Enzyme – Substrate interaction compatibility model formulation

# Problem Statement

Given *n*enzyme sequences that are predicted by *novoStoic* to perform a novel transformation of a substrate, rank the sequences in terms of their ***ability to catalyze the novel transformation.***

# Primary challenge

Unavailability of instances that represent lack of activity between enzyme and substrates restrict us from creating a training set with both positive and negative enzyme-substrate interaction instances, thereby preventing the use of traditional Machine Learning based discriminative classifiers.

# Sequence based modeling paradigms capable of solving the task

Although discriminative Machine Learning models in the supervised learning paradigm require both positive and negative training instances, there have been efforts to train a classifier using only positively labeled and unlabeled training examples [1]–[3], called PU learning. PU learning assumes that the unlabeled dataset contains both positive and negative training instances and tries to identify a set of reliable negative examples from the unlabeled set either automatically or through manual interference. For example, to ensure that the model can correctly discriminate between the two classes, the authors in [1] introduced a large number of irrelevant training examples which can be considered negative without any restrictions since they are not related to the problem they tried to solve. In our case however, such irrelevant examples may be hard to conjure.

A different approach to work with unlabeled or mislabeled training examples was proposed by Scholkopf et. al. [4], [5] where they adopted the discriminative classification algorithm, Support Vector Machines to the unsupervised learning domain and used it to detect outliers in their data. Although the algorithm does not require any labeled examples, the training set should contain both positive and negative data. Thus, along with the inclusion of active enzyme-substrate pair, we can also include random enzyme-substrate pair in our training set among which we hope that there are enough negative instances that would allow the model to detect an outlier enzyme-substrate pair.

While discriminative models such as SVMs try to learn the decision boundary between classes, generative models assume that the data comes from a probability distribution and tries to estimate that distribution [6]. For our purposes, generative models can be used to learn the probability distribution of substrates that lead to a certain enzyme. We could train a Generative Adversarial Network to predict encoded enzyme representations from numerical vector representation of substrates. Thereafter, we can use the encoded enzyme representation retrieved for a novel substrate to create a ranked list of enzymes by selecting those enzymes whose actual encoded representations are the most similar to the predicted encoded representation based on a distance metric such as Euclidean Distance.

Finally, among the unsupervised learning techniques, probability density estimation techniques such as the Kernel Density Estimation [7] can be used to learn the probability density function of our enzyme-substrate pair samples and given an unknown sample, we can use the estimated density function to predict how likely it is to observe that sample.

We believe the kernel density estimation might be the easiest to implement followed by outlier detection using unsupervised SVM algorithm. Please find below a brief sketch of the two methods.

# Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric method which can be used to estimate the probability density function of a multivariate random variable. Given a sample of independent and identically distributed observations, , of a random variable, the estimated density function at a point is:

Here, is the kernel function, is the bandwidth, a smoothing parameter, and is the number of observations. We could estimate the best kernel function and the ideal bandwidth required for our task using cross-validation. Compared to a Tanimoto index based search of relevant enzymes for a given substrate, KDE not only incorporates substrate encodings but also includes enzyme features and is capable of ranking enzyme-substrate pairs based on the density function estimate. Unlike other methods discussed in the above section that are capable of solving the task, KDE can work with only positive instances of enzyme-substrate interaction.

# One-Class SVM for outlier detection

The novelty detection using SVM algorithm, termed One-Class SVM, aims to estimate a function that is positive on a small subset of the dataset where most of the training data points belong and negative elsewhere. It uses a kernel function to map data into a feature space where it creates a separating hyperplane of maximal margin between the positive and negative regions. For a new instance, , the value of the function is determined based on which side of the hyperplane it belongs in the feature space. For our purposes of predicting propensity of enzyme-substrate activity, we could use the positive enzyme-substrate interaction instances along with manually created random enzyme-substrate instances and feed it to the One-Class SVM algorithm. Ideally, the algorithm will learn the positive instances as the primary subset where most of the data points would lie and define most of the random instances as outliers. Given a new enzyme-substrate instance from the test set, the algorithm will thus be able to predict whether it is most likely to belong to the positive instance subset or the random subset. One-Class SVMs are known to be highly sensitive to outliers and there are other outlier detection algorithms such as Isolation Forests, based on Random Forests, which have been shown to yield better performances and can be used to solve our task.

Diagram

Description automatically generated

Figure : The workflows of both Kernel Density Estimation (above) and Novelty Detection (below) using One-Class Support Vector Machines are illustrated here. Given information about interacting Enzyme-Substrate pairs, we can either estimate the density function of the concatenated feature space using Kernel Density Estimation or estimate the region where most of the datapoints lie using One Class SVMs. Henceforth, using the trained algorithms, we can score novel enzyme-substrate pairs according to their propensity of interaction.

# Structure based modeling paradigm capable of solving the task

Suppose each enzyme sequence needed to be ranked is and the corresponding EC number is .

## Workflow for scoring

**Step 1:** Perform a sequence alignment of each sequence against all sequences that belong to the EC number ECi that have at least one crystal structure in PDB. Choose a crystal structure with maximum sequence identity with as the template and construct a homology model for using Modeller [8].

Computational time ~ 6 hrs

**Step 2:** Extract the binding site for the homology model using a combination of active site annotation from Catalytic Site Atlas (CSA) [9] and *fpocket* [10]. The CSA will be first used to infer active site residues for the template and *fpocket* will be used to detect binding site pocket that overlaps with the active site residues. The assumption made here is that the binding pocket and active site residues within a given EC number are conserved. In absence of an inference from CSA or in cases where *fpocket* doesn’t detect any pockets that overlap with active sites inferred from CSA, all predictions from *fpocket* will be used.

Computational time ~ 1 hr

**Step 3:** Model the novel substrate inside the binding pocket(s) predicted from Step 2 above. For each binding pocket, perform docking search of the novel substrate using AutoDock Vina [11] followed by calculation of binding affinity (BA) scores. This will be followed by calculation of the *binding* *features* for each docking conformation – volume of binding site (V), shape complementarity of substrate in the binding site (SC) and the conformational flexibility (CF) [12] of active site residues calculated by rotamer scanning the residues by fixing the substrate conformation. Such features can be crucial to infer the compatibility of a novel substrate which may not necessarily have strong interactions, so is likely to be scored poorly if using binding affinity score alone. The values obtained for these features will be compared with those obtained for native substrates\* to infer the compatibility of the enzyme under consideration to catalyze the novel substrate.

Computational time ~ 1 hr

\*The *binding features* mentioned in Step 3 above will be computed for a diverse set of high-resolution crystal structures of native substrates bound to enzymes separately for each EC category to establish statistical distributions for each EC category.

## End result of workflow

Each enzyme sequence will have one of the following fates,

1. If no docking conformation with a negative binding affinity can be found for the novel substrate and the volume of ligand is greater than the volume of the binding site, then the sequence will be labeled as incompatible.
2. Else, the *binding affinity scores* along with the *binding features –* for all compatible sequences are output to the user for manual inspection.

An interactive application (similar to dG-predictor) can be built to perform all these calculations. Currently, no such tools exist to the best of our knowledge.

# References

[1] X. L. Li and B. Liu, “Learning from positive and unlabeled examples with different data distributions,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2005, vol. 3720 LNAI, doi: 10.1007/11564096\_24.

[2] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, “Building text classifiers using positive and unlabeled examples,” 2003, doi: 10.1109/icdm.2003.1250918.

[3] J. Bekker and J. Davis, “Learning from positive and unlabeled data: a survey,” *Mach. Learn.*, vol. 109, no. 4, 2020, doi: 10.1007/s10994-020-05877-5.

[4] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Piatt, “Support vector method for novelty detection,” 2000.

[5] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Comput.*, vol. 13, no. 7, 2001, doi: 10.1162/089976601750264965.

[6] G. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, “A comprehensive survey and analysis of generative models in machine learning,” *Comput. Sci. Rev.*, vol. 38, 2020, doi: 10.1016/j.cosrev.2020.100285.

[7] G. R. Terrell and D. W. Scott, “Variable Kernel Density Estimation,” *Ann. Stat.*, vol. 20, no. 3, 2007, doi: 10.1214/aos/1176348768.

[8] Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics. 2016 Jun 20;54:5.6.1-5.6.37. doi: 10.1002/cpbi.3. PMID: 27322406; PMCID: PMC5031415.

[9] Nicholas Furnham, Gemma L. Holliday, Tjaart A. P. de Beer, Julius O. B. Jacobsen, William R. Pearson, Janet M. Thornton, The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes, *Nucleic Acids Research*, Volume 42, Issue D1, 1 January 2014, Pages D485–D489

[10] Peter Schmidtke, Vincent Le Guilloux, Julien Maupetit, Pierre Tuffï¿½ry, fpocket: online tools for protein ensemble pocket detection and tracking, *Nucleic Acids Research*, Volume 38, Issue suppl\_2, 1 July 2010, Pages W582–W589

[11] Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009 Dec;30(16):2785-91. doi: 10.1002/jcc.21256. PMID: 19399780; PMCID: PMC2760638.

[12] Fogolari F, Maloku O, Dongmo Foumthuim CJ, Corazza A, Esposito G. PDB2ENTROPY and PDB2TRENT: Conformational and Translational-Rotational Entropy from Molecular Ensembles. J Chem Inf Model. 2018 Jul 23;58(7):1319-1324. doi: 10.1021/acs.jcim.8b00143. Epub 2018 Jul 6. PMID: 29897235.