-SNE of the embeddings of the 4-OT dataset.** The first two dimensions are mapped against the measured tautomerization scores, and while multiple smaller groups of points are found, also here, no correlation is seen between the locations and the measured score of each data point.

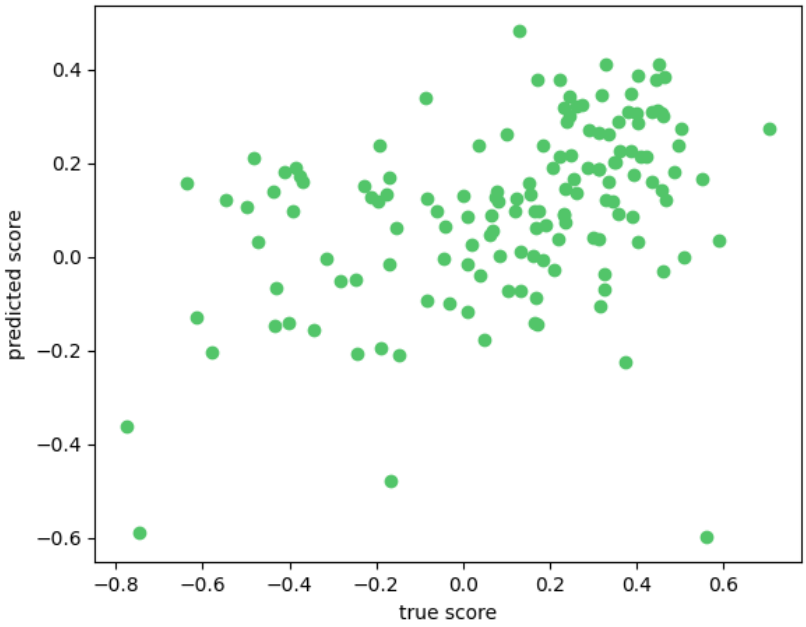

Figure 3.10: **Random Forest prediction of the stereoselectivity-scores of 4-OT based on the embeddings-data.** No high correlation could be obtained by the regressor for the embeddings data.

His6

Histidine 6 points towards the core of the hexamer. AAEs usually happen with polar amino acids, but the data analyzed here show an improvement of tautomerization and Michael addition of butanal to nitrostyrene upon introduction of Isoleucine, Leucine, or Methionine at this position. These AAEs also lead to a strong prevalence of the *2S3R* enantiomer of the Michael addition product. It is possible that these amino acid exchanges affect the pore water network, and might be beneficial for positioning of the aromatic substrates.

Arg11

As already pointed out, this residue has an important role in 4-OT's tautomerization mechanism (see chapter 1.2.2), which is also underlined by it's strong conservation (85%).

Leu31

Interestingly, an exchange of this residue to a lysine triples the tautomerization reaction, while for the two Michael addition reactions, most AAEs are detrimental. This might be because the residue interacts with the carboxylate group of phenylenolpyruvate.

Asp32

This residue is situated in a sharp turn rather distant from all active sites, and while proline is never observed at this position in nature, it triples tautomerization activity. However, on the other hand, Michael addition activity of butanal to nitrostyrene is strongly decreased in the D32P variant.

Ala33

This residue is found in the active site without outstanding conservation. Different AAEs can improve both, the tautomerization and the Michael addition reaction. Probably through interaction with the phosphoenolpyruvate carboxylate, positively charged residues improve the tautomerization reaction, while aspartate and glutamate show a positive effect for the two Michael additions, potentially through protonation of the aldehyde oxygen.

Arg39

This residue, which acts as general acid catalyst and lowers the pK_a of the active site (see chapter 1.2.2) shows no conservation, but AAEs at this position have no or a negative impact on all three reactions.

Ala46

This residue is very interesting, as it lies far away from relevant positions and is not conserved, but the introduction of a tyrosine residue at this position quadruples the tautomerization reaction rate.

Phe50

As pointed out in chapter 1.2.2, this residue is usually conserved aromatic, but leucine and valine at this position positively affect the tautomerization reaction and Michael addition of butanal to nitrostyrene. Aromatic residues at this position might force a wrong orientation of the aromatic substrates through π-π interactions, and therefore, hinder catalysis.

Gly54

This residue is highly conserved but Michael addition of acetaldehyde to nitrostyrene is strongly improved upon introduction of a tryptophane residue at this position, which might cause a reorientation of the loop and therefore a narrower active site, which is better suited for the small Michael donor.

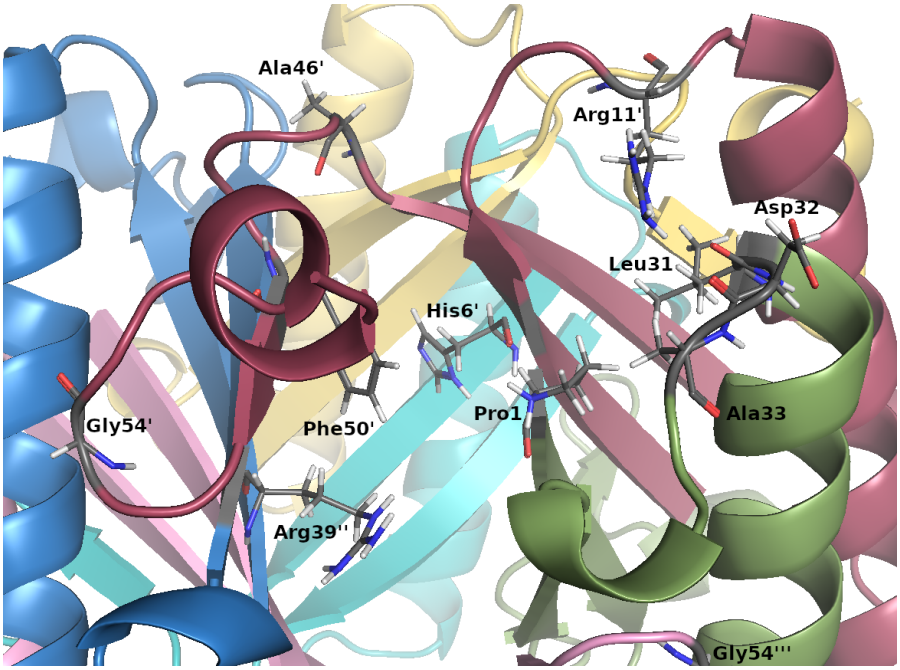

Figure 3.11: **Important catalytic residues according to rational analysis.** The positions of residues, which seem to have a notable contribution to 4-OT activities are shown here in the AlphFold Structure of the 4-OT wildtype.

4 Conclusion and Outlook

Two very different datasets were used here in order to try and predict enzyme activities based on their computationally determined physicochemical active site properties.
While the first dataset presented many problems that would supposedly not be present in the second dataset, no model could be found that was able to predict the activities of either of the two enzymes.
Apart from the various issues encountered with the activity data, the PAD dataset consists of very few data points compared to the large number of properties per variant, and therefore, an attempt was made to reduce this number in order to find a way to get meaningful results from this limited dataset. Often, laboratories have limited resources, that is in terms of staff, time, and equipment and therefore, being able to predict very active variants of an enzyme based on very little data would enable a more targeted search. However, from the results obtained here, it is not clear, whether the low predictive value of the model is caused by the activity data, the cavity data, or the model itself.
On the other hand, the 4-OT dataset is very detailed, as most of the variants had been produced, and their activities had been measured. But still, it was not possible to find a model and predict the activities of the variants. This suggests that the method tried here might not be capable of this task.
Of course, there are many other factors than just the enzyme's active site residues that contribute to catalysis and that have enough impact on enzyme activity to predict the latter without taking those factors into consideration. One of these factors is for sure the dynamics of the enzyme - opening and closing of the active site, getting substrate atoms closer together, etc - which however are way more complicated to compute than just the active site properties, and have therefore not been considered in this work. As found by Dr. Marco Cespugli, there are two amino acid residues (Asp32 and Ala46) that have a great impact on enzyme activities, but lie far away from the active site. Their role could be related to enzyme dynamics, but also an involvement in enzyme stability or induction of structural changes by the residue are conceivable.

In conclusion, the problem that had been presented here turned out to be more complicated than expected; at least in the two cases present here. The hypothesis that interactions between residues and the substrates in the active site mediate catalysis and that it should therefore be possible to predict an enzyme's performance based on it's active site properties could not be confirmed here. For both enzymes, the impacts of some residues on catalysis are known and can be explained through rational inspection of the sequence, but some remain hard to grasp.

With the growing importance of biocatalysis for industry and interest in optimizing enzymes

for various applications and by various means, the search for relationships between protein sequence and activity will become very important in the future and, with developments in computational biology, a solution to this problem is very likely to be found.

Bibliography

Amid, E., & Warmuth, M. K. (2019). Trimap: Large-scale dimensionality reduction using triplets. *CoRR*, *abs/1910.00204*. http://arxiv.org/abs/1910.00204

Bauer, K. K. F. (2024). Investigation of the substrate acceptance of decarboxylative enzymes.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Braun, M., Gruber, C. C., Krassnigg, A., Kummer, A., Lutz, S., Oberdorfer, G., Siirola, E., & Snajdrova, R. (2023). Accelerating biocatalysis discovery with machine learning: A paradigm shift in enzyme engineering, discovery, and design. *ACS Catalysis*, *13*(21), 14454–14469. https://doi.org/10.1021/acscatal.3c03417

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burks, E. A., Fleming, C. D., Mesecar, A. D., Whitman, C. P., & Pegan, S. D. (2010). Kinetic and structural characterization of a heterohexamer 4-oxalocrotonate tautomerase from chloroflexus aurantiacus j-10-fl: Implications for functional and structural diversity in the tautomerase superfamily, [PMID: 20465238]. *Biochemistry*, *49*(24), 5016–5027. https://doi.org/10.1021/bi100502z

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., ... Hassabis, D. (2022). Protein complex prediction with alphafold-multimer. *bioRxiv*. https://doi.org/10.1101/2021.10.04.463034

Frank, A., Eborall, W., Hyde, R., Hart, S., Turkenburg, J. P., & Grogan, G. (2012). Mutational analysis of phenolic acid decarboxylase from bacillus subtilis (bspad), which converts bio-derived phenolic acids to styrene derivatives. *Catal. Sci. Technol.*, *2*, 1568–1574. https://doi.org/10.1039/C2CY20015E

Gao, C. X., Dwyer, D., Zhu, Y., Smith, C. L., Du, L., Filia, K. M., Bayer, J., Menssink, J. M., Wang, T., Bergmeir, C., Wood, S., & Cotton, S. M. (2023). An overview of clustering methods with guidelines for application in mental health research. *Psychiatry Research*, *327*, 115265. https://doi.org/https://doi.org/10.1016/j.psychres.2023.115265

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422. https://doi.org/10.1023/A:1012487302797

Hendlich, M., Rippmann, F., & Barnickel, G. (1997). LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.*, *15*(6), 359–63, 389.

Hetmann, M., Langner, C., Durmaz, V., Cespugli, M., Köchl, K., Krassnigg, A., Blaschitz, K., Groiss, S., Loibner, M., Ruau, D., Zatloukal, K., Gruber, K., Steinkellner, G., & Gruber, C. C. (2023). Identification and validation of fusidic acid and flufenamic acid as inhibitors of sars-cov-2 replication using drugsolver cavitomix. *Scientific Reports*, *13*(1), 11783. https://doi.org/10.1038/s41598-023-39071-z

Jain, A. K. (2010). Data clustering: 50 years beyond k-means [Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)]. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/https://doi.org/10.1016/j.patrec.2009.09.011

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

McInnes, L., Healy, J., & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction. https://arxiv.org/abs/1802.03426

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). Colabfold: Making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Morley, K. L., Grosse, S., Leisch, H., & Lau, P. C. K. (2013). Antioxidant canolol production from a renewable feedstock via an engineered decarboxylase. *Green Chem.*, *15*, 3312–3317. https://doi.org/10.1039/C3GC40748A

Murel, P., Jacob, & Kavlakoglu, E. (2023). *What is ridge regression?* Retrieved August 21, 2024, from https://www.ibm.com/topics/ridge-regression

Myrtollari, K., Calderini, E., Kracher, D., Schöngaßner, T., Galušić, S., Slavica, A., Taden, A., Mokos, D., Schrüfer, A., Wirnsberger, G., Gruber, K., Daniel, B., & Kourist, R. (2024). Stability increase of phenolic acid decarboxylase by a combination of protein and solvent engineering unlocks applications at elevated temperatures. *ACS Sustainable Chemistry & Engineering*, *12*(9), 3575–3584. https://doi.org/10.1021/acssuschemeng.3c06513

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Poddar, H., Rahimi, M., Geertsema, E. M., Thunnissen, A.-M. W. H., & Poelarends, G. J. (2015). Evidence for the formation of an enamine species during aldol and michael-type addition reactions promiscuously catalyzed by 4-oxalocrotonate tautomerase. *ChemBioChem*, *16*(5), 738–741. https://doi.org/https://doi.org/10.1002/cbic.201402687

Poelarends, G. J., Veetil, V. P., & Whitman, C. P. (2008). The chemical versatility of the β–α–β fold: Catalytic promiscuity and divergent evolution in the tautomerase super-

family. *Cellular and Molecular Life Sciences*, *65*(22), 3606–3618. https://doi.org/10.1007/s00018-008-8285-x

Poelarends, G. J., & Whitman, C. P. (2004). Evolution of enzymatic activity in the tautomerase superfamily: Mechanistic and structural studies of the 1,3-dichloropropene catabolic enzymes [Mechanistic Enzymology]. *Bioorganic Chemistry*, *32*(5), 376–392. https://doi.org/https://doi.org/10.1016/j.bioorg.2004.05.006

Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion*, *27*, 111–125. https://doi.org/https://doi.org/10.1016/j.inffus.2015.06.005

Schubert, E., & Gertz, M. (2017). Intrinsic t-stochastic neighbor embedding for visualization and outlier detection. In C. Beecks, F. Borutta, P. Kröger, & T. Seidl (Eds.), *Similarity search and applications* (pp. 188–203). Springer International Publishing.

Scikit-learn user guide. (n.d.). Retrieved August 21, 2024, from https://scikit-learn.org/stable/user_guide.html

Selenium. (2024, March). Retrieved September 17, 2024, from https://www.selenium.dev/documentation/webdriver/

Sheng, X., Lind, M. E., & Himo, F. (2015). Theoretical study of the reaction mechanism of phenolic acid decarboxylase. *The FEBS Journal*, *282*(24), 4703–4713. https://doi.org/https://doi.org/10.1111/febs.13525

Uni-graz.catalophore.com. (2024). Retrieved September 17, 2024, from https://uni-graz.catalophore.com/CATALObase2/

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: A comparative review. *J Mach Learn Res*, *10*, 66–71.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

van der Meer, J.-Y., Poddar, H., Baas, B.-J., Miao, Y., Rahimi, M., Kunzendorf, A., van Merkerk, R., Tepper, P. G., Geertsema, E. M., Thunnissen, A.-M. W. H., Quax, W. J., & Poelarends, G. J. (2016). Using mutability landscapes of a promiscuous tautomerase to guide the engineering of enantioselective michaelases. *Nature Communications*, *7*(1), 10911. https://doi.org/10.1038/ncomms10911

What is boosting? (n.d.). Retrieved August 22, 2024, from https://www.ibm.com/topics/boosting

What is machine learning (ml)? (n.d.). Retrieved August 23, 2024, from https://www.ibm.com/topics/machine-learning

What is principal component analysis (pca)? (n.d.). Retrieved August 21, 2024, from https://www.ibm.com/topics/principal-component-analysis

Whitman, C. P. (2002). The 4-oxalocrotonate tautomerase family of enzymes: How nature makes new enzymes using a — structural motif. *Archives of Biochemistry and Biophysics*, *402*(1), 1–13. https://doi.org/https://doi.org/10.1016/S0003-9861(02)00052-8

Wu, W.-T., Li, Y.-J., Feng, A.-Z., Li, L., Huang, T., Xu, A.-D., & Lyu, J. (2021). Data mining in clinical big data: The frequently used databases, steps, and methodological models. *Military Medical Research*, *8*(1), 44. https://doi.org/10.1186/s40779-021-00338-z

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

Zandvoort, E., Baas, B.-J., Quax, W. J., & Poelarends, G. J. (2011). Systematic screening for catalytic promiscuity in 4-oxalocrotonate tautomerase: Enamine formation and aldolase activity. *ChemBioChem*, *12*(4), 602–609. https://doi.org/https://doi.org/10.1002/cbic.201000633

Appendix

Code Availability

The code used for all calculations outside of the Catalophore Platform is available on GitHub under https://github.com/ugSUBMARINE/cavityExploration.

AlphaFold Parameters

Table 4.1: Parameters for creation of AlphaFold structures:

Parameter	Value
num_relax	1
template_mode	none
msa_mode	mmseqs2_uniref_env
pair_mode	unpaired_paired
model_type auto	
num_recycles	3
recycle_early_stop_tolerance	auto
relax_max_iterations	200
paring_strategy	greedy
max_msa	auto
num_seeds	1
use_dropout	No

Parameters for Calculations on the Catalophore Platform

Homology Models

Table 4.2: Parameters for creation of the homology models of 4-OT:

Parameter	Value
Number of PSI-BLAST iterations	6
The maximum PSI-BLAST Evalue	0.5
Maximum number of templates to consider	5
Maximum number of ambiguous alignments to consider	5
Maximum oligomerization state	6
Maximum number of terminal loop residues	10
Use re-refined templates from PDB-Redo additionally	No
Homology model accuracy	Fast
Number of samples to try per loop	50
Use structure-based template profiles from RCSB	No

Cavity Procreation

Table 4.3: Parameters for Cavity Procreations:

Parameter	Value
Prepare structure	Yes
pH	7
Run CavFind	Yes
Min. cavity volume [A^3]	50/15/20 (1^{st} PAD/2^{nd} PAD/4-OT)
Max. cavity volume [A^3]	500/500/3700 (1^{st} PAD/2^{nd} PAD/4-OT)
Ligsite cutoff	5
Probe radius [A]	1.4
Grid Spacing	0.375
Softshell [A]	0.5
Remove hydrogens?	No
Remove hetgroups?	No
Remove waters?	No
Keep all alternates?	No
Calculate protein halos	No
Annotate cavities?	Yes
Detect chains?	Yes
Create hydrophicity cloud?	Yes
Create B-factor cloud?	Yes
Create accessibility cloud?	Yes
Collect statistics?	Yes
Create cavity picture?	Yes
Create histograms?	Yes
Create cavity preview image?	Yes
Detect ligands?	Yes
Create shaped cavities?	Yes
Rotate input structure randomly	No

Ligand Docking

Table 4.4: Parameters for Ligand Docking into PAD variants:

Parameter	Value
Run python analysis script?	Yes
Docking algorythm	VINA
ForceField	AMBER03
Docking box extension around cavity [A]	2
Docking runs	5
Clustering RMSD [A]	5
Rigid ligand?	No
Calculate cluster spreading?	Yes
Flexible receptor residues	None

Cavity Matching

Table 4.5: Parameters for Cavity Matching of PAD variants:

Parameter	Value
Maximum tries per entry	5
Query point cloud UUID	None
ICP iterations	100
ICP property steps	None
Timeout per match (s)	300
use sliding box	No
Property weights	
Cavity shape	0.01
Aromatic carbon point-cloud	1
Carbon point-cloud	1
Hydrogen-bond donor point-cloud	1
Non-hydrogen bonding nitrogens point-cloud	1
Nitrogen as H-bond acceptor point-cloud	1
Oxygen as H-bond acceptor point-cloud	1
Sulfur as H-bond acceptor point-cloud	1
Phosphor point-cloud	None
Desolvation point-cloud	0.1
Accessibility point-cloud	None
Electrostatics point-cloud	0.1
Hydrophobicity point-cloud	None
Flexibility point-cloud	None
Chains point-cloud	None
Sulfur point-cloud	None
Bromine point-cloud	None
Chlorine point-cloud	None
Fluorine point-cloud	None
Iodine point-cloud	None
To be found out	None

Classification Analyses

Classification Results

Table 4.6: Precision and Recall of the Classification per residue and cavity split. Means and Standard deviations were calculated over 3 splits and 2 repeats. Cavity splits are labelled according to section 2.5.1.

Reaction	Mean Precision			Mean Recall		
	CAc	FAc	SAc	CAc	FAc	SAc
catHalves	0.2±0.4	0.4±0.2	0.4±0.1	0.1±0.1	0.4±0.3	0.5±0.1
catRes	0.5±0.3	0.4±0.2	0.4±0.1	0.3±0.1	0.4±0.2	0.5±0.1
AAE	0.3±0.2	0.2±0.1	0.2±0.1	0.2±0.2	0.2±0.2	0.3±0.2

Table 4.7: Overall Precision and Recall of the Classification. Means and Standard deviations were calculated over 3 splits and 2 repeats. Cavity splits are labelled according to section 2.5.1.

Reaction	Precision	Recall
catHalves	0.367±0.120	0.367±0.120
catRes	0.400±0.077	0.400±0.077
AAE	0.255±0.081	0.255±0.081

Regression Analyses

GSCV

The following values were used for cross-validation, though not all of the options were allowed at all times to reduce running times of the scripts.

Table 4.8: Parameters optimized using GSCV:

Parameter	Value
n_estimators	100, 200, 300, 400, 500, 600, 700
criterion	squared_error, absolute_error
max_features	1, log2, sqrt
max_depth	1, 2, 3, 4, 5, 6, None
min_samples_split	2, 3, 4, 5, 6

Regression Results

PAD data were for each cavity split and ligand. Abbreviations for the cavity splits are as folows: catHalves: points of the cavity part close to the (assumed) position the ligand were divided in two groups based on their vicinity to either end of the cavity; catRes: cavity points close to the respective catalytically important residue; AAE: cavity points in the vicinity of positions 28, 79, and 92, respectively (see also chapter 2.5.1 and Figure 2.1). Differently, 4-OT data were predicted for four different reactions, which are abbreviated as follows: Tautomerization of phenylenolpyruvate to phenylpyruvate: Tautomerization, and Michael additions of acetaldehyde to *trans*-β-nitrostyrene: Acetaldehyde-M., Michael addition of butanal to *trans*-β-nitrostyrene: Butanal-M., and stereoselectivity of the latter Michael addition: Stereos.

Table 4.9: Mean pearson R values obtained with the different regressors for PAD reactions. Means and Standard deviations were calculated over 5 splits and 2 repeats.

Cavity Split	Ligand	Ridge	RF
catHalves	CAc	0.138±0.471	0.199±0.346
	FAc	0.154±0.389	0.009±0.187
	SAc	-0.085±0.263	0.014±0.354
catRes	CAc	-0.118±0.511	0.094±0.281
	FAc	0.322±0.341	0.214±0.291
	SAc	0.174±0.289	-0.001±0.298
AAE	CAc	0.104±0.404	0.127±0.428
	FAc	0.274±0.263	0.089±0.403
	SAc	0.297±0.238	0.401±0.244

Table 4.10: Mean pearson R values obtained with the different regressors for PAD reactions (clustered properties). Means and Standard deviations were calculated over 5 splits and 2 repeats.

Cavity Split	Ligand	Ridge	RF
catHalves	CAc	0.092±0.438	0.116±0.348
	FAc	0.104±0.370	-0.013±0.346
	SAc	-0.074±0.250	-0.194±0.275
catRes	CAc	-0.027±0.397	-0.037±0.333
	FAc	0.058±0.401	0.127±0.352
	SAc	0.200±0.267	0.064±0.209
AAE	CAc	0.038±0.278	-0.065±0.257
	FAc	0.262±0.270	0.130±0.229
	SAc	0.178±0.269	0.250±0.239

Table 4.11: Mean pearson R values obtained with the different regressors for PAD reactions (manually chosen properties). Means and Standard deviations were calculated over 5 splits and 2 repeats.

Cavity Split	Ligand	Ridge	RF
catHalves	CAc	0.302±0.329	0.166±0.456
	FAc	0.238±0.256	0.019±0.405
	SAc	0.233±0.295	0.041±0.320
catRes	CAc	-0.005±0.347	-0.019±0.260
	FAc	0.018±0.269	0.136±0.199
	SAc	0.036±0.348	0.034±0.270
AAE	CAc	-0.066±0.266	-0.007±0.314
	FAc	0.210±0.175	0.186±0.233
	SAc	0.055±0.279	0.134±0.257

Table 4.12: Mean pearson R values obtained with the different regressors for 4-OT reactions (Point Cloud Data). Means and Standard deviations were calculated over 5 splits and 2 repeats.

Reaction	KNN	Ridge	RF	GB
Tautomerization	-0.001±0.087	0.165±0.090	0.243±0.089	0.200±0.095
Acetaldehyde-M.	0.181±0.075	0.301±0.063	0.346±0.104	0.331±0.106
Butanal-M.	0.015±0.048	0.138±0.087	0.163±0.099	0.179±0.051
Stereos.	0.340±0.081	0.372±0.090	0.548±0.089	0.554±0.071

Table 4.13: Mean pearson R values obtained with the different regressors for 4-OT reactions (Point Cloud Data, best residues per reaction). Means and Standard deviations were calculated over 5 splits and 2 repeats.

Reaction	KNN	Ridge	RF	GB
Tautomerization	-0.021±0.060	0.110±0.067	0.191±0.056	0.166±0.071
Acetaldehyde-M.	0.241±0.076	0.326±0.054	0.306±0.070	0.263±0.086
Butanal-M.	-0.017±0.063	0.030±0.051	0.090±0.047	0.060±0.069
Stereos.	0.338±0.102	0.393±0.072	0.524±0.055	0.489±0.062

Table 4.14: Mean pearson R values obtained with the different regressors for 4-OT reactions (Point Cloud Data, clustered properties). Means and Standard deviations were calculated over 5 splits and 2 repeats.

Reaction	KNN	Ridge	RF	GB
Tautomerization	0.021±0.066	0.149±0.082	0.292±0.054	0.278±0.078
Acetaldehyde-M.	0.180±0.053	0.288±0.053	0.346±0.050	0.319±0.068
Butanal-M.	0.058±0.086	0.081±0.074	0.171±0.080	0.121±0.088
Stereos.	0.350±0.092	0.370±0.089	0.511±0.086	0.499±0.103

Table 4.15: Mean pearson R values obtained with the different regressors for 4-OT reactions (Embeddings). Means and Standard deviations were calculated over 5 splits and 2 repeats.

Reaction	KNN	Ridge	RF	GB
Tautomerization	0.198±0.067	0.221±0.066	0.240±0.088	0.226±0.071
Acetaldehyde-M	0.312±0.088	0.298±0.074	0.374±0.057	0.340±0.064
Butanal-M	0.220±0.079	0.173±0.039	0.247±0.049	0.250±0.071
Stereos.	0.420±0.097	0.288±0.099	0.447±0.049	0.429±0.106

List of Figures

List of Tables