7

CASE STUDIES

In this thesis, we present applications of the methodology of Chapter 3 and ingredients of Chapters 4, 5, and 6. In particular, we focus on two novel approaches to differential equation parameter identification and their inference, and Poisson process sensing. The latter is a problem for which we defined a whole treatment under RKHS assumption in Sec. 7.4.

In Sec. 7.1, we provide a high-impact application of the framework in protein design and some considerations in this context. We provide applications for estimating linear functionals in Sec. 7.3. Classically, one estimates the whole function, but we showcase that much less is needed for many applications and we can demonstrate improved complexity.

Contributions: Sec. 7.4 and application of learning linear functional in Sec. 7.3 form one of the main contributions of this thesis. These are based on my published works (Mutný and Krause 2020, 2021) and (Mutný and Krause 2022), respectively. Sec. 7.1 is based on a large collaboration with the Ward and Panke group. I was responsible for the data analysis and decision-making for this project. The writeup is in preparation (Vornholt et al. 2023).

7.1 ENZYME DESIGN

Biocatalysis and metabolic engineering provide eco-friendly methods for producing various important compounds, thereby holding the potential to revolutionize diverse industries. Nonetheless, significant enzyme modification is often necessary to develop an effective biocatalyst for specific uses. Implementing machine learning techniques to create a model that maps the connection between protein structure and function can enhance the efficiency of enzyme modification. This approach also boosts the chances of finding the best solution. Hence, the concept of machine

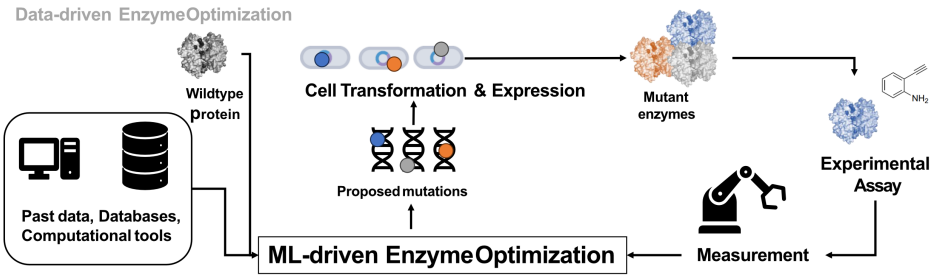

FIGURE 7.1: An overview of the enzyme design pipeline using modern experiment design.

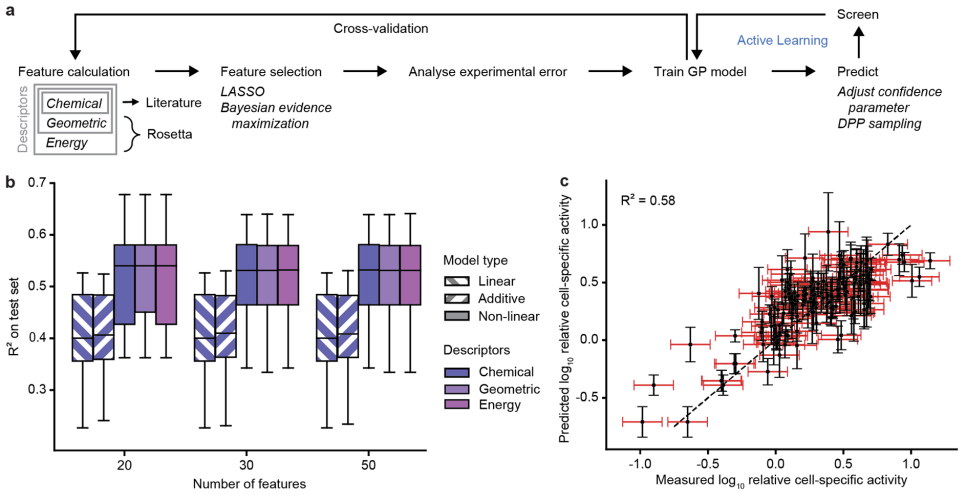

FIGURE 7.2: **a**, Overview of the machine learning pipeline. Model selection and model fitting were benchmarked using cross-validation. **b**, Performance of various models based on cross-validation analysis on different splits. The influence of feature size (x-axis), model type (fill pattern), and descriptors (colour) was investigated. Concerning the model types, we investigated linear, additive, non-linear, and non-additive non-linear models. **c**, Performance of the GP model using chemical descriptors and 20 features on an exemplary cross-validation split. The measurement uncertainty is displayed in red, while the uncertainty of the model is in black. The R2 value of this particular cross-validation split is provided.

learning-assisted directed evolution (MLDE) has gained substantial interest lately (Gelman et al. 2021; Romero, Krause, and Arnold 2013; Wu et al. 2019). Generally, MLDE begins with an initial screening phase, recording both sequence and function for various enzyme versions that are randomly selected. This sequence-function information is then utilized to train a model. This model subsequently directs further test phases in the experimental design process, as shown in Fig. 7.1.

An interesting use case for MLDE is artificial metalloenzymes (ArMs). These hybrid catalysts promise to significantly increase the number of reactions available in biocatalysis by equipping enzymes with the catalytic versatility of metals (Vornholt and Jeschek 2020). This can be achieved by repurposing natural enzymes (Bordeaux, Tyagi, and Fasan 2015; Kan et al. 2016), designing metal binding sites (Drienovská et al. 2017), or incorporating organometallic cofactors into proteins (Yang et al. 2018). Most ArMs have low initial activity, and extensive protein engineering is required to obtain biocatalysts suitable for real-world applications. Using this strategy, we aim to improve previously designed ArM for gold-catalyzed hydroamination (Vornholt et al. 2021).

INITIAL HIGH-THROUGHPUT SAMPLING We focus our optimization efforts on a sequence region of a gene encoding Streptavidin (Sav) amino-acids 111,112,118,119

and 121 in the canonical ordering. The previous study optimized the positions 112 and 121 and found an 8-fold improvement in the whole sequence space. Our collaborators at Basel performed the activity assay, sequenced 32 96-well plates using our NGS-based strategy, and successfully retrieved the relevant sequence information for 2,663 of 2,880 wells containing Sav mutants. After excluding variants with introduced stop codons and wells containing more than one variant, sequence-activity data for 2,164 clones remained. The library displayed high sequence diversity, with every amino acid appearing in every position. As there were previously recorded sequence-activity data for 400 Sav double mutants (S112X K121X) that are part of the same sequence space (Vornholt et al. 2021), we added these older data to the measurements of the triple and quadruple mutants obtained herein. Consequently, 2,992 data points covering 2,435 distinct ArM variants were available as training data for machine learning.

7.1.1 *Problem Statement and Modeling Choices*

To construct a model, we relied on RKHS models (Sec. 2.1) and especially their probabilistic counterparts, Gaussian processes (Rasmussen and Williams 2006). We comment on the relation to RKHS models in Sec. 2.1.3. The only difference to our exposition in the thesis is that instead of using the tools of Chapter 5, we use Bayesian methodology to determine Θ as a credible set. In this case, the kernel measures the similarity between different enzyme variants as a function of their sequence. Proper kernel selection is paramount to achieve good performance and sample efficiency – i.e., predict accurately with little data. Thus, we performed a benchmarking process and found that the Matérn kernel performed best in our case. We used evidence maximization for this purpose. This technique is extensively reviewed in Rasmussen and Williams (2006, Chapter 6). The Matérn kernel operated on design features that we refer to as descriptors. We considered features that reflect the chemical properties of amino acids (as in Wu et al. (2019)), as well as features that were extracted from Sav mutant structures predicted using the Rosetta software (Kellogg, Leaver-Fay, and Baker 2010). The latter include both geometric features (e.g., solvent-accessible surface area, number of hydrogen bonds, partial charge, dihedral angles, etc.) and energy terms. Note that the geometric and energy-based descriptors were compiled to be strict supersets of the chemical descriptors.

MODEL SELECTION Given a large number of features (125 chemical, 682 geometric, and 161 energy features), we sought to select a subset of highly predictive and parsimonious features to ensure data efficiency and eliminate redundancy amongst the features. Due to the non-linearity of the evidence maximization optimization, we first reduced the feature set using LASSO. More precisely, we fitted a linear model and then selected features with non-zero coefficients for automatic relevance detection using Bayesian evidence maximization with the non-linear model. This allowed us to decrease the initial pool of features and speed up the evidence maximization step, which required multiple optimization restarts to ensure an adequate maximum was achieved.

EXPERIMENTAL NOISE To understand the stochastic nature of experiments (in the spirit of Chapter 5) we estimated the experimental error by analysis of variants appearing multiple times in the screening. This revealed that the deviation of these replicates from the per-variant mean seems to follow a normal distribution (in log transformation). To estimate the standard deviation of the Gaussian distribution, we made the simplifying assumption that the variance of the measurement remains constant across the different rounds of our screening. In the first round, we encountered a total of 511 variants that had at least one repetition and used these values to determine a standard deviation for the first round. We repeated this analysis after each round. It turns out that the experimental error (in the log activity) was quite substantial at levels of 40% of activity for wild-type activity. As this experimental error is high, properly incorporating it into the decision framework should allow us to consider this.

INITIAL MODEL PERFORMANCE We trained different models with different descriptors on the initial data set and evaluated their performance using 15-fold cross-validation. We included a linear and an additive non-linear model based on chemical descriptors for comparison. The nonlinear additive model is restricted to adding a potentially non-linear effect of the individual descriptors on the activity additively. Interestingly, the chemical, geometric, and energy-based descriptors displayed a comparable performance, and a set of 20 features proved to be sufficient in all cases (Fig. 7.2). Notably, the linear and additive models performed considerably worse, confirming that non-linear and non-additive methods are required to capture the sequence-activity relationships in the data accurately.

7.1.2 *Experiment Design and Results*

Now, every experiment design starts by considering a utility. In this case, we know that our experimental budget is limited. Namely, we know that we can operate only in 3 rounds, where the first round has already been used for initial exploration. Hence, we decide to use different utilities in the second and third rounds. We call the second, *exploration* round and the third, *exploitation* round.

EXPLORATION ROUND In the second round, we performed an exploratory screening round to improve the model's accuracy and ability to generalize across the entire sequence space. We designed a new library consisting of 720 variants that maximized the D-design objective on the whole 3.2 million mutants space and hence should improve the predicted activity among all 3.2 million mutants. To optimize this objective, we used *greedy* algorithm. We can see that the average width of the confidence set has decreased in Fig. 7.3b).

EXPLOITATION ROUND After the exploration round, we incorporated the newly acquired data and set out to test whether we can discover very active ArMs. We designed a third library of 720 variants predicted to be of high activity and generated part of the predictions using our primary model based on chemical descriptors. We also used models based on the alternative descriptors that were trained on the

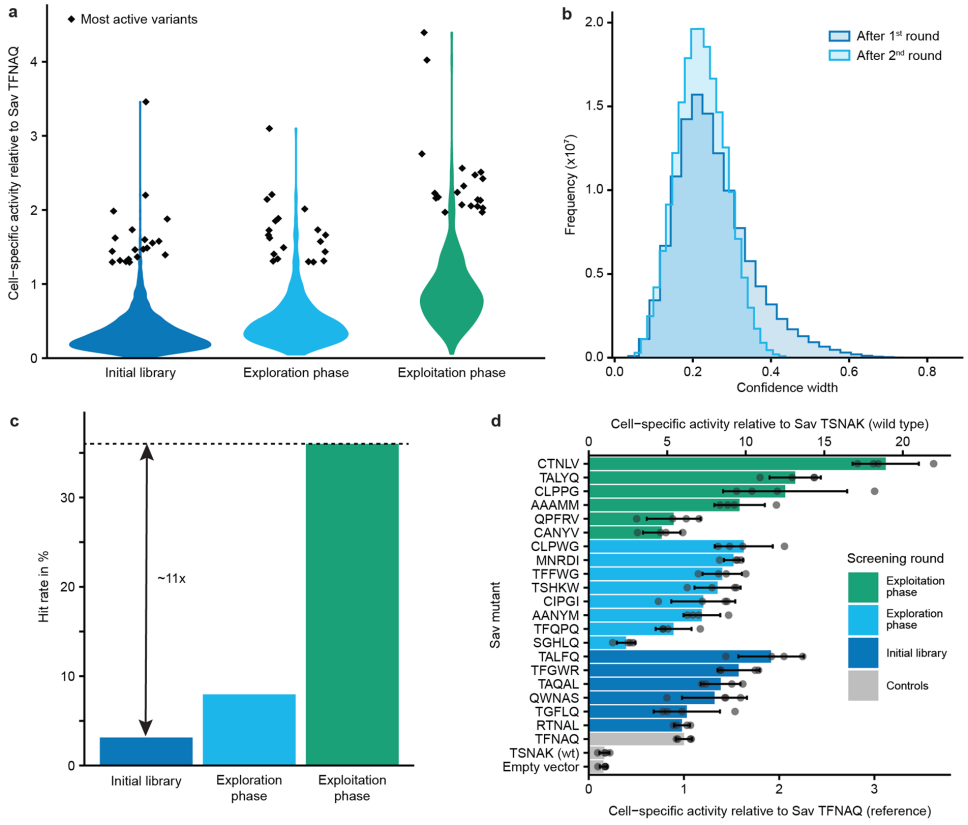

FIGURE 7.3: **a**, Activity distribution in the three screening rounds displayed as violin plots. The 20 most active variants in each round are shown as diamonds. Activity is displayed relative to the reference variant (Sav TFNAQ). **b**, Histograms of the standard deviations of predictions after the first and second rounds of experimentation. **c**, Hit rate in the three screening rounds.**d**, The five most active variants from each of the three screening rounds were tested again in four replicates. The five-letter codes denote the respective variants' amino acids in positions 111, 112, 118, 119, and 121.

complete data set (including the exploration round). Additionally, we employed a diversification step to avoid choosing only variants that are highly similar to each other. This provides a safeguard against inaccuracies in the top predictions and ensures we obtain variants with diverse properties besides activity (e.g. thermostability, solubility, or activity under alternative conditions). To this end, we used a notion of diversity known as determinantal point processes (Kulesza, Taskar, et al. 2012), which use the kernel to determine which variants are similar to each other (see Sec. 6.3.2). We obtained the designed library as an oligonucleotide pool and acquired experimental data for 468 distinct variants. Notably, this third library displayed a clear shift towards higher activities compared to the first two rounds, both in terms of average activity and top activities (Fig. 7.3). We further analyzed the hit rate in the screening rounds, which we define here as the fraction of ArM variants with higher activity than the reference variant, which is the most active variant identified in a previous study.

We chose to measure our performance based on the discovery of a hit. A hit is considered a variant whose fitness is higher than a previously discovered best variant (Vornholt et al. 2021). While in the initial round, only 3 % of variants were hits, this rate reached 36 % in the exploitation phase, an increase of approximately 11-fold (Fig. 7.3). This demonstrates that the models learned a meaningful representation of the activity landscape and can reliably identify active ArMs. The chemical descriptors in further analysis proved to work best but might indicate that the model based on chemical descriptors profited more from the exploration round, as the variants tested in this round were chosen to be informative based on this model. To reinforce our argument that optimized experiment design is better than random design, we created a simulated study in Fig. 7.4. This analysis indicates that acquiring training data by random sampling has diminishing returns. Approximately 40 % of the initial data set is sufficient to achieve a similar performance as a model trained on the initial data set. This might suggest that additional random screening rounds would not benefit much. In contrast, the model-guided exploration round, which consisted of only 20% additional variants, improved the hit rate in the subsequent exploitation round from 20 % to 48 %. This indicates that the data gathered in this round were substantially more informative than in the prior round. Thus, experiment design and model-guided exploration are powerful strategies for developing to minimize the experimental effort.

As the screening results mainly consisted of single measurements per variant and the presence of large experimental noise, our collaborators tested the most promising variants from all rounds again in four replicates each (Fig. 7.3). This revealed that the variant of Sav 111C 112T 118N 119L 121V (abbreviated Sav CTNLV) was the most active variant, reaching an 18-fold higher activity than the wild type (Sav TSNAK) and a three-fold higher cell-specific activity than the previously best-discovered variant.