



# Enhanced Sequence-Activity Mapping and Evolution of Artificial Metalloenzymes by Active Learning

## Journal Article

### Author(s):

Vornholt, Tobias; Mutný, Mojmír; [Schmidt, Gregor](#) ; Schellhaas, Christian; Tachibana, Ryo; [Panke, Sven](#) ; Ward, Thomas R.; Krause, Andreas; Jeschek, Markus

### Publication date:

2024-07-24

### Permanent link:

<https://doi.org/10.3929/ethz-b-000676410>

### Rights / license:

[Creative Commons Attribution 4.0 International](#)

### Originally published in:

ACS Central Science 10(7), <https://doi.org/10.1021/acscentsci.4c00258>

### Funding acknowledgement:

180544 - NCCR Catalysis (phase I) (SNF)

# Enhanced Sequence-Activity Mapping and Evolution of Artificial Metalloenzymes by Active Learning

Tobias Vornholt,<sup>#</sup> Mojmir Mutný,<sup>#</sup> Gregor W. Schmidt, Christian Schellhaas, Ryo Tachibana, Sven Panke, Thomas R. Ward,<sup>\*</sup> Andreas Krause,<sup>\*</sup> and Markus Jeschek<sup>\*</sup>



Cite This: *ACS Cent. Sci.* 2024, 10, 1357–1370



Read Online

ACCESS |



Metrics & More



Article Recommendations



Supporting Information

**ABSTRACT:** Tailored enzymes are crucial for the transition to a sustainable bioeconomy. However, enzyme engineering is laborious and failure-prone due to its reliance on serendipity. The efficiency and success rates of engineering campaigns may be improved by applying machine learning to map the sequence-activity landscape based on small experimental data sets. Yet, it often proves challenging to reliably model large sequence spaces while keeping the experimental effort tractable. To address this challenge, we present an integrated pipeline combining large-scale screening with active machine learning, which we applied to engineer an artificial metalloenzyme (ArM) catalyzing a new-to-nature hydroamination reaction. Combining lab automation and next-generation sequencing, we acquired sequence-activity data for several thousand ArM variants. We then used Gaussian process regression to model the activity landscape and guide further screening rounds. Critical characteristics of our pipeline include the cost-effective generation of information-rich data sets, the integration of an explorative round to improve the model's performance, and the inclusion of experimental noise. Our approach led to an order-of-magnitude boost in the hit rate while making efficient use of experimental resources. Search strategies like this should find broad utility in enzyme engineering and accelerate the development of novel biocatalysts.



## INTRODUCTION

Biocatalysis and metabolic engineering offer sustainable production routes for many compounds of interest and thus hold the potential to transform various industries. However, extensive enzyme engineering is typically required to obtain a suitable biocatalyst for a desired application. This is often a time-consuming, empirical process whose outcome is subject to chance, as classical methods are agnostic to the topology of the underlying sequence-activity landscape. Engineering strategies that incorporate machine learning to model this landscape could render enzyme engineering more efficient and increase the likelihood of identifying an optimal solution. Accordingly, machine learning-assisted directed evolution (MLDE) has attracted significant attention in recent years.<sup>1–3</sup>

In general, MLDE starts with an initial screening round in which both sequence and activity are recorded for a number of enzyme variants. These sequence-activity data are then used to train a machine learning model, with the objective of predicting the activity of untested variants directly from their sequence. If successful, such models can suggest variants that are likely to be highly active and thus support further screening rounds by *in silico* library design.<sup>1</sup> Further, the model can be iteratively updated with new data to improve its predictive performance, a strategy referred to as active learning. While several studies have demonstrated the general feasibility of such approaches,<sup>4–12</sup> there are still various challenges that

need to be addressed to maximize the success rate and efficiency of MLDE and enable its widespread implementation. This pertains to various aspects such as library design, experimental data acquisition, model development, and the strategy for sampling the sequence space.

With regard to library design, the crucial challenge is to create a library that is as information-dense as possible to allow for the development of accurate models while keeping the screening effort manageable. In the initial stages of model development, this calls for libraries that exhibit a high degree of sequence diversity to provide adequate information on the underlying sequence space, while at the same time containing a sufficient number of active mutants.<sup>13</sup> These requirements can be difficult to reconcile, as simultaneous randomization of multiple residues commonly results in a large fraction of inactive mutants, from which little to no meaningful information for model training can be extracted.

Once a library has been generated, it is often challenging to measure a sufficiently large set of sequence-activity data. In

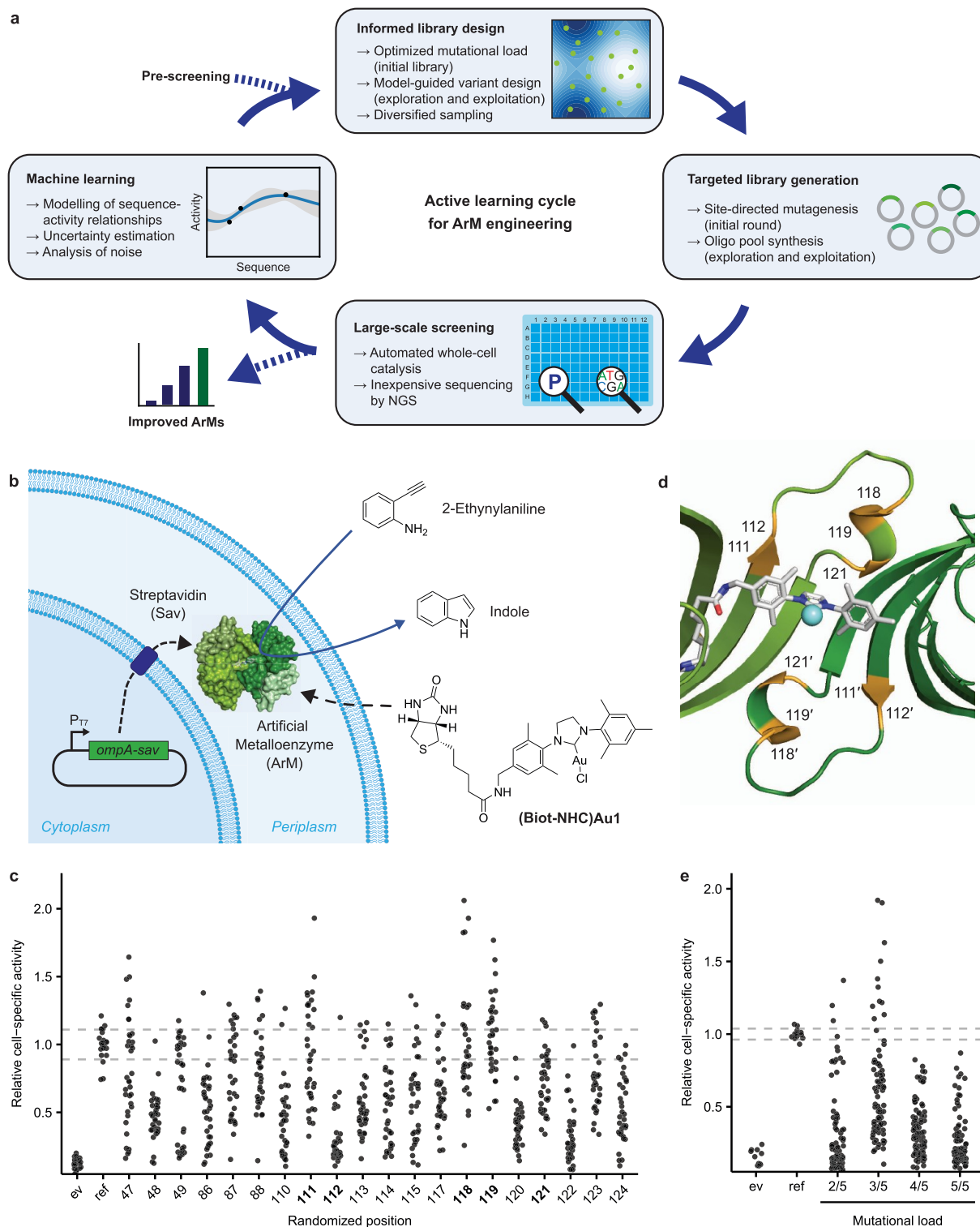
**Received:** February 15, 2024

**Revised:** April 22, 2024

**Accepted:** May 2, 2024

**Published:** May 22, 2024





**Figure 1.** Engineering strategy and library design for ArMs catalyzing hydroamination. **a.** Illustration of the active learning strategy for ArM engineering. An iterative process of library design, cloning, large-scale screening, and machine learning was used to model the sequence-activity landscape and identify improved ArMs. Crucial steps and considerations are highlighted and are explained in the main text. **b.** Illustration of whole-cell biocatalysis using an ArM in the periplasm of *E. coli*. Sav is exported to the periplasm by means of an N-terminal OmpA signal peptide, where it binds the biotinylated cofactor (Biot-NHC)Au1. The resulting ArM converts 2-ethynylaniline to indole in a new-to-nature hydroamination reaction. Indole can subsequently be quantified using a colorimetric assay. **c.** Single site-saturation mutagenesis to identify influential amino acid residues with respect to ArM activity. Starting from the reference variant Sav S112F K121Q, 20 residues in Sav were individually mutated using degenerate NDT codons. The activity of the resulting variants is displayed relative to the mean activity of the reference variant (“ref”). Dashed lines indicate one standard deviation around the mean activity of the reference variant, which was measured in triplicate in each 96-well plate. A strain lacking Sav, i.e., containing an empty vector (“ev”), was included as a control ( $n = 3$  per 96-well plate). The five positions selected for combinatorial randomization are highlighted in bold. Note that no improvement was expected at positions 112 and 121, as the reference variant had already been optimized with regard to these positions.<sup>40</sup> **d.** Residues selected for randomization (highlighted in orange) in a ribbon model of Sav harboring a

Figure 1. continued

metathesis catalyst (PDB SIRA). For clarity, only two biotin-binding sites of two opposing Sav monomers (a so-called functional dimer) are displayed. e. Effect of different multisite randomization strategies on the activity distribution of ArM libraries. Starting from the reference variant, either two, three, four or five residues among positions 111, 112, 118, 119, and 121 were randomized simultaneously. Hydroamination activity is displayed relative to the average activity of the reference variant ("ref",  $n = 3$  per 96-well plate) for 90 variants from each library. A strain containing an empty vector ("ev") was included as a control ( $n = 3$  per 96-well plate).

some cases, high-throughput assays such as fluorescence-activated cell sorting can be combined with deep sequencing to obtain very large data sets.<sup>14,15</sup> However, most enzymatic reactions of industrial relevance require more laborious analytical procedures to obtain a readout for activity. Moreover, the need to also obtain sequence information on all tested variants can lead to prohibitive costs if conventional Sanger sequencing is used. Consequently, most studies to date have relied on small data sets ( $10^1$ – $10^2$  variants).<sup>4–10</sup> While this has led to several successful demonstrations of MLDE, larger data sets are likely to lead to more accurate machine learning models and improve the chances of identifying variants with the desired properties,<sup>11</sup> particularly as the search space increases in size.

Beyond these experimental considerations, several critical decisions have to be made regarding the machine learning strategy. Prominent examples in this regard include the encoding strategy for the protein sequences and the choice of a suitable machine learning algorithm. Many encoding strategies have been suggested for creating a meaningful representation of protein variants, ranging from simple one-hot encoding and descriptors based on amino acid properties<sup>16–18</sup> to structure-based descriptors<sup>19,20</sup> and learned embeddings.<sup>21,22</sup> Similarly, various machine learning algorithms have been employed or suggested for MLDE, including linear regression,<sup>23–25</sup> Gaussian processes,<sup>4,7–9,25,26</sup> and neural networks.<sup>12</sup> While the best strategy depends on the data set and task at hand, Gaussian processes have repeatedly revealed their utility for active learning.<sup>8,9,25</sup>

Less attention has been devoted to other aspects of the machine learning process, such as the handling of experimental noise or the sampling strategy during ML-guided screening rounds, both of which are critical to the success and efficiency of MLDE. With regard to the sampling strategy, many studies have relied on a single training phase followed by greedy sampling of the top predictions of the resulting model. Due to inevitable biases in library generation and the limitations in generating sufficient sequence-activity data, this is unlikely to result in a comprehensive and accurate representation of the sequence-activity landscape. Consequently, such models may be "blind" for promising regions of the sequence space, leading to suboptimal outcomes such as low hit rates. Active learning strategies that improve the model in iterative cycles of experiments and machine learning may help to develop a better representation of the sequence-activity landscape, as these can converge to the optimal solution over time.<sup>27</sup> However, the aforementioned bottleneck in experimental data generation makes performing many iterations undesirable. Thus, resources invested into model improvement (i.e., exploration) must be carefully weighed against the focus on regions of the sequence space that are likely to contain active variants but might only comprise local optima (exploitation). In addition, activity may not be the only selection criterion during exploitation. Instead, it is often desirable to sample various potential optima to obtain a diverse set of variants,

which requires more elaborate approaches than simple greedy selection of top predictions.<sup>28</sup> Hence, smart sampling strategies for active learning are required to maximize the chances of success at a given experimental budget.

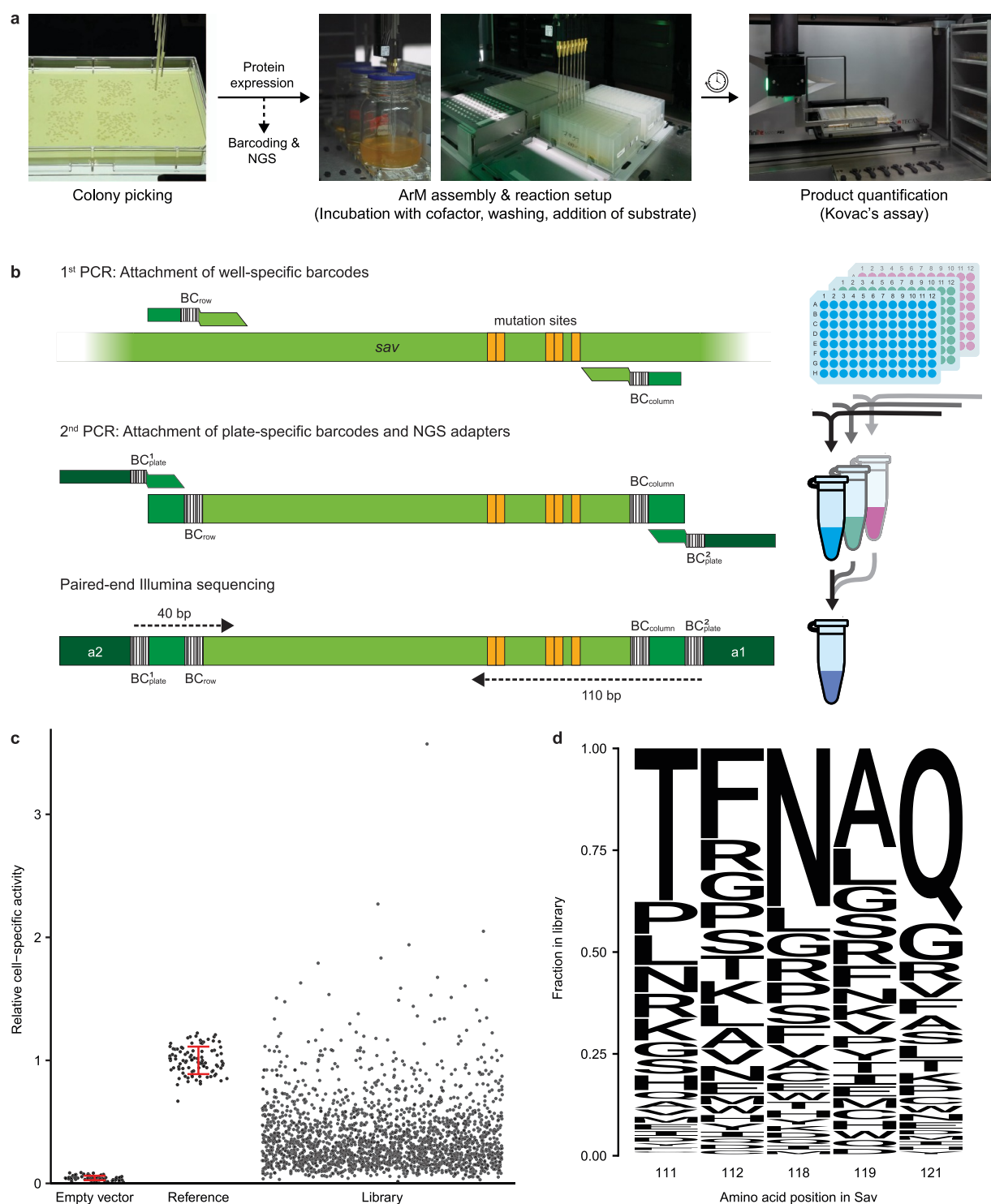
In this study, we introduce an integrated experimental and computational pipeline that addresses critical limitations in the MLDE of enzymes. Specifically, we combine informed library design with large-scale screening and a novel active machine-learning strategy. As an impactful testbed, we selected an artificial metalloenzyme (ArM) for gold-catalyzed hydroamination, a new-to-nature reaction for atom-economical C–N bond formation. We simultaneously engineered five crucial amino acid residues in this ArM, corresponding to a search space of 3 200 000 possible variants. To sample this space, we combined lab automation with a cost-efficient next-generation sequencing (NGS) strategy, which allowed us to acquire sequence-activity data on more than 2000 ArM variants. Furthermore, we developed a machine learning model based on Gaussian process regression that incorporates optimized descriptors and estimates of experimental noise to efficiently navigate the sequence space. Guided by the model's uncertainty estimates, we performed a second screening round focused on exploration and model refinement. Importantly, our results demonstrate that this targeted exploration substantially improved the model's performance. The optimized model reliably proposed highly active ArM variants in a final exploitation round, as illustrated by a 12-fold increased hit rate compared to the initial library.

## RESULTS

**Design of an Information-Dense ArM Library.** ArMs are hybrid catalysts that promise to significantly increase the number of reactions available in biocatalysis by equipping enzymes with the catalytic versatility of abiological transition metal cofactors.<sup>29</sup> ArMs have been created for a variety of natural and non-natural reactions,<sup>30–35</sup> and some have demonstrated catalytic prowess comparable to that of natural enzymes.<sup>36–39</sup> However, most ArMs initially display a low activity, and extensive protein engineering is required to identify catalytically proficient variants. This engineering is typically a labor-intensive and slow process. Therefore, ArMs represent an impactful yet challenging use case for MLDE.

A particularly versatile strategy for creating ArMs is to incorporate an organometallic cofactor into the tetrameric protein streptavidin (Sav) using a biotin moiety as the anchor. Using this approach, we have previously engineered an ArM for gold-catalyzed hydroamination by exhaustively screening a library of 400 Sav double mutants (Sav S112X K121X) using a whole-cell assay in 96-well plates.<sup>40</sup> While this represents an attractive starting point, extending the search space to more positions offers the opportunity to achieve further improvements, which will be crucial for adapting ArMs for real-world applications. However, exhaustive screening quickly becomes intractable in this case, and smart heuristics for the efficient





**Figure 2.** Large-scale acquisition of sequence-activity data for ArMs. **a.** Depiction of the critical automated steps in the screening workflow. Colony picking, ArM assembly, reaction setup, and product quantification were performed on a lab automation platform. The less labor-intensive protein expression protocol was performed manually. In parallel to the activity assay, samples of the starter cultures were processed further for NGS. **b.** PCR-based barcoding strategy for cost-effective sequencing of Sav variants in 96-well plates by NGS. First, the mutated region of the Sav gene is amplified using primers with row- ( $BC_{row}$ ) and column-specific ( $BC_{column}$ ) DNA barcodes. This step is performed in PCR plates using heat-treated bacterial cultures as templates. After pooling all samples from one plate, a second PCR is performed to add two plate-specific barcodes ( $BC_{plate}$ ) as well as adapters required for Illumina sequencing (a1 and a2). Subsequently, all samples are pooled and sequenced via paired-end reading to cover all barcodes and mutation sites. **c.** Cell-specific hydroamination activity of 2164 ArM variants from the initial library obtained by automated screening of 32 96-well plates. Only variants that were included for model training are displayed. Controls (empty vector and reference variant) are displayed with their standard deviation in red. **d.** Fraction of amino acids at the five randomized positions in Sav. Note that the amino acids of the reference variant (Sav 111T 112F 118N 119A 121Q, abbreviated Sav TFNAQ) are the most abundant, as the library was derived from this variant and contained at most four amino acid substitutions per variant.

exploration of the underlying sequence-activity landscape are essential.<sup>41</sup>

To navigate the sequence-activity landscape of the ArM, we devised an iterative active learning cycle involving library design, cloning, screening, and machine learning (Figure 1a). With regard to library design, the first step is to choose the target residues and a randomization scheme. To maximize the potential impact of the screening campaign, we aimed to find important positions in Sav besides the previously identified residues S112 and K121.<sup>40</sup> Thus, we individually randomized the 20 residues closest to the biotinylated gold cofactor in Sav S112F K121Q, which is the most active variant we had observed before<sup>40</sup> (referred to as “reference variant” herein). Randomization was performed using degenerate NDT (N = A, C, G or T; D = A, G or T) codons, which encode 12 amino acids covering all chemical classes of amino acids, a strategy that has revealed high success rates at a reduced screening effort.<sup>40</sup> Subsequently, we measured hydroamination activity using our previously established protocol relying on periplasmic catalysis in *Escherichia coli* (Figure 1b).<sup>40</sup> We tested 36 clones per randomized position to achieve a statistical library coverage of approximately 95%.<sup>42</sup> As expected, most variants displayed reduced activity compared to the reference variant (Figure 1c). Notably, positions 111, 118, and 119 revealed the highest potential for improvement upon mutagenesis, with several variants outperforming the reference variant. Consequently, we selected these positions for further engineering. In addition, we chose to also randomize positions 112 and 121 again, as our observations had indicated that epistatic effects play an important role in highly active ArM mutants.<sup>40</sup>

Next, we sought to create a combinatorial library of the five selected positions (111, 112, 118, 119, and 121, Figure 1d), which, upon full randomization, corresponds to a search space of  $20^5 = 3\,200\,000$  variants. This greatly exceeds the capacity of typical activity assays and well plate-based screenings. Thus, navigating the underlying sequence-activity landscape represents a significant challenge. In order to model this space for MLDE, it is crucial to design a library that offers a good coverage of the targeted sequence space and at the same time maintains a sufficient proportion of active variants.<sup>13</sup> While simultaneous randomization of all five residues would fulfill the first criterion, we anticipated that the high mutational load would likely lead to a large fraction of inactive variants. This would not only diminish the chances of identifying improved variants but also, importantly, would be uninformative for machine learning. Upon initial tests, we indeed observed a marked drop in the activity distribution when randomizing more than three of the five positions simultaneously (Figure 1e). Accordingly, we set out to construct a library with three to four mutations distributed across the five target residues as a good compromise between high sequence-diversity and sufficient residual activity. In other words, the constructed library covers all five target positions, but individual variants contain at most four amino acid substitutions relative to the reference variant Sav S112F K121Q, which served as the parent of the library (Figure S1 of the Supporting Information). This was achieved by site-directed mutagenesis PCR using various sets of primers containing degenerate NNK (K = G or T) codons at different positions and subsequent mixing of the resulting sublibraries (see Methods).

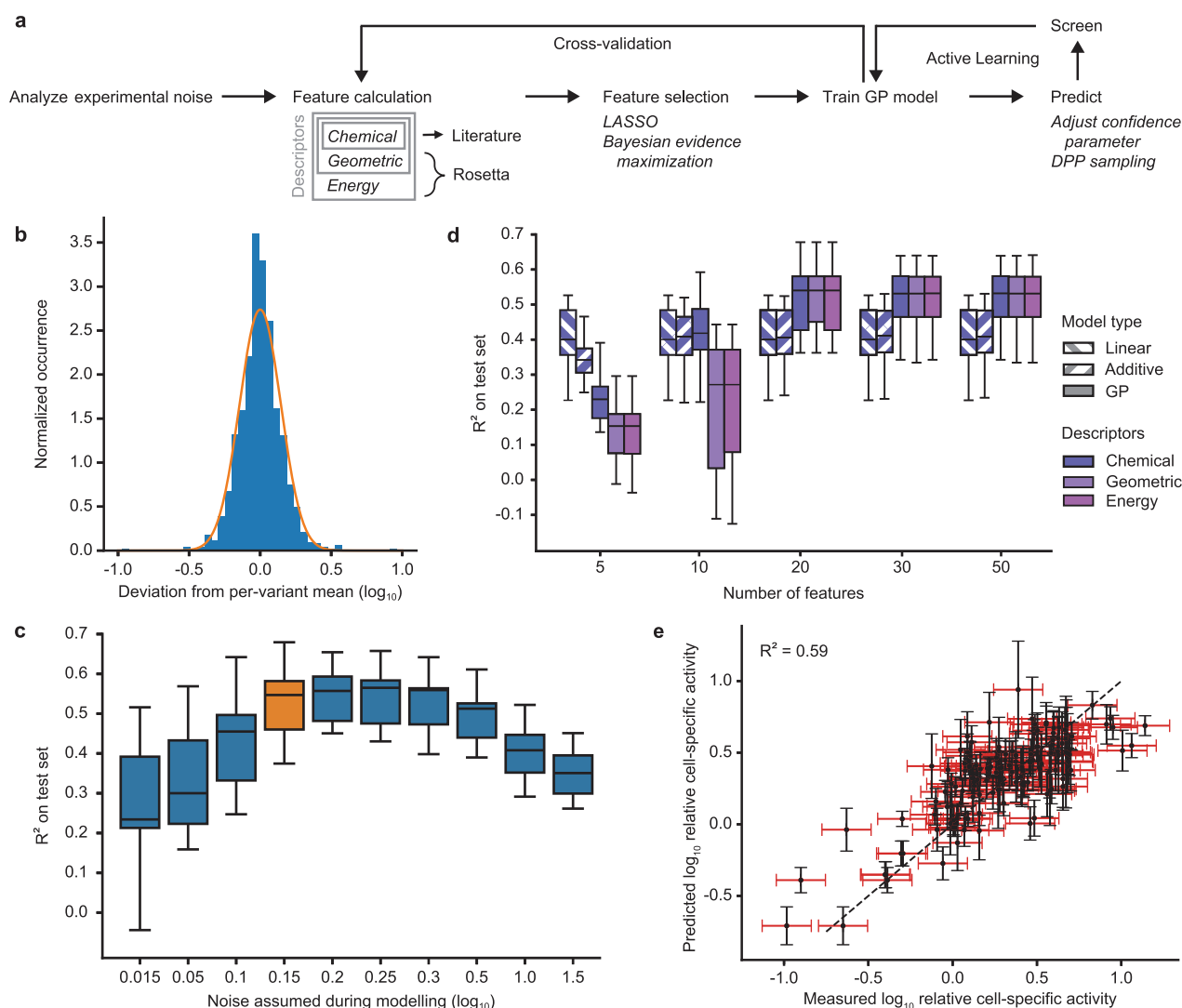
**Large-Scale Acquisition of Sequence-Activity Data.** Our previously established whole-cell screening protocol for

ArMs relied on periplasmic Sav expression, ArM assembly, and catalysis in 96-well plate format. By combining this protocol with conventional Sanger sequencing, we were able to obtain sequence-activity data for a few hundred variants.<sup>40</sup> Although this platform was more flexible and simpler than comparable screening strategies involving protein purification, it still required considerable manual labor, particularly for product quantification. Additionally, when larger data sets are required, Sanger sequencing rapidly leads to prohibitively high sequencing costs. To facilitate the generation of larger data sets for MLDE, we thus sought to minimize manual intervention in the activity assay and develop more cost-efficient means of obtaining the sequence information for each functionally characterized variant.

First, we automated all steps in the assay protocol that are labor-intensive (and thus limiting in terms of throughput) or critical for reproducibility. Specifically, we made use of a Tecan EVO 200 platform for all steps from colony picking to product quantification, with the exception of Sav expression in 96-deep well plates, which only requires a small number of pipetting steps (Figure 2a). The most important addition to our previous semiautomated pipeline<sup>40</sup> is the photometric quantification of the product indole. While this is a laborious procedure when carried out manually, the automated version simplifies screenings and proved to be very reproducible (Figure S2). As the robotic platform can handle up to eight 96-well plates at the same time, it greatly accelerates the acquisition of large data sets.

Besides the activity assay, another critical barrier to obtaining sufficiently large sets of sequence-activity data can be the cost of sequencing. Obtaining the sequences of several thousand protein variants by Sanger sequencing typically costs more than USD 10 000, which is prohibitive for most academic laboratories. In principle, the cost per variant can be reduced significantly by relying on NGS, which quickly becomes more cost-efficient than Sanger sequencing as the library size increases. However, in NGS all variants are sequenced in bulk, which means a method to retroactively link each sequence to the corresponding activity measurement is required. Previously, the use of DNA barcodes has been suggested to enable NGS of protein variants distributed across 96-well plates.<sup>43–45</sup> Building on these strategies, we established a two-step PCR protocol for the barcoding of Sav variants that is compatible with the Illumina NGS platforms (Figure 2b). In the first step, which is carried out in 96-well plates, the randomized region of the Sav gene is amplified using primers that append a well-specific barcode combination as well as constant regions to the ends of the PCR products. This is achieved using eight forward (representing the plate's rows) and 12 reverse primers (representing the columns). For simplicity, heat-treated samples of bacterial cultures serve as templates, avoiding the need for laborious and costly plasmid purification.

Subsequently, PCR products are pooled by plate, and each pool is gel-purified and used as a template for a second PCR. In this step, primers binding to the previously added terminal constant regions are used for amplification. These primers contain overhangs to append plate-specific barcodes as well as the adapters required for NGS. Through the combination of well- (1<sup>st</sup> step) and plate-specific (2<sup>nd</sup> step) barcodes, it is possible to sequence thousands of variants from multiple plates in a single, low-cost NGS run and to assign the obtained sequences to the corresponding activity value obtained in the



**Figure 3.** Development of the initial GP model. **a.** Overview of the machine learning pipeline. Initially, the standard deviation of the activity measurements was estimated to account for experimental noise. Subsequently, three feature sets were calculated and reduced sets were obtained by applying LASSO and Bayesian evidence maximization. The resulting descriptors were then used to train GP models. Model selection and model fitting were benchmarked using cross-validation. Ultimately, the GP model can be used to navigate the sequence space in active learning cycles. **b.** Histogram of the deviation between replicates in the initial library. The distribution of residuals can be conservatively approximated by a normal distribution with a specific variance (orange). **c.** Influence of the noise estimate on the predictive performance of the resulting GP model. The value chosen based on Figure 3b is highlighted in orange. The models used here were based on chemical descriptors with 20 features (see Figure 3d) and were evaluated using 15-fold cross-validation. The box plots display the 25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentile with whiskers denoting the 1.5-fold interquartile range. **d.** Influence of feature number ( $x$ -axis), model type (fill pattern), and descriptors (color) on the performance of machine learning models analyzed by 15-fold cross-validation. The box plots display the 25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentile with whiskers denoting the 1.5-fold interquartile range. A comparison of linear models based on different descriptors can be found in Figure S8. **e.** Performance of the GP model using chemical descriptors and 20 features on an exemplary cross-validation split. The measurement uncertainty (one standard deviation) is displayed in red, while the uncertainty of the model is in black. The  $R^2$  value of this particular cross-validation split is displayed.

functional assay. In our specific case, paired-end sequencing of 40 bp from one end and 110 bp from the other end of the final PCR product was sufficient to read all well- and plate-specific barcodes as well as the five mutation sites in the Sav gene at a high read coverage (average of >100-fold per variant) and low cost (see Discussion).

Relying on the combination of automated activity assay and NGS, we screened 32 96-well plates containing variants from the aforementioned library of Sav. As each plate contained six controls (empty vector and reference variant in triplicate), this amounts to a total of 2880 variants. Excluding mutants that failed to grow, we obtained activity data on 2790 variants. Most of these displayed an intermediate activity between the

background level of cells lacking Sav (empty vector) and the reference variant Sav S112F K121Q (Figure 2c). Notably, approximately 3% of all mutants were more active than the reference. Using the NGS-based strategy, we retrieved the sequences for 2663 out of 2880 wells containing Sav mutants. After excluding variants with nonsense mutations and wells containing more than one variant, sequence-activity data for 2164 clones were obtained, of which 2035 were distinct variants. Notably, for variants appearing in multiple wells, the deviation between these replicate activity measurements was generally low, corroborating the high robustness of the assay (Figure S3). Importantly, the library displayed a high degree of sequence diversity, with every amino acid appearing in every

position (Figure 2d) and an average Hamming distance of 4.3 between the mutants. Note that the amino acids of the reference variant were the most abundant in each position, as we did not randomize all five positions simultaneously. Thus, the library exhibited a high degree of variability both in terms of activity distribution (including a low fraction of inactive variants) as well as sequence diversity. This indicated that the aforementioned design goals for the library were met, providing a promising data basis for modeling the sequence-activity landscape by machine learning.

As we had previously recorded sequence-activity data for 400 Sav double mutants (S112X K121X) that are part of the same sequence space,<sup>40</sup> we added these older data to the measurements obtained herein. As a result, a total of 2992 data points covering 2435 distinct ArM variants were available as initial training data for machine learning.

**Development of an Initial Machine Learning Model of ArM Activity.** To construct a model that can reliably predict the activity of untested ArM variants and guide further screening rounds, we relied on Gaussian process (GP) regression.<sup>46</sup> This machine learning technique can capture highly nonlinear relationships and has the distinct advantage of being probabilistic, which means that it predicts a probability distribution rather than a point estimate, and thus provides an estimate for the confidence of each prediction. This feature can not only help users assess the uncertainty of individual predictions, but also is ideally suited for active learning strategies. In this scenario, the model's uncertainty estimates can be used to guide subsequent screening rounds toward uncertain regions of sequence space with the goal of improving the model (i.e., exploration), before suggesting highly active variants in later rounds (i.e., exploitation).

GPs are characterized by a mean and a covariance function, which is commonly referred to as kernel. In our case, as we operate on the space of protein sequences, the kernel measures the similarity between different ArM variants. Since the selection of a suitable kernel is of paramount importance for good performance and sample efficiency (i.e., predicting accurately with little data), we performed a benchmarking process and found that the nonlinear Matérn kernel<sup>46</sup> performed best in our case (see Methods).

Moreover, our model development pipeline included steps to account for experimental noise and to select suitable descriptors (Figure 3a). Considering the inherent noise in biological experiments during modeling is crucial to ensure that decisions are not influenced by random fluctuations. To distinguish the genuine signal from these fluctuations, it is necessary to define a probabilistic model for data generation, known as the likelihood. This step involves specifying the likelihood and its parameters, which is essential for applying Bayes' theorem to calculate the posterior distribution (see Methods). To elucidate the form of the likelihood, we relied on the variants appearing multiple times in the screening. This revealed that the deviation of these replicates from the per-variant mean closely follows a log-normal distribution, which can be viewed as a conservative estimate of the experimental noise in the data (Figure 3b). Considering the log-transformed values, this implies a Gaussian likelihood. Next, we used the replicate measurements to determine a standard deviation, which is a key element in defining the data likelihood. We made the simplifying assumption that the variance of the measurement remains constant across the different ArM variants and repeated this analysis after each round of

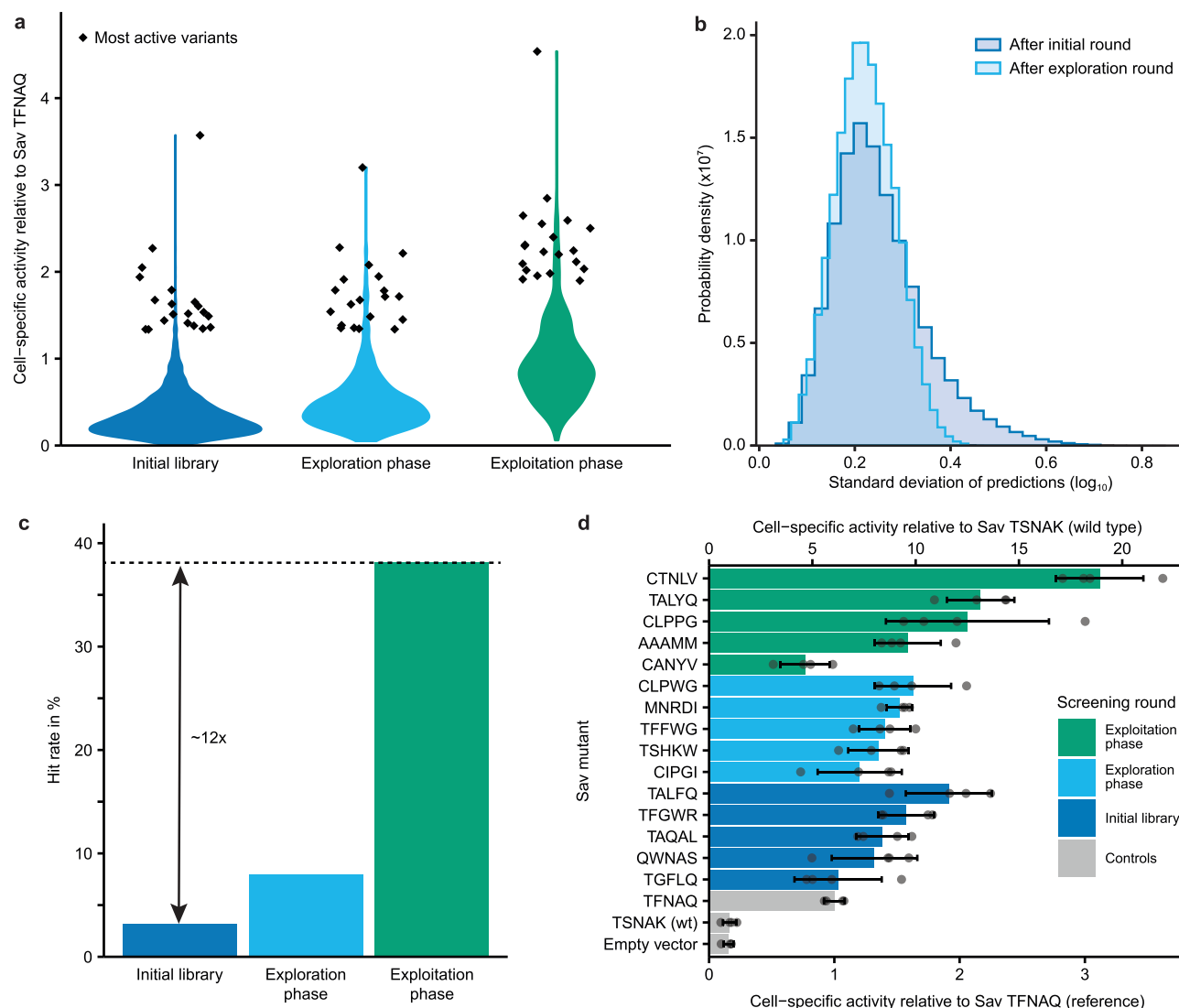
screening. As illustrated in Figure 3c, under- or overestimating the experimental noise leads to a drastically reduced performance of the resulting model, likely due to overfitting to noise in the data. In contrast, the procedure applied here results in a robust performance in the face of noisy data.

With regard to the descriptors that represent the ArM variants during training, we considered features that reflect chemical properties of amino acids<sup>11</sup> as well as features that were extracted from Sav mutant structures predicted with the Rosetta software.<sup>47</sup> The latter included 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 descriptors were compiled to be strict supersets of the chemical descriptors (i.e., they also included the chemical descriptors), and similarly the energy-based descriptors are strict supersets of the geometric descriptors. Given the large number of features (125 chemical, 682 geometric, and 161 energy features), we sought to select subsets that are parsimonious while still highly predictive to ensure data efficiency and eliminate redundancy. To this end, we relied on Bayesian evidence maximization (see Methods). Due to the nonlinearity of the optimization challenge, we first reduced the feature sets using LASSO, which performed best in a benchmarking test (Figure S4). More precisely, we fitted a linear model and selected features with nonzero coefficients for automatic relevance detection using Bayesian evidence maximization with a Gaussian process. This allowed us to reduce the initial pool of features to 20–100 and speed up the evidence maximization step, which required multiple optimization restarts to ensure that an adequate maximum was achieved.

Finally, we trained GP models using the different reduced feature sets on the available sequence-activity data and evaluated model performance using 15-fold cross-validation. For comparison, we included a linear and an additive, nonlinear model based on chemical descriptors. The latter is restricted to treating potentially nonlinear effects on the activity additively and is therefore not capable of modeling epistatic effects. Notably, the linear and additive models performed considerably worse than the GP models (Figure 3d), confirming that advanced methods such as GP models are required to accurately capture the sequence-activity relationships in the data. 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. The most influential features based on automatic relevance detection are listed in Table S1 (see Figure S5 for an analysis of their influence).

As computationally expensive structural calculations are required to generate the geometric and energy-based features and no clear benefit over models relying only on chemical descriptors was observable, we chose to continue with the subset of 20 chemical features as our primary encoding strategy for further modeling. The resulting model displayed a good predictive performance, with a median  $R^2$  of 0.54 based on 15-fold cross-validation (see Figure 3e and Figure S6 for exemplary validation splits). While leaving room for improvement, this degree of correlation has previously been shown to be suitable for guiding directed evolution campaigns.<sup>11</sup> Moreover, the median Spearman correlation of 0.68 demonstrates that the relative ranking of variants was largely reproduced by the model (Figure S7), which is important for confident selection of high-activity variants.



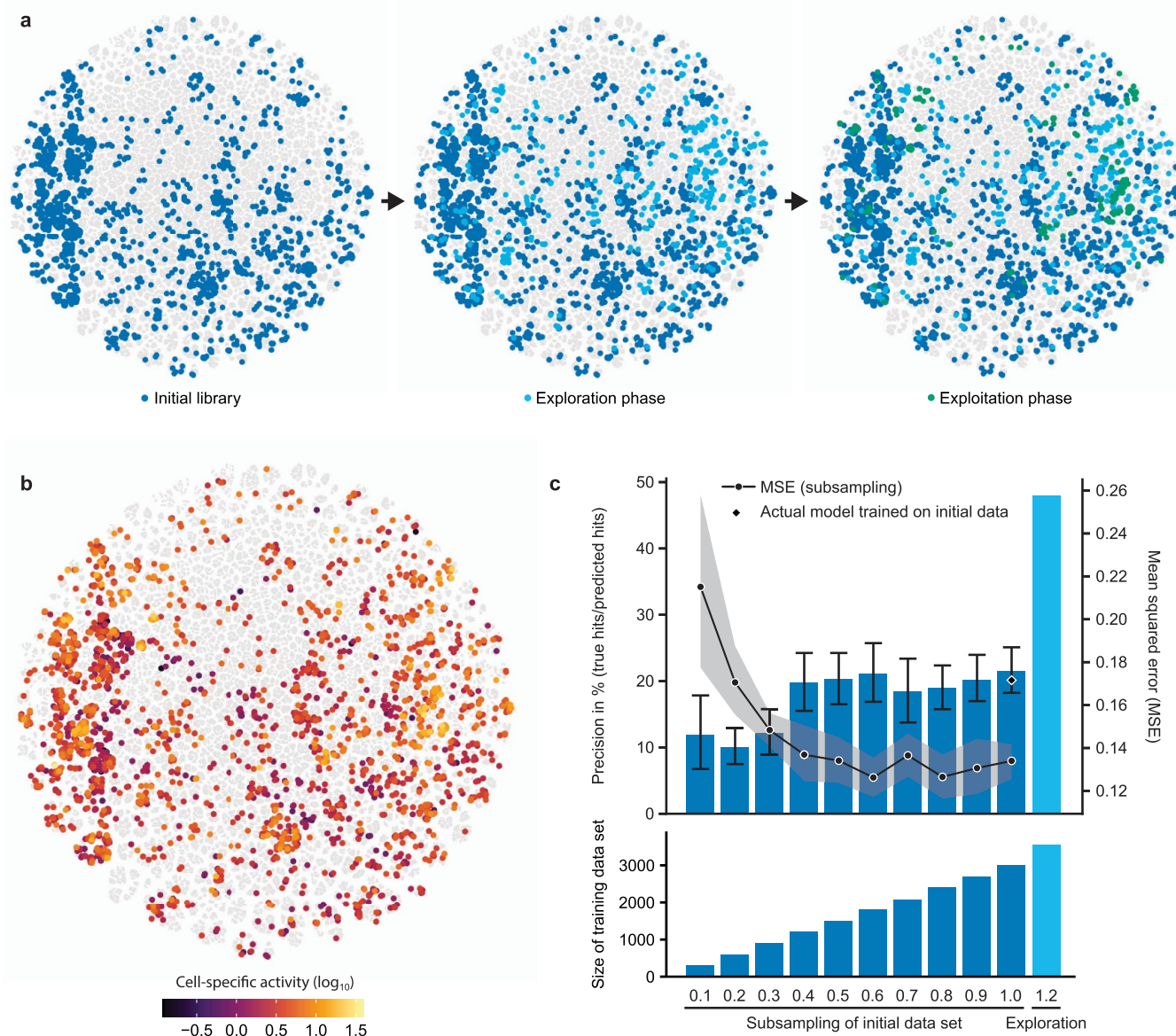


**Figure 4.** ArM engineering by means of active learning. **a.** Activity distributions in the three screening rounds displayed as violin plots. The 20 most active variants in each round are depicted as diamonds. Activity is displayed relative to the reference variant (Sav TFNAQ). **b.** Normalized histograms of the standard deviations of predictions across all 3.2 million variants after the first and second round of screening. **c.** Hit rate in the three screening rounds. Here, any variant with a higher cell-specific activity than the reference variant is considered a hit. The hit rate represents the fraction of hits among all variants screened in the respective round. Note that the hit rate in the initial library was calculated based on the triple and quadruple mutants, excluding the double mutants that had been tested previously.<sup>40</sup> In the third round, chimeric variants that were not part of the computationally designed library were excluded to provide a better analysis of the models' performance. **d.** The five most active variants from each screening round were tested again in four replicates. The five-letter codes denote the amino acids in positions 111, 112, 118, 119, and 121 for the respective variants.

**Model Refinement by Active Learning.** The aforementioned performance parameters indicate that the initial GP model can predict ArM activity with reasonable accuracy. However, due to the vast sequence space, the random sampling from this space during the generation of training data, as well as inevitable biases in experimental library construction, it is likely that this initial model will not generalize well across the entire sequence-activity landscape. Consequently, it may be “blind” for certain underexplored regions containing highly active ArMs. Therefore, we performed a second, exploratory screening round with the goal of improving the model's accuracy and ability to generalize across the entire sequence space. To this end, we designed a new library consisting of 720 variants that were primarily selected to be “informative”. Specifically, we utilized the uncertainty estimates of the GP model and selected the variants with the highest uncertainty in

the predicted activity among all 3.2 million mutants.<sup>48,49</sup> This selection was performed in an iterative manner, meaning the uncertainty was recalculated every time after selecting a single variant (see [Methods](#)).

We generated these variants based on a pool of oligonucleotides obtained through commercial synthesis on arrays, a method that allows for the cost-efficient construction of large and targeted libraries<sup>50</sup> and is therefore highly useful for active learning with large batch sizes. After cloning the oligonucleotides into the Sav expression plasmid, we screened the resulting exploration library relying on the automated pipeline in combination with NGS as described above. This exploratory round yielded sequence-activity data on 465 additional variants. It should be noted that this library also contained chimeric variants with amino acid combinations that were not planned in the computational design, likely due to



**Figure 5.** Enhanced sequence-activity mapping through active learning. a. t-SNE visualization of the sequence space. ArM variants that were tested in the three screening rounds are highlighted in different colors. To generate this visualization, all 3.2 million mutants were considered, and a uniform subsample of untested variants was plotted in gray. The similarity metric used was derived from the GP model (see [Methods](#) for details). b. t-SNE visualization of the sequence space with color encoding the activity of experimentally tested variants. The clustering is identical to that in [Figure 5a](#). c. Precision in identifying hits and mean squared error (MSE) of predictions as a function of the size of the training data set. The dark-blue bars in the upper graph indicate the average precision of models that were trained on different fractions of the initial data set (screening round 1). The diamond at 1.0 represents the precision of the model used to inform experiments. The light-blue bar on the right represents the model refined by model-guided exploration (screening round 2). Note that the precision is not identical to the experimentally determined hit rate (see [Methods](#)). The lower graph depicts the size of the data sets used to train the respective models.

PCR-mediated recombination between variants.<sup>51,52</sup> While unintended, these additional variants can also be used to augment the machine learning model and were therefore included for training. If desired, chimera formation can be minimized by optimizing the PCR conditions.<sup>51,52</sup>

The exploration library displayed a similar activity distribution as the initial training data ([Figure 4a](#)), which is in line with the focus on informative instead of active variants. Importantly, these new data led to a decrease in the standard deviation of the predictions, most prominently for variants that had previously exhibited a high uncertainty ([Figure 4b](#)). While this observation alone is not a proof of increased accuracy, it hints toward an improved representation of previously

underexplored regions of the sequence space, which we examined in more detail in subsequent analyses (see below).

**Active Learning Increases the Efficiency of Directed Evolution.** Following model refinement in the exploration round, we set out to test whether our model-guided approach can indeed aid in the discovery of active ArMs. With this goal in mind, we designed a third library of 720 variants predicted to be of high activity. Additionally, we employed an in silico diversification step to avoid choosing only variants with highly similar sequences. This provides a safeguard against inaccuracies in the top predictions and increases the likelihood of obtaining 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 (DPPs),<sup>48,49</sup> which use the GP kernel to determine which variants are similar to each other (see [Methods](#) and [Figure S9a](#)). In short, this approach treats the descriptors of the Sav variants as vectors in Euclidian space and attempts to select a set of vectors that are as orthogonal to each other as possible. We applied this process to a set of variants with the highest predicted activity to obtain a subset of active and yet sequence-diverse variants. This led to a more diverse set of variants compared to a simple greedy selection of the variants with the highest predicted activity as assessed by three different metrics of diversity ([Figure S9b](#)). Note that this procedure was not required in the exploration round, as the iterative selection of informative variants naturally leads to a diverse set of mutants.

As described for the exploration round, we obtained the designed library based on an oligonucleotide pool and acquired experimental data for 349 distinct variants. Gratifyingly, this third library displayed a clear shift toward higher activities compared to the first two rounds, both in terms of the average as well as the top activities ([Figure 4a](#)). 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.<sup>40</sup> While only 3% of the initial library were hits, this rate reached 38% in the exploitation phase, amounting to an approximately 12-fold increase ([Figure 4c](#)). This demonstrates that the model acquired a meaningful representation of the activity landscape and can reliably predict active ArMs.

To confirm the results from the different screening rounds, which were performed in single measurements, we tested the most promising variants from all three rounds again in four replicates ([Figure 4d](#)). This revealed that Sav 111C 112T 118N 119L 121 V (abbreviated Sav CTNLV) was the most active variant, reaching an 18-fold higher cell-specific hydroamination activity than the wild type (Sav TSNAK) and a 3-fold higher cell-specific activity than the reference variant (Sav TFNAQ). In addition, we purified the most active variants from our whole-cell screening to test whether they also display an increased total turnover number *in vitro*, which was the case for five of the seven variants tested ([Figure S10](#)). As observed before,<sup>40</sup> the ranking of the variants changed *in vitro*, which can be expected due to the different reaction environments and varying expression levels in the periplasmic screening.

Notably, the Sav CTNLV mutant does not retain the S112F K121Q mutations that were found to be optimal in the previous double mutant screening.<sup>40</sup> Likewise, all other variants evaluated in the validation experiment ([Figure 4d](#)) retain neither or only one of these two mutations. This highlights the importance of epistatic effects, which can only be adequately considered through combinatorial library designs and nonadditive models. Strikingly, several highly active variants contain a cysteine at position 111, which seems counterintuitive as cysteine has been repeatedly shown to have a pronounced inhibitory effect on gold-catalyzed hydroamination.<sup>53</sup> However, residue 111 is pointed away from the metal, presumably preventing the thiol from interfering with catalysis. Notably, the beneficial impact of this mutation was not obvious from the initial data set, but became increasingly apparent in subsequent rounds. This indicates that active learning can traverse the mutational space more broadly than

alternative methods and enable the identification of counterintuitive effects on activity.

To further corroborate this hypothesis, we performed more detailed analyses to investigate whether the active learning strategy with a model-guided exploration round indeed led to a better representation of the available sequence space. We visualized the sequence space (using t-SNE<sup>54</sup> on the kernel matrix, see [Methods](#)) to analyze how the tested variants are distributed across this space ([Figure 5a, b](#)). While care must be taken when interpreting such low-dimensional projections, this analysis indicates that the initial library did indeed not cover the sequence space uniformly. The subsequent exploration round filled in several of the “gaps” in accordance with the design goal of this phase. The exploitation phase focused on a number of regions of high activity, indicating that the selection criteria of high activity and sequence diversity were met. The emergence of multiple clusters of active variants is compatible with the notion of a “rugged” activity landscape with many local optima. Such landscapes can be challenging to navigate using classical methodologies, which frequently follow a single “uphill” trajectory. In contrast, the GP model developed here acquires a holistic understanding of the entire space of 3.2 million ArM variants and allows us to sample various potential optima, increasing the chances of finding suitable variants.

Lastly, we sought to quantify the effect of the applied sampling strategy in relation to the size of the training data set. A crucial question in this regard is whether the active learning strategy suggested here provides a significant benefit over a comparable increase in the size of the training data set by random sampling of variants. To investigate this, we trained models on different fractions of the initial data set using the same model development pipeline as before. As a proxy for an experimentally determined hit rate, we analyzed the models’ precision in identifying hits among the variants tested in the exploitation phase (i.e., the percentage of true hits among variants predicted to be hits). As illustrated in [Figure 5c](#), this analysis indicates that acquiring training data by random sampling is accompanied by strong diminishing returns: Approximately 40% of the initial data set size (equivalent to ~1200 data points) is sufficient to achieve a similar performance (in terms of precision and mean squared error (MSE)) as a model trained on the entire initial data set (~3000 data points). This suggests that additional random screening rounds of similar size would not have led to noteworthy improvements of the model. In contrast, the model-guided exploration round, which consisted of only 564 additional data points (an increase of less than 20% in data volume), improved the precision in identifying hits from ~20% to 48%. This increase is significantly beyond any improvement that can be anticipated due to the mere increase in data volume, emphasizing the fact that this round was substantially more informative than random sampling. This confirms the validity of the suggested active learning and model-guided exploration strategies, pointing to a high potential for enhancing MLDE campaigns while at the same time minimizing the experimental effort.

## DISCUSSION

MLDE is a highly promising strategy for engineering enzymes and other proteins.<sup>1–3,11</sup> However, the success and efficiency of such engineering campaigns hinges on the ability to generate sufficiently large and informative data sets, the use of smart



sampling strategies, and the choice of suitable machine learning techniques that optimally leverage the resulting data.

Many studies on MLDE have relied on small data sets<sup>4–10</sup> and a single training phase,<sup>4,5,10,55,56</sup> which may be attributed to experimental limitations. This bears the risk that the resulting models do not accurately represent the sequence space, and thus are likely to leave significant potential hidden within this space untapped. Here, we applied lab automation and NGS to acquire large data sets in a simple and cost-efficient manner, and directed our sampling to the most informative data by means of advanced active learning techniques.

Lab automation greatly increases the throughput of screenings and is, at the same time, highly adaptable to various reactions and target proteins. In this study, we performed some experimental steps manually, but a fully automated workflow could also be implemented. Similarly, the computational pipeline is largely automated, and thus it is conceivable to conduct ArM engineering with minimal human intervention, as was recently demonstrated for the thermostability of a natural enzyme.<sup>57</sup> Importantly, recent developments such as academic biofoundries and cloud laboratories are making such approaches more widely accessible.<sup>58</sup>

The NGS strategy employed here enables the sequencing of thousands of protein variants for the cost of a small Illumina run and PCR reagents. The former is available for a few hundred dollars (e.g., MiSeq Nano, yielding approximately 1 million reads) and will likely continue to get cheaper. If combined with other samples and run on an instrument with a large capacity, the prorated costs may even be in the range of a few dollars. Regarding the PCR reagents, primer synthesis costs are low as only 20 primers are required to address all 96 positions in a well plate. Similarly, the use of two plate barcodes means that 12 primers for the second PCR are sufficient to distinguish 36 plates. Thus, the required number of primers is lower than in alternative barcoding strategies,<sup>45</sup> leading to improved scalability. Nonetheless, other methods may be advantageous in specific cases (e.g., when several target genes need to be sequenced). Overall, this workflow enables sequencing at a cost of less than one cent per variant.

Combined, automation and NGS are ideally suited to generate large data sets for MLDE. At the same time, it is also crucial to design information-dense libraries to maximize the efficiency of experimental screening rounds. In the initial round, we achieved this by optimizing the mutational load in the library, which is a straightforward and broadly applicable strategy. Alternatively, zero-shot methods, for example based on  $\Delta\Delta G$  calculations,<sup>13</sup> can be applied as well. In subsequent rounds, library design can be guided by the machine learning model. While it may seem attractive to apply an exploitation-focused strategy to quickly identify active variants, we hypothesized that a model-guided exploration round could substantially improve the predictive performance and thus increase the chances of identifying suitable variants in a subsequent round. Indeed, we observed that the exploration round improved the model's ability to identify active variants far beyond what would be expected due to the increase in data volume alone. This demonstrates that active learning is a highly effective and efficient strategy for developing accurate models of sequence-activity landscapes. Moreover, the separation into exploration and exploitation phases provides a transparent and practical solution to the exploration-exploitation dilemma, as it allows for a clear and plannable

resource allocation. In addition, our study introduces DPP sampling as a strategy for diversifying the selection of active variants, which increases the robustness of MLDE to possible model inaccuracies and may be beneficial with regard to secondary properties beyond activity.

Active learning with large batch sizes, as employed here, may be most attractive when navigating large, rugged sequence-activity landscapes, which is challenging using conventional methodologies. If the screening throughput is limited, for example because of costly reagents or slow analytical procedures, smaller batch sizes can also be used with our methodology. In this case, additional rounds could be performed to increase the predictive performance and the chance of identifying promising variants.

In terms of the machine learning approach, this study corroborates that Gaussian process regression is an attractive choice for MLDE, particularly when strong epistatic effects are present in the sequence-activity landscape. Moreover, it is well-suited for active learning strategies, as the uncertainty quantification is computationally simple, which constitutes an advantage over alternative methods such as deep learning. Our results demonstrate that simple and computationally efficient descriptors are sufficient for nontrivial improvements to engineering campaigns, which is in line with other literature on the subject.<sup>59,60</sup> Nonetheless, it might be possible to further boost the predictive performance, for example by employing improved structure prediction algorithms or descriptors from modern protein language models.<sup>61,62</sup> Lastly, our results highlight that accurately accounting for experimental noise is crucial during model development, an aspect that has frequently been neglected.<sup>63</sup>

The application of these strategies to the engineering of ArMs for gold-catalyzed hydroamination led to the identification of a variant with 18-fold higher cell-specific activity than the wild type. Compared to our previous screening of double mutants,<sup>40</sup> extending the search space to five positions led to a 3-fold improvement. Further rounds of active learning could potentially lead to the discovery of even more active variants. Moreover, the strategies developed here could be used to target additional positions. In this case, minor modifications to the established methods may be required. Most importantly, when engineering more than approximately seven residues simultaneously, the computational search for the most informative or most active variants needs to be restricted (e.g., based on Hamming distance to a parent variant), as exhaustive calculations become impossible due to the exponential increase in the number of possible amino acid combinations. It should be noted that this ArM is likely a challenging engineering target due to the relatively exposed location of the cofactor in Sav. Therefore, applying this engineering strategy to alternative scaffolds with a more shielded active site might enable larger improvements.<sup>64</sup>

Currently, artificial (metallo)enzymes are typically limited by their rather modest activity. Thus, the field could profit greatly from advanced machine learning-guided engineering strategies, as demonstrated here. Similarly, the active learning approach described here could be applied to tailor natural enzymes for industrial applications, or to engineer other proteins such as antibodies, biosensors, or transporters.



## ■ ASSOCIATED CONTENT

### Data Availability Statement

The data and code created as part of this study are available under <https://github.com/lasgroup/ml-protein-design-sav-gold>.

### SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acscentsci.4c00258>.

Materials and methods, Figures S1–S10 (including additional validation of the pipeline), and Tables S1–S15 (including primer sequences and calculated features) (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Authors

**Thomas R. Ward** – National Centre of Competence in Research (NCCR) Molecular Systems Engineering, 4056 Basel, Switzerland; Department of Chemistry, University of Basel, 4058 Basel, Switzerland; [orcid.org/0000-0001-8602-5468](https://orcid.org/0000-0001-8602-5468); Email: [thomas.ward@unibas.ch](mailto:thomas.ward@unibas.ch)

**Andreas Krause** – Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland; Email: [krausea@ethz.ch](mailto:krausea@ethz.ch)

**Markus Jeschek** – Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland; Institute of Microbiology, University of Regensburg, 93053 Regensburg, Germany; Email: [markus.jeschek@ur.de](mailto:markus.jeschek@ur.de)

### Authors

**Tobias Vornholt** – Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland; National Centre of Competence in Research (NCCR) Molecular Systems Engineering, 4056 Basel, Switzerland; [orcid.org/0000-0001-9700-2384](https://orcid.org/0000-0001-9700-2384)

**Mojmír Mutný** – Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland

**Gregor W. Schmidt** – Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland

**Christian Schellhaas** – Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland; [orcid.org/0000-0001-7367-1620](https://orcid.org/0000-0001-7367-1620)

**Ryo Tachibana** – Department of Chemistry, University of Basel, 4058 Basel, Switzerland; [orcid.org/0000-0002-7229-3370](https://orcid.org/0000-0002-7229-3370)

**Sven Panke** – Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland; National Centre of Competence in Research (NCCR) Molecular Systems Engineering, 4056 Basel, Switzerland

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acscentsci.4c00258>

### Author Contributions

<sup>#</sup>T.V. and M.M. contributed equally to this manuscript. T.V. and M.J. conceived the project. T.V. and G.S. developed the automated screening methods. T.V. and C.S. performed experiments. T.V. analyzed screening results and NGS data. M.M. developed, applied, and analyzed the machine learning pipeline. R.T. developed initial computational models. M.J., S.P., and T.R.W. supervised experimental work. A.K. supervised machine learning aspects. T.V., M.M., and M.J. wrote the manuscript with input from all authors.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors thank Fadri Christoffel for synthesizing the gold cofactor. This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. Additional funding was provided by the NCCR Molecular Systems Engineering (grant number 200021\_178760). R.T. acknowledges a grant from the Naito Foundation.

## ■ REFERENCES

- (1) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **2019**, *16*, 687–694.
- (2) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020**, *10*, 1210–1223.
- (3) Freschlin, C. R.; Fahlberg, S. A.; Romero, P. A. Machine learning to navigate fitness landscapes for protein engineering. *Curr. Opin. Biotechnol.* **2022**, *75*, 102713.
- (4) Saito, Y.; et al. Machine-learning-guided mutagenesis for directed evolution of fluorescent proteins. *ACS Synth. Biol.* **2018**, *7*, 2014–2022.
- (5) Biswas, S.; Khimulya, G.; Alley, E. C.; Esvelt, K. M.; Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **2021**, *18*, 389–396.
- (6) Bedbrook, C. N.; Yang, K. K.; Rice, A. J.; Gradinaru, V.; Arnold, F. H. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLoS Comput. Biol.* **2017**, *13*, No. e1005786.
- (7) Bedbrook, C. N.; et al. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. Methods* **2019**, *16*, 1176–1184.
- (8) Romero, P. A.; Krause, A.; Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, No. E193.
- (9) Greenhalgh, J. C.; Fahlberg, S. A.; Pfleger, B. F.; Romero, P. A. Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. *Nat. Commun.* **2021**, *12*, 1–10.
- (10) Li, G.; et al. Machine learning enables selection of epistatic enzyme mutants for stability against unfolding and detrimental aggregation. *ChemBioChem.* **2021**, *22*, 904–914.
- (11) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 8852–8858.
- (12) Gelman, S.; Fahlberg, S. A.; Heinzelman, P.; Romero, P. A.; Gitter, A. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, No. e2104878118.
- (13) Wittmann, B. J.; Yue, Y.; Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **2021**, *12*, 1026–1045.
- (14) Atwal, G. S.; Kinney, J. B. Learning quantitative sequence–function relationships from massively parallel experiments. *J. Stat. Phys.* **2016**, *162*, 1203–1243.
- (15) Höllerer, S.; Desczyk, C.; Muro, R. F.; Jeschek, M. From sequence to function and back – High-throughput sequence-function mapping in synthetic biology. *Curr. Opin. Syst. Biol.* **2024**, *37*, 100499.
- (16) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481–2491.
- (17) Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z. A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers* **2005**, *80*, 775–786.

- (18) Barley, M. H.; Turner, N. J.; Goodacre, R. Improved descriptors for the quantitative structure-activity relationship modeling of peptides and proteins. *J. Chem. Inf. Model.* **2018**, *58*, 234–243.
- (19) Somnath, V. R.; Bunne, C.; Krause, A. Multi-Scale Representation Learning on Proteins. In *Advances in Neural Information Processing Systems*; Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J. W., Eds.; MIT Press: Cambridge, MA, 2021.
- (20) Gainza, P.; et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **2020**, *17*, 184–192.
- (21) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16*, 1315–1322.
- (22) Yang, K. K.; Wu, Z.; Bedbrook, C. N.; Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics* **2018**, *34*, 2642–2648.
- (23) Fox, R. J.; et al. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **2007**, *25*, 338–344.
- (24) Cadet, X. F.; Gelly, J. C.; van Noord, A.; Cadet, F.; Acevedo-Rocha, C. G. Learning strategies in protein directed evolution. In *Directed Evolution: Methods and Protocols*; Currin, A., Swainston, N., Eds.; Springer US: New York, NY, 2022; pp 225–275 DOI: 10.1007/978-1-0716-2152-3\_15.
- (25) Saito, Y.; et al. Machine-learning-guided library design cycle for directed evolution of enzymes: The effects of training data composition on sequence space exploration. *ACS Catal.* **2021**, *11*, 14615–14624.
- (26) Buchler, J.; Malca, S. H.; Patsch, D.; Voss, M.; Turner, N. J.; Bornscheuer, U. T.; Allemann, O.; Le Chapelain, C.; Lumbroso, A.; Loiseleur, O.; Buller, R.; et al. Algorithm-aided engineering of aliphatic halogenase WelOS\* for the asymmetric late-stage functionalization of soraphens. *Nat. Commun.* **2022**, *13*, 371.
- (27) Srinivas, N.; Krause, A.; Kakade, S. M.; Seeger, M. W. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theory* **2012**, *58*, 3250–3265.
- (28) Hie, B. L.; Yang, K. K. Adaptive machine learning for protein engineering. *Curr. Opin. Struct. Biol.* **2022**, *72*, 145–152.
- (29) Vornholt, T.; Jeschek, M. The quest for xenobiotic enzymes: From new enzymes for chemistry to a novel chemistry of life. *ChemBioChem.* **2020**, *21*, 2241–2249.
- (30) Kan, S. B. J.; Lewis, R. D.; Chen, K.; Arnold, F. H. Directed evolution of cytochrome c for carbon–silicon bond formation: Bringing silicon to life. *Science* **2016**, *354*, 1048–1051.
- (31) Zhou, Z.; Roelfes, G. Synergistic catalysis in an artificial enzyme by simultaneous action of two abiological catalytic sites. *Nat. Catal.* **2020**, *3*, 289–294.
- (32) Yang, H.; et al. Evolving artificial metalloenzymes via random mutagenesis. *Nat. Chem.* **2018**, *10*, 318–324.
- (33) Jeschek, M.; et al. Directed evolution of artificial metalloenzymes for in vivo metathesis. *Nature* **2016**, *537*, 661–665.
- (34) Key, H. M.; Dydio, P.; Clark, D. S.; Hartwig, J. F. Abiological catalysis by artificial haem proteins containing noble metals in place of iron. *Nature* **2016**, *534*, 534–537.
- (35) Song, W. J.; Tezcan, F. A. A designed supramolecular protein assembly with in vivo enzymatic activity. *Science* **2014**, *346*, 1525–1528.
- (36) Bordeaux, M.; Tyagi, V.; Fasan, R. Highly diastereoselective and enantioselective olefin cyclopropanation using engineered myoglobin-based catalysts. *Angew. Chem., Int. Ed.* **2015**, *54*, 1744–1748.
- (37) Kan, S. B. J.; Huang, X.; Gumulya, Y.; Chen, K.; Arnold, F. H. Genetically programmed chiral organoborane synthesis. *Nature* **2017**, *552*, 132–136.
- (38) Dydio, P.; Key, H. M.; Nazarenko, A.; Rha, J. Y.-E.; Seyedkazemi, V.; Clark, D. S.; Hartwig, J. F.; et al. An artificial metalloenzyme with the kinetics of native enzymes. *Science* **2016**, *354*, 102–106.
- (39) Studer, S.; Hansen, D. A.; Pianowski, Z. L.; Mittl, P. R. E.; Debon, A.; Guffy, S. L.; Der, B. S.; Kuhlman, B.; Hilvert, D.; et al. Evolution of a highly active and enantiospecific metalloenzyme from short peptides. *Science* **2018**, *362*, 1285–1288.
- (40) Vornholt, T.; Christoffel, F.; Pellizzoni, M. M.; Panke, S.; Ward, T. R.; Jeschek, M.; et al. Systematic engineering of artificial metalloenzymes for new-to-nature reactions. *Sci. Adv.* **2021**, *7*, No. eabe4208.
- (41) Currin, A.; Swainston, N.; Day, P. J.; Kell, D. B. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.* **2015**, *44*, 1172–1239.
- (42) Reetz, M. T.; Kahakeaw, D.; Lohmer, R. Addressing the numbers problem in directed evolution. *ChemBioChem.* **2008**, *9*, 1797–1804.
- (43) Chen, Y.; et al. Barcoded sequencing workflow for high throughput digitization of hybridoma antibody variable domain sequences. *J. Immunol. Methods* **2018**, *455*, 88–94.
- (44) Glenn, T. C.; et al. Adapterama II: Universal amplicon sequencing on Illumina platforms (TaggMatrix). *PeerJ.* **2019**, *7*, No. e7786.
- (45) Wittmann, B. J.; Johnston, K. E.; Almhjell, P. J.; Arnold, F. H. evSeq: Cost-effective amplicon sequencing of every variant in a protein library. *ACS Synth. Biol.* **2022**, *11*, 1313–1324.
- (46) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, 2005 DOI: 10.7551/mitpress/3206.001.0001.
- (47) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **2011**, *79*, 830–838.
- (48) Kulesza, A.; Taskar, B. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.* **2012**, *5*, 123–286.
- (49) Nava, E.; Mutný, M.; Krause, A. Diversified sampling for batched Bayesian optimization with determinantal point processes. In *Proceedings of the 25<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2022.
- (50) Kuiper, B. P.; Prins, R. C.; Billerbeck, S. Oligo pools as an affordable source of synthetic DNA for cost-effective library construction in protein- and metabolic pathway engineering. *ChemBioChem.* **2022**, *23*, No. e202100507.
- (51) Omelina, E. S.; Ivankin, A. V.; Letiagina, A. E.; Pindyurin, A. V. Optimized PCR conditions minimizing the formation of chimeric DNA molecules from MPRA plasmid libraries. *BMC Genomics* **2019**, *20*, 1–10.
- (52) Thompson, J. R.; Marcelino, L. A.; Polz, M. F. Heteroduplexes in mixed-template amplifications: Formation, consequence and elimination by ‘reconditioning PCR’. *Nucleic Acids Res.* **2002**, *30*, 2083–2088.
- (53) Burgener, S.; Dačević, B.; Zhang, X.; Ward, T. R. Binding interactions and inhibition mechanisms of gold complexes in thiamine diphosphate-dependent enzymes. *Biochemistry* **2023**, *62*, 3303–3311.
- (54) Van Der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
- (55) Xu, Y.; et al. Deep dive into machine learning models for protein engineering. *J. Chem. Inf. Model.* **2020**, *60*, 2773–2790.
- (56) Ma, E. J.; et al. Machine-directed evolution of an imine reductase for activity and stereoselectivity. *ACS Catal.* **2021**, *11*, 12433–12445.
- (57) Rapp, J. T.; Bremer, B. J.; Romero, P. A. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat. Chem. Eng.* **2024**, *1*, 97–107.
- (58) Carbonell, P.; Radivojevic, T.; García Martín, H. Opportunities at the intersection of synthetic biology, machine learning, and automation. *ACS Synth. Biol.* **2019**, *8*, 1474–1477.
- (59) Hsu, C.; Nisonoff, H.; Fannjiang, C.; Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* **2022**, *40*, 1114–1122.
- (60) Shanehsazadeh, A.; Belanger, D.; Dohan, D. Is transfer learning necessary for protein landscape prediction? *Quant. Biol. Biomol.* **2020**, DOI: 10.48550/arXiv.2011.03443.

(61) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, 379, 1123–1130.

(62) Brandes, N.; Goldman, G.; Wang, C. H.; Ye, C. J.; Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **2023**, 55, 1512–1522.

(63) Sundar, V.; Tu, B.; Guan, L.; Esvelt, K. FLIGHTED: Inferring Fitness Landscapes from Noisy High-Throughput Experimental Data. *NeurIPS* **2023**.

(64) Christoffel, F.; et al. Design and evolution of chimeric streptavidin for protein-enabled dual gold catalysis. *Nat. Catal.* **2021**, 4, 643–653.