nature portfolio

Peer Review File

Accelerated enzyme engineering by machine-learning guided cell-free expression

Corresponding Author: Professor Michael Jewett

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The article "Accelerated enzyme engineering by machine-learning guided cell-free expression" discusses a new platform that integrates machine learning (ML) with cell-free expression systems to accelerate enzyme engineering. Specifically, the researchers developed a high-throughput, machine-learning orientated approach to facilitate the design of biocatalysts. This approach allows for the exploration of adaptive landscapes across multiple regions of chemical space, thereby contributing to the forward design of biocatalysts. The key is the utilisation of a cell-free gene expression (CFE) system, which allows for the rapid synthesis and functional testing of proteins in a design-build-test-learn (DBTL) workflow.
The researchers identified first-order mutant sequence-function relationships for enzyme variants of specific chemical transformations by evaluating the substrate specificity of the enzyme, and then used these data to train an enhanced ridge regression machine learning model combined with an evolutionary zero-sub-adaptation predictor to infer high-level mutants with higher activity. In the study, they focused specifically on McbA enzymes, an ATP-dependent amide bond synthase from Marinactinospora thermotolerans that is involved in the biosynthesis of marinacarboline secondary metabolites.
In this way, the researchers were able to evolve McbA divergently into a number of different specialised enzymes. For example, for the drugs moclobemide and metoclopramide, multiple mutants with significant activity were obtained by iterative saturation mutagenesis (ISM). For cinchocaine, on the other hand, difficulties were encountered in advancing, which supports the importance of introducing machine learning models in the framework to capture phenotypic interactions.
In addition, the researchers tested different amino acid coding strategies, including one-hot coding, Georgiev features, VHSE features, z-scale features, and physical descriptors. These coding strategies attempted to capture the physicochemical properties of amino acids to provide richer information to the model. Also, various null prediction methods such as Evolutionary Scale Modelling (ESM)-1b, EVmutation-based probability density model and MAESTRO were considered in the study to estimate the free energy of structural changes.
Taken together, this study demonstrates how the use of cell-free expression systems and machine learning can be used to accelerate enzyme engineering and appears to provide a powerful tool for biocatalyst development.

Major Issues
1. Description of the Hot Spot Screening (HSS) method is insufficient. Authors should describe in more detail how Hot Spot Screening (HSS) is performed, including the selection criteria, the experimental methodology, and its plausibility analyses, in order to help readers better understand the plausibility and validity of the methodology.
2. Lack of an adequate review of the existing literature. There are a number of current studies on enzyme-directed evolution, and the authors should outline the strengths and weaknesses of these articles and clearly state the innovations and improvements of the present study compared to these existing works. At a minimum, the following articles should be included:
Structure- and Data-Driven Protein Engineering of Transaminasesfor Improving Activity and Stereoselectivity
Multi-modal deep learning enables effcientand accurate annotation of enzymaticactive sites
3. The article mentions the use of Mean Squared Error (MSE) to evaluate model performance during training, but does not report specific MSE values. Also, the authors should explicitly report the MSE values and their variations under different experimental or model parameters so that readers can better understand the predictive performance and validity of the model.
4. Clarification on the Choice of Augmented Ridge Regression Model. The authors should detail why the augmented ridge regression machine learning model was chosen as the primary prediction model. It is recommended that the authors discuss the advantages of this model over other machine learning models or deep learning models (e.g., random forests, support

vector machines, neural networks, etc.).
5. A variety of amino acid coding approaches were used and their performance compared in the current article, and authors should further explore and compare the performance of amino acid representations generated using large models of the protein language (e.g., ESM or ProtTrans), which provide richer semantic information and may improve prediction accuracy and generalisation. Authors should experiment with encoding using these advanced models and report their performance results to justify why a particular encoding method and model structure was chosen.
6. This study relies on machine learning models trained from a single mutant dataset to predict higher order mutants. Although the authors demonstrate the success of the models on this particular dataset, the ability of these models to generalise across different enzymes and reaction conditions requires further validation. In particular, the representativeness and broad applicability of the current dataset may be insufficient when confronted with more complex or diverse enzyme reactions. Specifically, if a mutant of a particular enzyme lacks significant activity enhancement, or if there are only a small number of better-performing mutants in the dataset, it is a question that needs to be explored as to whether the model can still provide an effective mutation scheme to enhance enzyme activity. The authors should detail how the model maintains its predictive power in the presence of sparse or noisy data, and whether there is a risk of model failure or prediction bias under these conditions. Further experimental or simulation data could help to verify the robustness and generalisation ability of the model in different contexts.
7. The sections of the article that refer to machine learning models and algorithm implementations do not provide detailed instructions or guides for code use, which may affect the ability of other researchers to reproduce the results of this study. It is recommended that the authors provide more detailed instructions for using the code or accompany the publicly available code base with appropriate tutorials.
8. The authors should more clearly articulate how existing protein structure prediction tools (e.g., AlphaFold3, etc.) can be applied to enzyme engineering and what promise or impact they can bring to the field.

Minor Issues
1 In line 127 of the article, the author should provide the full name for the term "UV absorbance" when it is first mentioned. The full name is "ultraviolet absorbance."
2 The authors made predictions on the entire combined dataset (n = 160,000), but do not provide information on the time required to complete these predictions. Understanding the computational efficiency of the model is important to assess its feasibility in real-world applications. It is recommended that the authors report the time taken to make predictions on the entire combined dataset and clearly state the hardware configuration (e.g., CPU or GPU model, memory size, etc.).
3 The authors should perform a more detailed statistical analysis of the predicted values for the entire combined dataset to demonstrate the overall predictive performance of the model. This could include statistics such as the distribution of predicted values, mean, median, standard deviation, skewness and kurtosis.
4 Lack of Detailed Description of the Machine Learning Model Performance. The authors lack a detailed description of the performance of the machine learning model, especially in Figures 4A and 4B, where the process on how to obtain the P-value is not clear. It is suggested that the authors should provide a detailed description of how the P-value was calculated and a more detailed statistical analysis process in the Methods section.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Summary

This manuscript reports a workflow which combines machine learning and cell-free gene expression to expedite enzyme engineering. The authors showcase the workflow by improving the catalytic activity of the generalist amide bond synthetase McbA towards a number of different substrates. The authors performed site-saturation mutagenesis on the active site and assumed substrate tunnels with single mismatch primers to generate a library of linear expression templates (LETs). They screened these LETs using cell free protein expression with three different pairs of substrates and used LC-MS to quantify product yield of the respective three products. These sequence-fitness landscapes were then used as input for a ridge-regression machine learning algorithm. The hyperparameters, encoding strategies, and fitness predictors were all optimized using these datasets. They compared the results of their machine learning predictions with ones they obtained from performing iterative site-saturation mutagenesis (ISM), and found that the ISM predictions were nearly always predicted by the machine learning method, and conversely, the top-performing ML results were often not able to be predicted from ISM strategy. They employed this same ML workflow on six additional desired products and were able to improve yield for all nine of the identified products with enzyme engineering. This reviewer believes the authors present an effective use of ML and CFE, but need to more deeply consider the biochemical implications of their findings, highlight the generalizability of their workflow, and compare it to similar approaches. If the authors address the below major and minor revisions, the paper should be published.

Major Comments

1. The results from the machine learning-aided enzyme engineering campaigns would be strengthened by offering some biochemical intuition that can be gained from the use of machine learning that may otherwise not have been uncovered. Is there anything about the enzyme mechanism or function of particular residues that is revealed by looking at which residues conferred beneficial mutations? This is alluded to in Fig. S25, but could use a lot more elaboration. Similarly, it would be

beneficial to have a more systematic understanding of the substrate scope (e.g. why were aliphatic acids poorly tolerated?), and how it changes after the mutagenesis.

2. Many datasets for site saturation mutagenesis and hot spot screens have been generated from previous enzyme engineering efforts. Can the generalizability of this approach be shown using an existing, publicly available dataset for a new enzyme? This would also be a helpful way to evaluate the efficacy of this model against other, previously published enzyme engineering efforts.
The motivation for including Fig. 5 and the corresponding results section needs to be clearer. Is there new knowledge being conveyed, or is this just the same workflow detailed in the prior sections of the paper repeated verbatim for 6 new products?

3. The discussion does a poor job at contextualizing this workflow with other similar enzyme engineering/evolution approaches. The manuscript would be greatly strengthened by a quantitative comparison of the accessible design space, throughput, confidence (how many predicted variants must be tested), time, and cost for different workflows.

4. Methods have no details on the cell-free protein expression step.

Minor Comments

1. The abstract and introduction could benefit from saying how much the McbA enzyme was improved in specific reactions to help evaluate the efficacy of the method (between 1.6 and 34-fold improvements across six compounds).

2. The manuscript is lacking detail that may help the reader follow along in several places. For example, Figure 3A would be strengthened by more explanation and interpretation of the workflow and result of the NDCG plot. NDCG is used in the x-axis of the plot here, but there is no reference to this plot in the results and no explanation of what NDCG is in the figure legend. NDCG does not appear until the next page after Figure 3A is referenced. What is the takeaway from this panel for the reader? How should they understand the result of the zero shot selection?

3. Similarly, it is not clear what the reader is meant to gain by the right hand plots in Figures 4A and 4B (Spearman vs. NDCG). This plot is not referenced in the results. Presumably, the same information is in Figure S15.

4. It is difficult to assess how good the ML model is based on Figures 4A and S15. How does this model fare against similar ML models? What is the p-value of the correlation? Additionally, “rho” should be defined in the figure legend.

5. Figure S15 has a minor typo: “Pyhiscal Descriptors” is written instead of “Physical Descriptors”.

6. Fig. 2D highlights products that are currently inaccessible, and seems to indicate that the paper will, at some point, focus on expanding the enzyme’s catalytic abilities to include these products. However, that seems to be the only mention of these inaccessible products. Inclusion of this subfigure seems misleading unless there is a section focusing on engineering the enzyme to better produce these.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
• Summary

In this manuscript, the authors present a pipeline to screen amide synthetase variants for new reactivities in high-throughput, the latter of which is enabled through cell-free expression of candidates. Relying on the resulting sequence-activity data, they train reaction-specific ML models with the goal to predict and design variants with high activity for the respective reaction in silico across a larger combinatorial sequence space of 4^20=160k variants.

• General assessment/criticism

In general, the topic is timely and impactful. The study is conducted in an appropriate fashion and it is easy to follow the major narrative/rationale. My main criticism is towards several claims made for the developed ML models, in particular on performance and capability to extrapolate/generalize across the entire sequence space. These claims do not withhold closer scrutiny and/or are not sufficiently backed by experimental/statistical evidence in my opinion. I provide specific points to that end below. If these claims are put into perspective and analyzed more carefully in a revision, I’d be supportive of publication.

• Main points

- It is unclear how a model trained only on single-mutant data could infer epistatic (i.e. non-linear) interactions between the positions. I see how the zero-shot parts could help to that end, but then these do not contain any info on the target activity for the tested substrate-product pairs. The authors are encouraged to clarify how this could work (conceptually) as such extrapolations are considered very difficult in ML at least without adding any additional (e.g. biophysical) information. Could it be that the selected positions behave more additively and thus the model becomes quite good without being able to

resemble residue interaction? The fact that the simple linear models are not much worse (in some cases better) that augmented ones may indicate that. [see also comments below on generalizability across the entire 20^4 sequence space]

- L247-248: “The evaluated augmented models outperformed the ridge regression model alone.” Looking at Fig. S15, I would argue that this is not consistently the case and depends on the reaction and performance metric. E.g. for Moc, the improvements over the simple model are e.g. minor. For Meto, the simple model outperforms all augmented ones in terms of Spearman ranking. The authors should put this claim into perspective and provide error bars for the performance metrics in Fig. S15 for the different training/validation splits performed. The latter would allow to evaluate whether the improvements are within error or not.

- Design of new variants:

- the improvements over top variants from ISM (“M4”) are rather small in Fig. 4 C+D. This is surprising because it implies that the sequence space of 20^4 variants does not contain any significantly better variants than the ones found in ISM (which covers <1% of total space).
For Moc, only three variants with miniscule improvements (probably within error) were found. For meto, most of the designed variants are inactive. Crucially, comparing the ranks in Fig. 4C-E with actual performance of the best-predicted variants, there seems to be little rank correlation. The authors are encouraged to report the ranking correlations also for these designed variants and put the claims made in terms of model performance in perspective.

- Generalizability of the model across the sequence space: This is strongly related to the point before. Looking more closely at top predictions, one finds that the variants have a very high sequence similarity, i.e. are very close in sequence space to each other. Examples (see Tab. S1):
--> Moc: M4 = SSFL, and the three slightly better variants are SSFV, SSFI, SSFT.
--> Meto: M4 = SCFT, and the only better variant is SCFS (SAFS and SSFT are close as well)
In other words, the predictions deviate often only by one mutation from the ISM path. The found replacements are often conservative (“trivial”) (see above, L to I or T to S; the later is highlighted as “surprising finding” in the main text). Further, there is a strong tendency of the model to suggest amino acids found to be best/good in the single-site screens. These are strong indications that there may be strong overfitting to small regions of the entire sequence space close to the training data/ISM path. A way to test this would be to construct a set of variants that is further away and uniformly distributed in sequence space, and then test the models’ performance on these variants.
With the presented results at hand, I would argue a justified claim would be that the ML models aid the search around hypothetical ISM trajectories and thus might help to reduce effort and increases success rates. But good generalizability across the entire 20^4 space it is not tested for and seems highly unlikely given the high sequence similarity amongst predicted top variants.

- Designed variants for six additional compounds: some of the improvements found are nice, but a comparison to ISM is missing, which unfortunately does not allow to make any conclusions on the “effectiveness” of the ML-guided search over the conventional approach in these cases. Moreover, there is a high degree of sequence similarity amongst predicted top variants also in these cases.

- Discussion:
- L341-344: The claim “…we comprehensively mapped the sequence-function landscape of McbA” is unjustified, since only four residues of McbA were looked at and even within the 20^4 subspace generalizability remains undemonstrated (see comment above on generalizability and overfitting).
- L352-356: The claims made in this part are in my opinion a strong exaggeration of the actual findings. It is merely speculative and likely not true that the entire space of 20^N was comprehensively mapped. All tested variants are very close in sequence space and generalizability is not tested for.

• Minor points
- Generally, the figures/captions lack critical information. Few examples are: what the training and test data were for models; quantitative information on activity metrics; what the displayed different activity metrics area and whether/how they are normalized (e.g. Fig. 4A/B); for which encoding strategy the model performance metrics are shown in Fig. 3A etc.). This should be carefully revised to facilitate understanding for the readers.

- Typo in Fig. 1: “accesible”

- It would be helpful to add the ATP concentration in Fig. 1A

- Fig. 2C: It is not specified how high the conversion was for these compounds (only a range is given in the main text). This info should be added to the figure.

- Fig. 3a: Should be 64 x 19 = 1216 mutants (instead of 1280) for the HSS?

- Fig. 3a: typo “extrpolate”

- Fig. 3b-d: the colour scale has no label, ticks etc. specifying what the activity metric is here (caption says it should be % conv.)

- It was not clear from Fig. 3 or main text which were the top4 residues selected as training data (n=77) for the initial models. I believe this is somewhere in the SI, but it would be helpful to give this information in the main manuscript.

- L259-263: “the augmented EVmutation model”: upon reading it was not entirely clear whether it is really one model trained for the two reactions (multi-objective learning) or two models trained separately for each reaction. According to Fig. 3A and after further reading, I’m quite sure it is the later, but it would help understanding to make this clear early on.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
1. The authors have still not fully described how the HSS (particularly the active site selection process) is performed. Please describe the methodology used to identify active sites, including any criteria or algorithms employed.
2. The GitHub page lacks a comprehensive Readme file to guide readers through the project. Although the authors have provided all necessary raw data, source code, and an environment.yml file, there is no overview of the entire workflow.

(Remarks on code availability)
The GitHub page lacks a comprehensive Readme file to guide readers through the project. Although the authors have provided all necessary raw data, source code, and an environment.yml file, there is no overview of the entire workflow.

Reviewer #2

(Remarks to the Author)
The authors have addressed all of my concerns.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have added critical information and analyses during the revision of their manuscript. Further, they have addressed my previous concerns by discussing some of the main claims and/or putting them into perspective. Thus, I am supportive of publication.

(Remarks on code availability)

Reviewer #1 (Remarks to the Author):

The article "Accelerated enzyme engineering by machine-learning guided cell-free expression" discusses a new platform that integrates machine learning (ML) with cell-free expression systems to accelerate enzyme engineering. Specifically, the researchers developed a high-throughput, machine-learning orientated approach to facilitate the design of biocatalysts. This approach allows for the exploration of adaptive landscapes across multiple regions of chemical space, thereby contributing to the forward design of biocatalysts. The key is the utilisation of a cell-free gene expression (CFE) system, which allows for the rapid synthesis and functional testing of proteins in a design-build-test-learn (DBTL) workflow.
The researchers identified first-order mutant sequence-function relationships for enzyme variants of specific chemical transformations by evaluating the substrate specificity of the enzyme, and then used these data to train an enhanced ridge regression machine learning model combined with an evolutionary zero-sub-adaptation predictor to infer high-level mutants with higher activity. In the study, they focused specifically on McbA enzymes, an ATP-dependent amide bond synthase from Marinactinospora thermotolerans that is involved in the biosynthesis of marinacarboline secondary metabolites.
In this way, the researchers were able to evolve McbA divergently into a number of different specialised enzymes. For example, for the drugs moclobemide and metoclopramide, multiple mutants with significant activity were obtained by iterative saturation mutagenesis (ISM). For cinchocaine, on the other hand, difficulties were encountered in advancing, which supports the importance of introducing machine learning models in the framework to capture phenotypic interactions.
In addition, the researchers tested different amino acid coding strategies, including one-hot coding, Georgiev features, VHSE features, z-scale features, and physical descriptors. These coding strategies attempted to capture the physicochemical properties of amino acids to provide richer information to the model. Also, various null prediction methods such as Evolutionary Scale Modelling (ESM)-1b, EVmutation-based probability density model and MAESTRO were considered in the study to estimate the free energy of structural changes.
Taken together, this study demonstrates how the use of cell-free expression systems and machine learning can be used to accelerate enzyme engineering and appears to provide a powerful tool for biocatalyst development.

We thank you for celebrating how our work provides a powerful platform and important tool for biocatalyst development.

Major Issues
1. Description of the Hot Spot Screening (HSS) method is insufficient. Authors should describe in more detail how Hot Spot Screening (HSS) is performed, including the selection criteria, the experimental methodology, and its plausibility analyses, in order to help readers better understand the plausibility and validity of the methodology.

We thank you for raising the need for additional clarification on the hot spot screen (HSS). For the selection of the initial 64 residues, we used the crystal structure of McbA and general

heuristics (i.e., capturing the active site and substrate tunnels) to guide our selection of residues.

Specifically, we added text on p. 7:

> *"Guided by the crystal structure of McbA (PDB: 6SQ8), we selected 64 residues that completely enclosed the active site and putative substrate tunnels (e.g., residues within 10 Å of the docked native substrates)."*

For following ISM rounds, we selected the most highly active residues based on ease of experimentation. For moclobemide, we selected six residues, which were all above a threshold of 1.5-fold activity over wild type. For metoclopramide, we selected 10 residues, which were all above a threshold of 1.25-fold activity over wild type. However, there was no biological motivation for these selections, and it was dependent on each substrates sequence-fitness landscape. We believe that our data indicate the importance of screening a large amount of sequence space to investigate residue-catalysis relationships (e.g., a solvent exposed residue away from the active site increased yields in some instances). Our accompanying method for site saturation mutagenesis is valid and makes it easier for the broader community to explore diverse sequence spaces.

In the revised manuscript, we now highlight how the residues were selected for the HSS on p. 9-10:

> *"For moclobemide, we selected six residues from the HSS above a threshold of 1.5-fold activity over wild type to mutate over three rounds of ISM (**Fig. S7**). We first fixed the top mutation from the HSS (V177S) and performed site-saturation mutagenesis on the five additional residues. By reintroducing previously fixed mutations in subsequent rounds, we explored potential epistatic interactions (e.g., S177 was saturated in ISM round 2, given V177S was incorporated before A323F). In addition, we completely explored all combinatorial double mutants of the top two residues, which showed directly additive impacts for moclobemide synthesis (**Fig. S7**). After three rounds of ISM, whereby we selected the most highly active mutant in each round to start the next round, we identified…"*
>
> *"For metoclopramide, we selected 10 residues for the ISM campaign, which were all above a threshold of 1.25-fold activity over wild type. Three rounds of ISM for metoclopramide yielded a quadruple mutant that displayed nearly 30-fold activity over wt-McbA."*

In terms of the experimental methodology, this is described in the methods section for the HSS and ISM campaigns. There, we provide a detailed description of our cell-free library generation protocol, the experimental methodology of the amide synthetase activity assay, the analytics protocol using LC-MS, and all specific primers and reagents used in the methods section. Importantly, we newly add a section on cell-free gene expression conditions that was

inadvertently missed before (and noted by Reviewer 2, point 4). We believe this covers all aspects of our methodology and readily enables facile replication and/or adaptation.

2. Lack of an adequate review of the existing literature. There are a number of current studies on enzyme-directed evolution, and the authors should outline the strengths and weaknesses of these articles and clearly state the innovations and improvements of the present study compared to these existing works. At a minimum, the following articles should be included: Structure- and Data-Driven Protein Engineering of Transaminases for lmproving Activity and Stereoselectivity Multi-modal deep learning enables effcient and accurate annotation of enzymatic active sites

We thank you for bringing to our attention these additional manuscripts in directed evolution. In the revised manuscript, we have expanded the description of the power of directed evolution and how our approach offers a complement.

Specifically, we added text in the introduction on p. 3:

> *"Our approach offers a compliment to the growing toolbox of directed evolution strategies, such as those that predict catalytic features of enzymes."*
>
> *And cite the three examples of this:*
> *https://www.science.org/doi/10.1126/science.adf2465*
> *https://onlinelibrary.wiley.com/doi/full/10.1002/anie.202301660*
> *https://www.nature.com/articles/s41467-024-51511-6*

3. The article mentions the use of Mean Squared Error (MSE) to evaluate model performance during training, but does not report specific MSE values. Also, the authors should explicitly report the MSE values and their variations under different experimental or model parameters so that readers can better understand the predictive performance and validity of the model.

We thank you for requesting this information. In the revised manuscript, we now clarify the methods section to state that MSE was used specifically for hyperparameter tuning of α using the scikit-learn GridSearchCV() function with a 5-fold cross-validation. We searched for α values ranging from 0.01-100, which would be cumbersome to show all values of MSE for in the supplement. Given the fact that NDCG aligns with our experimental goals of discovering high-fitness variants with minimal screening burden, we used the NDCG value to evaluate model performance and select a top-performing model (and not MSE). For readers convenience and to support replicability and understanding of model performance, MSE values for each model variation are retrievable through the provided GitHub code.

Specifically, we added the text on p. 20:

> *Augmented models (using combinations of the above zero-shot predictors and amino acid encodings) were trained on the single site saturation libraries for four residues (n ≈*

80) and tested on the withheld higher-order mutants from the additional rounds of saturation mutagenesis (n ≈ 200). The four residues (hot spots) were determined by selecting the four residues that contained mutations with highest improvement in yield among the 64 residues tested. Hyperparameter turning of α was performed using repeated 5-fold cross-validation (with 20 repeats) by randomly sampling 80% of the training data and testing on the withheld 20%; model performance for hyperparameter tuning was scored using mean squared error (MSE).

4. Clarification on the Choice of Augmented Ridge Regression Model. The authors should detail why the augmented ridge regression machine learning model was chosen as the primary prediction model. It is recommended that the authors discuss the advantages of this model over other machine learning models or deep learning models (e.g., random forests, support vector machines, neural networks, etc.).

In the revised manuscript, we better detail why the augmented ridge regression machine learning model was chosen as the primary prediction model. In short, we chose this simple linear regression model as it was previously shown to outperform more sophisticated models, especially in a regime where limited data is available (https://www.nature.com/articles/s41587-021-01146-5). We believe this criterion fits our goal of training a model on ~80 single mutants to predict an entire combinatorial space of 20^4 mutants.

In the revised manuscript, we now discuss this previous work in the results section when stating why we chose augmented ridge regression. We also clarified in the discussion section that there are alternative complex models (random forests, SVMs, etc.) that may perform better, especially in instances where more data is collected for training.

Specifically, we add text on p. 9:

"The key idea was to use single mutant data from the HSS to extrapolate higher order mutants with increased activity (Fig. 3A). To achieve this goal, we chose to fit augmented ridge regression models with our data—as such models are simple and have been previously shown to outperform more sophisticated models for protein engineering[43]—allowing us to predict higher-order mutants with increased activity."

Furthermore, we add on p.13:

"We selected an augmented ridge regression machine learning model because it had previously been shown to perform well compared to more sophisticated approaches and was able to extrapolate from single to higher order mutations, an observation that is consistent with our data."

As well as on p. 13:

"There may be instances where more complex models (e.g., random forests, support vector machines, neural networks, etc.) that can generalize better to the entire sequence

space will be necessary to navigate complex fitness landscapes. Despite this, the ML framework used here aided in the search around hypothetical ISM trajectories, reducing effort and increasing success rates in enzyme engineering campaigns."

5. A variety of amino acid coding approaches were used and their performance compared in the current article, and authors should further explore and compare the performance of amino acid representations generated using large models of the protein language (e.g., ESM or ProtTrans), which provide richer semantic information and may improve prediction accuracy and generalisation. Authors should experiment with encoding using these advanced models and report their performance results to justify why a particular encoding method and model structure was chosen.

We thank the reviewer for the suggestion. As noted in our rationale above on the selection of augmented ridge regression, we believe in the power and simplicity of the model, as demonstrated in our 9 unique engineering campaigns. That said, we appreciate that new models are being developed rapidly. For example, Hsu et al. (https://www.nature.com/articles/s41587-021-01146-5) compared augmented ridge regression against protein LLM models and found augmented ridge regression to be superior in many tested datasets. Wittmann et al. showed several examples where LLM encodings having no additional benefit over one-hot encodings (https://www.sciencedirect.com/science/article/pii/S2405471221002866?via%3Dihub#fig2) and Jian et al. observed that LLMs negatively correlated with experimental fitness landscapes (https://www.biorxiv.org/content/10.1101/2024.07.17.604015v1). While we agree with the reviewer that protein LLMs are incredibly exciting and promising for the field, it seems like we are still figuring out how to best implement them for machine-learning guided directed evolution. Exploring more advanced models and balancing power, accuracy, and simplicity, serve as future extensions of our work.

In the revised manuscript, we try to capture this sentiment in the discussion by referencing these manuscripts and stating on p. 14:

"There are numerous other protein variant effect predictors continuously pushing the state-of-the-art forward that could improve our predictions[57]. More complex sequence encodings based on natural language processing may also outperform augmented encodings[49,58]."

6. This study relies on machine learning models trained from a single mutant dataset to predict higher order mutants. Although the authors demonstrate the success of the models on this particular dataset, the ability of these models to generalise across different enzymes and reaction conditions requires further validation. In particular, the representativeness and broad applicability of the current dataset may be insufficient when confronted with more complex or diverse enzyme reactions. Specifically, if a mutant of a particular enzyme lacks significant activity enhancement, or if there are only a small number of better-performing mutants in the dataset, it is a question that needs to be explored as to whether the model can still provide an

effective mutation scheme to enhance enzyme activity. The authors should detail how the model maintains its predictive power in the presence of sparse or noisy data, and whether there is a risk of model failure or prediction bias under these conditions. Further experimental or simulation data could help to verify the robustness and generalisation ability of the model in different contexts.

We share the reviewer's enthusiasm for generalizability. We believe that the concept of combining site saturation data with machine learning is readily generalizable. For example, we were able to apply our model to accelerate the design of 9 protein catalysts for 9 distinct chemistries by extrapolating single-mutant data to higher-order mutations. This is consistent with a recent *Nature Biotechnology* publication that developed the idea of augmented ridge regression (https://www.nature.com/articles/s41587-021-01146-5), which demonstrates the broad generalizability of such a model in both instances of predicting single mutations as well as extrapolating to higher-order (quadruple) mutations for different enzyme classes. In addition, in terms of our methods, we demonstrated site saturation mutagenesis was robust, by showing it worked on two different proteins, both GFP and McbA.

That said, we would not expect this exact same model to work for all enzyme classes, nor was this the goal of our work. Rather, we sought to highlight the creation of a ML-guided platform that integrates cell-free DNA assembly, cell-free gene expression, and functional assays to rapidly map fitness landscapes across protein sequence space of a single enzyme and optimize enzymes for multiple, distinct chemical reactions. In this way, the ML-models can be adapted to specific use cases, but the workflow remains the same. Our work therefore addresses a key gap in ML-assisted engineering methods. Namely, that rapidly building datasets to navigate vast sequence space remains a major challenge for the field (Wittmann, B. J., Yue, Y. & Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. Cell Syst 12, 1026-1045.e7 (2021)), especially considering most genotype-phenotype links are lost in high-throughput enzyme engineering campaigns (Bell, E. L. et al. Directed evolution of an efficient and thermostable PET depolymerase. Nat. Catal. 5, 673–681 (2022)).

In terms of reaction complexity, we believe that the amide synthetase McbA is a good example of a complex reaction mechanism. The reaction is a three-substrate (acid, amine, and ATP) and three product (amide, ADP, and phosphate) reaction that exhibits ping-pong kinetics. However, as stated above, the goal was not to predict across diverse enzyme reaction classes. We agree with the reviewer that we would like to study many more enzymes but believe this goes beyond the scope of any one paper. As it is, this study took an extensive amount of time to complete and the data generated for the hot spot screens alone, 64 residues X 20 amino acids X 9 amide compounds (11,520 reactions), resulted in a dataset reported that is orders of magnitude higher than traditional directed evolution campaigns that only sample the winner or few winners. Indeed, most directed evolution studies of enzymes quantify activity of only a handful of variants, which are insufficient for ML models. For example, recent work from Sarai et al. (https://www.science.org/doi/10.1126/science.adi5554), reported data for just ~10 leading variants, not more than 1000 enzyme variants as we do here.

We agree that having an enriched data-set (mutants that have non-zero activity) is crucial for model performance, as recent work from Frances Arnold has shown (https://pubmed.ncbi.nlm.nih.gov/34416172/). Given that we performed many ML-guided engineering campaigns, we have instances where the data generation is very high quality (e.g., moclobemide) and instances where the fitness landscape is flat (many zero activity mutants and not many mutants that drastically improve activity) and the signal-to-noise ratio is high (e.g., declopramide). While model performance does vary in these cases, we still demonstrate the ability to predict mutants with higher activity than wild type. This provides confidence that the approach is robust.

We have added the following sentence to the discussion on p. 13:

> *"…. in each of the nine test cases, ML-predicted enzyme variants with 4 mutations had greater activity than the combination of the four most active single-residue mutants alone. This observation holds in instances where the data generation is high quality (e.g., moclobemide) and instances where the fitness landscape is flat (i.e., many zero activity mutants and not many mutants that improve activity) and the signal-to-noise ratio is high (e.g., declopramide), which highlights the robustness of our approach."*

7. The sections of the article that refer to machine learning models and algorithm implementations do not provide detailed instructions or guides for code use, which may affect the ability of other researchers to reproduce the results of this study. It is recommended that the authors provide more detailed instructions for using the code or accompany the publicly available code base with appropriate tutorials.

We thank you for raising this point of clarification. In the accompanying Github code, we provide all necessary raw data, source code, and an environment.yml file needed to replicate every machine-learning guided experiment performed in the manuscript. Detailed jupyter notebooks are also included that now describe and walk through every step of the process with more detailed instructions to aid in replicability.

8. The authors should more clearly articulate how existing protein structure prediction tools (e.g., AlphaFold3, etc.) can be applied to enzyme engineering and what promise or impact they can bring to the field.

We agree with the reviewer that it is an exciting time for protein structure prediction tools. While AlphaFold3 (and others like ESM Fold) are being used successfully for protein engineering campaigns for protein-protein and protein-small molecule binders, dynamic protein structures and transition states needed for a deeper understanding of catalysis are not yet modeled in the current stationary structures.

In the revised manuscript, we now also highlight a few of the earliest examples of de novo enzyme design (https://www.nature.com/articles/s41586-023-05696-3; https://www.biorxiv.org/content/10.1101/2024.08.29.610411v1).

Specifically, we now write on p. 14 with these new references:

> *"Beyond the cell-free framework, our work also highlights the versatility of the amide synthetase McbA to be directed to catalyze many unique reactions of interest, including those used in small-molecule pharmaceutical production. Looking forward, we anticipate that the approach described here, especially when augmented with de novo protein design[10,59], will accelerate enzyme engineering campaigns to unlock specialized enzymes with diverse functions and properties."*

Minor Issues

1 In line 127 of the article, the author should provide the full name for the term "UV absorbance" when it is first mentioned. The full name is "ultraviolet absorbance."

We thank you for noticing this mistake and have corrected the text.

2 The authors made predictions on the entire combined dataset (n = 160,000), but do not provide information on the time required to complete these predictions. Understanding the computational efficiency of the model is important to assess its feasibility in real-world applications. It is recommended that the authors report the time taken to make predictions on the entire combined dataset and clearly state the hardware configuration (e.g., CPU or GPU model, memory size, etc.).

In the revised manuscript, we provide additional details of hardware configuration in the methods section and include rough compute times. We state on p. 20:

> *"Given the already trained EVmutation probability density model and the low dimensionality of the encodings, model training and predictions for the entire combinatorial dataset could be made in minutes running on 12th Gen Intel Core i7 with 32 GB RAM without GPU acceleration."*

A key reason for selecting the augmented ridge regression model was its ease of implementation and short computational time (without GPU acceleration), and we thank the reviewer for allowing us to highlight this further. We additionally state in the introduction on p. 3 that:

> *"Importantly, the ML models can be run on the central processing unit (CPU) of a typical computer making our entire approach user-friendly and accessible."*

3 The authors should perform a more detailed statistical analysis of the predicted values for the entire combined dataset to demonstrate the overall predictive performance of the model. This could include statistics such as the distribution of predicted values, mean, median, standard deviation, skewness and kurtosis.

We share the reviewers desire to accurately depict model predictive performance, and we hope to have done so by presenting standard used values of spearman correlation and NDCG. The only way to truly test model predictive performance is to compare experimental data with model predictions. In the case of the entire predicted set of 20^4, we only have experimentally tested the top 25 predictions in each instance, so we do not believe it would be fair or accurate to determine model performance by comparing the distribution of the 20^4 predictions. Additionally, while model accuracy is important, our goal is to effectively predict higher performing mutants, quickly, with minimal screening burden. Validating additional variants, especially those predicted to be mildly active or non-functional, may give us more confidence in the model but won't necessarily drive the directed-evolution campaigns forward.

4 Lack of Detailed Description of the Machine Learning Model Performance. The authors lack a detailed description of the performance of the machine learning model, especially in Figures 4A and 4B, where the process on how to obtain the P-value is not clear. It is suggested that the authors should provide a detailed description of how the P-value was calculated and a more detailed statistical analysis process in the Methods section.

In the revised manuscript, we have clarified that rho (ρ) refers to spearman correlation and not p-value in the caption of Fig. 4 to prevent this confusion. Spearman correlation is calculated according to standard methods and is provided in our accompanying code on GitHub.

Reviewer #2 (Remarks to the Author):

Summary

This manuscript reports a workflow which combines machine learning and cell-free gene expression to expedite enzyme engineering. The authors showcase the workflow by improving the catalytic activity of the generalist amide bond synthetase McbA towards a number of different substrates. The authors performed site-saturation mutagenesis on the active site and assumed substrate tunnels with single mismatch primers to generate a library of linear expression templates (LETs). They screened these LETs using cell free protein expression with three different pairs of substrates and used LC-MS to quantify product yield of the respective three products. These sequence-fitness landscapes were then used as input for a ridge-regression machine learning algorithm. The hyperparameters, encoding strategies, and fitness predictors were all optimized using these datasets. They compared the results of their machine learning predictions with ones they obtained from performing iterative site-saturation mutagenesis (ISM), and found that the ISM predictions were nearly always predicted by the machine learning method, and conversely, the top-performing ML results were often not able to be predicted from ISM strategy. They employed this same ML workflow on six additional desired products and were able to improve yield for all nine of the identified products with enzyme engineering. This reviewer believes the authors present an effective use of ML and CFE, but need to more deeply consider the biochemical implications of their findings, highlight the generalizability of their workflow, and compare it to similar approaches. If the authors address the below major and minor revisions, the paper should be published.

We thank you for celebrating our effort as an effective use of ML and cell-free gene expression.

Major Comments

1. The results from the machine learning-aided enzyme engineering campaigns would be strengthened by offering some biochemical intuition that can be gained from the use of machine learning that may otherwise not have been uncovered. Is there anything about the enzyme mechanism or function of particular residues that is revealed by looking at which residues conferred beneficial mutations? This is alluded to in Fig. S25, but could use a lot more elaboration. Similarly, it would be beneficial to have a more systematic understanding of the substrate scope (e.g. why were aliphatic acids poorly tolerated?), and how it changes after the mutagenesis.

We thank you for making this suggestion and agree that a great use of the large data sets we generated is to tease apart and propose some biochemical interactions. To find residues that impact acid or amine affinities, we can look at instances where, for example, the acid component is changed while the amine component is held constant (or vice versa). The best example of this is for the amide products metoclopramide, cinchocaine, procainaimde, and declopramide, which all contain *N,N*-diethylethylenediamine but different acids. One could propose that V177S is a beneficial mutation for this reaction, but it is only a generalization and not universal as procainamide did not contain this mutation. Given that we engineered nine distinct enzymes and the combinatorial space of potential reactions is as large as it is, it is hard to conclusively make claims like this without more evidence. However, in light of the above example and similar, our data seems to indicate that residue selection (hot spots) are more strongly related to the overall reaction and not the acid or amine fragments on their own. This is indicated by the large number of unique hot spots that we witness for each reaction, even among those with the same acid or amine fragment.

Per your request, we have added a comment on the biological implications of our findings in the discussion. Specifically, we write on p. 13:

> *"We identified 19 unique residue positions within McbA that significantly impact biocatalysis, with each reaction yielding a unique set of these hot spot residues. Across all nine final engineered McbA variants, we made a total of 21 different mutations occurring across 14 different residues (Fig. S25). While many of the substrate pairs contain the same acid or amine, it is difficult to rationalize why certain mutations arise as they are not conserved among many of these enzymes. For example, the amide products metoclopramide, cinchocaine, procainamide, and declopramide all contain N,N-diethylethylenediamine but different acids. One could propose that V177S is a beneficial mutation for this reaction, but it is only a generalization and not universal as procainamide did not contain this mutation. However, our data seems to indicate that residue selection (hot spots) are more strongly related to the overall reaction and not the acid or amine fragments on their own."*

2. Many datasets for site saturation mutagenesis and hot spot screens have been generated from previous enzyme engineering efforts. Can the generalizability of this approach be shown using an existing, publicly available dataset for a new enzyme? This would also be a helpful way to evaluate the efficacy of this model against other, previously published enzyme engineering efforts.

We share the reviewer's enthusiasm for generalizability. In terms of extending our approach to other data, a recent *Nature Biotechnology* publication that proposed the idea of augmented ridge regression (https://www.nature.com/articles/s41587-021-01146-5) as a simple and effective machine-learning strategy has demonstrated the broad generalizability of the model in both instances of predicting single mutations as well as extrapolating to higher-order (quadruple) mutations. Thus, while a different context, generalizability has been shown.

In the revised manuscript on p. 14, we now make several claims in the text that discuss when this model may be applicable to other enzymes (generalizable) and when other models may be advantageous (not generalizable):

> *“There may be instances where more complex models (e.g., random forests, support vector machines, neural networks, etc.) that can generalize better to the entire sequence space will be necessary to navigate complex fitness landscapes, especially where more data is collected for training. Despite this, the ML modeling framework used here aided in the search around hypothetical ISM trajectories, reducing effort and increasing success rates in enzyme engineering campaigns.”*
>
> *"Our approach can, in theory, be applied to any enzyme but will require reaction-specific fine-tuning around data collection and ML model generation."*
>
> *"Engineering campaigns with different proteins may also warrant exploring various ML models and parameters. While we observed excellent performance with the augmented EVmutation model, alternative fitness goals may also require alternative fitness predictors. For example, if the goal is to engineer stability, it would reason that a structural-based fitness predictor may be superior. There are numerous other protein variant effect predictors continuously pushing the state-of-the-art forward that could improve our predictions[56]. More complex sequence encodings based on natural language processing may also outperform augmented fitness predictions[48,57]. Finally, we note that training the ML models on more residues, multi-mutant data (including data from multiple rounds of mutagenesis or random combinations of mutants that diversify amino acids across the entire protein of interest), or kinetic measurements could be beneficial in engineering better catalysts.”*

The motivation for including Fig. 5 and the corresponding results section needs to be clearer. Is there new knowledge being conveyed, or is this just the same workflow detailed in the prior sections of the paper repeated verbatim for 6 new products?

We thank you for allowing us to clarify the motivation for and results of Figure 5. Overall, we hoped to convey our workflow and that we can generate broad site saturation mutagenesis data (Fig. 3), then use this data to validate ML-guided directed evolution on a small test set (Fig. 4), and finally apply this method to a larger set of reactions of interest (Fig. 5). In Figure 5, we intended to demonstrate that our method can be used for parallelized engineering of multiple reactions simultaneously to expedite enzyme engineering campaigns. The 150 predictions (6 compounds x 25 predictions) were ordered, tested, and validated within just a few days. When considering traditional enzyme engineering campaigns can take weeks to months or more for a single reaction, we believe this final demonstration supports our claims that our method can accelerate enzyme engineering, at least by providing a toehold on where to start. Specifically, in the discussion on p. 13 we now state that:

> *"This framework uniquely integrates a cell-free gene expression and mutagenesis method, ML to expedite directed evolution campaigns, and divergent evolution to convert a generalist enzyme into multiple specialists. We showcased this framework by rapidly navigating nine protein engineering campaigns for the amide synthetase McbA, six of which were performed simultaneously."*

3. The discussion does a poor job at contextualizing this workflow with other similar enzyme engineering/evolution approaches. The manuscript would be greatly strengthened by a quantitative comparison of the accessible design space, throughput, confidence (how many predicted variants must be tested), time, and cost for different workflows.

We thank the reviewer for this suggestion on how to greatly strengthen the manuscript.

In the revised version of the manuscript, we have now highlighted the advantages of using our workflow against traditional different methods. We present a general timeline for going through one round of our workflow, and stress the time saved from CFE over *in vivo* expression, the ability to connect sequence to function without sequencing, and the ability to avoid creating/buying large complex combinatorial libraries by using ML-guided predictions. Specifically, we state on p. 14:

> *"Additionally, cell-free expression from linear expression templates expedites the process of exploring sequence-function landscapes since the entire process for protein expression and assessment can be done without cells and we could avoid laborious cloning steps, taking hours instead of days. These features speed the pace of engineering relative to ISM alone; our framework enabled six enzyme engineering campaigns to be simultaneously completed in just one week per compound."*

While it would be a nice and useful comparison, we do not believe it is fair to calculate the time, cost, confidence, accessible design space, etc. for every state-of-the-art directed evolution method in the literature. To our knowledge, we cannot find any directed evolution workflow manuscript that presents data in this manner. Moreover, while we know phage assisted continuous evolution (https://www.nature.com/articles/nature09929) is a powerful directed

evolution method, we don't know ourselves what complex equipment and expertise would be necessary to perform this method and a necessary selection for amide bond formation by McbA. That said, we now direct the reader to our recent Nature Reviews Genetics paper (https://www.nature.com/articles/s41576-019-0186-3) which describes typical costs of cell-free systems and how they are reasonably inexpensive (i.e., ~$3/mL and our reactions were performed often performed on a 10-µL reaction, or just cents per test).

Specifically, we add in the discussion on p. 14 that:

> *"Furthermore, we also note the low cost (i.e., cents per 10-µL reaction) and high scalability of CFE enabling our workflow."*

4. Methods have no details on the cell-free protein expression step.

We thank you for noticing this and have added the details of the cell-free expression system to the methods section. Specifically, we write on p. 15:

> *"Crude cell extracts were prepared as previously described using E. coli BL21 Star (DE3) cells (Invitrogen)61. CFPE reactions were performed based on the PANOx-SP system and carried out in 384-well PCR plates (Bio-Rad) as 10-µL reactions with 1 µL of LET serving as the DNA template. Reactions were incubated at 30 ºC for 16 h."*

Minor Comments

1. The abstract and introduction could benefit from saying how much the McbA enzyme was improved in specific reactions to help evaluate the efficacy of the method (between 1.6 and 34-fold improvements across six compounds).

In the revised manuscript, we add a statement to the introduction to highlight the results of our parallelized engineering campaigns, focusing on all nine compounds (not just six) where improvements ranged from 1.6- to 42-fold. Specifically, we state:

In the abstract:

> *"Over these nine compounds, ML-predicted enzyme variants demonstrated 1.6- to 42-fold improved activity relative to the parent."*

On p. 4:

> *"Our ML-guided approach was able to improve the McbA enzyme between 1.6- and 42-fold across nine compounds."*

2. The manuscript is lacking detail that may help the reader follow along in several places. For example, Figure 3A would be strengthened by more explanation and interpretation of the

workflow and result of the NDCG plot. NDCG is used in the x-axis of the plot here, but there is no reference to this plot in the results and no explanation of what NDCG is in the figure legend. NDCG does not appear until the next page after Figure 3A is referenced. What is the takeaway from this panel for the reader? How should they understand the result of the zero shot selection?

We thank you for helping us clarify the narrative of the manuscript in this key place. We have added a description of NDCG and the zero-shot selector graph in the Figure 3 legend. This panel is used to support our choice for using an EVmutation (EC) zero shot predictor, as it performed well (indicated by NDCG) for both compounds we validated with - with the full validation found in the supplementary information.

> *"Zero shot selection was compared against all encoding strategies (Georgiev is shown) and ranked by NDCG.* ***(B-D)*** *Hot spot screen of 64 identified residues in McbA showing fold-change percent conversion of (**B**) moclobemide, (**C**) metoclopramide, and (**D**) cinchocaine normalized to WT (n = 1; blue indicates low percent conversion, red indicates high percent conversion). The top four highly mutable sites, or hot spots, are highlighted in orange and starred. They are used as the training set for model validation. Full validation results are provided in Supplementary Figure S15."*

We additionally state on p. 10:

> *"Moving forward, we decided to use the site saturation dataset and the augmented EVmutation model with Georgiev encodings given the strong predictive performance among both compounds and the fact that the already-trained probability density model simplified application to other compounds (**Fig. 3A**)."*

3. Similarly, it is not clear what the reader is meant to gain by the right hand plots in Figures 4A and 4B (Spearman vs. NDCG). This plot is not referenced in the results. Presumably, the same information is in Figure S15.

We thank you for allowing us to clarify Figure 4. While Figure S15 provides evidence to support our selection of an augment EV mutation model with Georgiev encodings, we intend Figure 4A-B to provide the reader with insight into how different training datasets result in changes in model performance (a unique observation not shown in Figure S15). Many enzyme engineering campaigns are constrained to collecting smaller data sets and exploring a limited sequence-fitness landscape. Some of these strategies include building libraries with reduced codon usage (NDT, NRT) or performing "scans" (e.g., an "alanine scan"). We provide evidence in Figure 4A-B that data sets built from these traditional directed evolution approaches lead to models with worse predictive performance compared to a full site-saturation mutagenesis library. We added additional clarification on p. 10:

> "Lastly, we tested the necessity of the entire site saturation dataset (*n* = 77) for training models to achieve high predictive performance and to evaluate whether smaller datasets commonly used in protein engineering would be sufficient."

4. It is difficult to assess how good the ML model is based on Figures 4A and S15. How does this model fare against similar ML models? What is the p-value of the correlation? Additionally, "rho" should be defined in the figure legend.

In the revised manuscript, we provide clarity by defining ρ in the figure 4 legend as the spearman correlation.

> *"Analysis of model fidelity with training sets built with smaller libraries than saturation mutagenesis, including reduced codon sets (NDT, NRT) and reduced amino acid alphabets based on BLOSUM50, quantified by spearman correlation (ρ) and NDCG."*

As is commonly used in the field (https://www.nature.com/articles/s41587-021-01146-5, https://www.sciencedirect.com/science/article/pii/S2405471221002866?via%3Dihub), we assess model predictive power by using NDCG and spearman correlation. Figures 4A and S15 are good ways to qualitatively compare different encoding strategies and zero-shot predictors. We believe true model assessment comes in the form of experimentally testing the top predictions, as our goal is to engineer an enzyme to be better than wildtype. We do not perform a statistical analysis and therefore do not have p-values to represent.

5. Figure S15 has a minor typo: "Pyhiscal Descriptors" is written instead of "Physical Descriptors".

We thank you for noticing this typo and have corrected it.

6. Fig. 2D highlights products that are currently inaccessible, and seems to indicate that the paper will, at some point, focus on expanding the enzyme's catalytic abilities to include these products. However, that seems to be the only mention of these inaccessible products. Inclusion of this subfigure seems misleading unless there is a section focusing on engineering the enzyme to better produce these.

We thank you for requesting this clarification and apologize for potentially misleading the reader into thinking they will be a focus of the manuscript. The utility in including this subsection is to provide concrete examples of molecules that are not able to be synthesized by McbA in direct contrast with those that are, so that the reader can better understand the innate preferences of McbA. For example, including nonivamide and capsaicin (in contrast to those molecules in panel C) demonstrate the inability of McbA to react with aliphatic and fatty acids. Likewise, inclusion of imatinib and nilotinib demonstrate potential size or steric constraints of acceptable substrates. We hope these molecular examples, along with the heatmap, provide readers with a better understanding of the substrate scope of McbA and steer future enzyme engineering work.

We add clarification text on p. 5:

> *"Moreover, several important molecules are not synthesized by wild-type McbA (**Fig. 2D**). These inaccessible products suggest that McbA may not be able to react with aliphatic and fatty acids (e.g., nonivamide, capsaicin) or particular large substrates (e.g., imatinib, nilotinib). A better understanding of substrate scope can guide enzyme engineering work."*

Reviewer #3 (Remarks to the Author):

• Summary

In this manuscript, the authors present a pipeline to screen amide synthetase variants for new reactivities in high-throughput, the latter of which is enabled through cell-free expression of candidates. Relying on the resulting sequence-activity data, they train reaction-specific ML models with the goal to predict and design variants with high activity for the respective reaction in silico across a larger combinatorial sequence space of 4^20=160k variants.

• General assessment/criticism

In general, the topic is timely and impactful. The study is conducted in an appropriate fashion and it is easy to follow the major narrative/rationale. My main criticism is towards several claims made for the developed ML models, in particular on performance and capability to extrapolate/generalize across the entire sequence space. These claims do not withhold closer scrutiny and/or are not sufficiently backed by experimental/statistical evidence in my opinion. I provide specific points to that end below. If these claims are put into perspective and analyzed more carefully in a revision, I'd be supportive of publication.

We thank you for supporting the impact and timeliness of our work. As requested, we put our claims into more perspective with more careful perspective and analysis in our revised manuscript.

• Main points

- It is unclear how a model trained only on single-mutant data could infer epistatic (i.e. nonlinear) interactions between the positions. I see how the zero-shot parts could help to that end, but then these do not contain any info on the target activity for the tested substrate-product pairs. The authors are encouraged to clarify how this could work (conceptually) as such extrapolations are considered very difficult in ML at least without adding any additional (e.g. biophysical) information. Could it be that the selected positions behave more additively and thus the model becomes quite good without being able to resemble residue interaction? The fact that the simple linear models are not much worse (in some cases better) that augmented ones may indicate that. [see also comments below on generalizability across the entire 20^4 sequence space]

We thank you for raising these points. We share your interest in understanding more how the model is successfully predicting combinations of mutations. As we better emphasize in our revised work, we in part selected the linear augmented ridge regression model because Hsu et al. showed that it could predict single mutations as well as extrapolate higher-order (up to quadruple) mutations. In their publication (https://www.nature.com/articles/s41587-021-01146-5), they were unable to ascertain how the model works, but they did offer that "the extrapolative performance should depend on how much epistasis contributes to the fitness landscape, and also on the ability of a model to capture such epistasis. The poor extrapolative performance on UBE4B may also stem from its evolutionary data not providing as much relevant information to the property of interest (that is, the assayed property)."

Given this backdrop, we agree with you that the zero-shot predictor is indeed helping. The general idea of augmenting the linear regression model with a zero-shot predictor is to enhance its ability to predict unseen sequence space by capturing evolutionary (or other non-assayed) data. Conceptually, one could imagine the following scenario of epistasis. Two residues that, when mutated individually, are systematically detrimental to the measured activity of a reaction of interest. However, in the evolutionary trajectory of the enzyme, there was an instance where mutating these two residues simultaneously resulted in a functional enzyme, as captured by a number of sequences enriched in these mutations. Thus, the zero-shot predictor can predict a functional enzyme when these mutations are made, despite the only labeled data being negative. We believe this is the general core principle of why augmented ridge regression performs as well as it does against very complex models. It doesn't mean that the model "understands" epistasis, as it is a linear regression model, but that also doesn't prevent the model from finding areas of sequence space where epistasis is present.

As you mention, we would expect positive epistasis to be rare. However, without knowing the full 20^4 combinatorial landscape, and being unable to access this space of 160,000 mutants experimentally in a time and cost dependent manner, we feel it is not possible to speculate if most beneficial interactions will just be additive or if epistasis is more dominant in the landscape. In the revised manuscript, we have decided to not make claims that we do not have data to support. However, we have re-emphasized that the models do help predict mutants with improved activity over wildtype, which was the goal of our study.

Specifically, we write on p. 11:

> *"These results show that our ML-guided strategy can discover high fitness variants for a variety of molecules using the same starting enzyme. While it is unclear if the model can directly infer nonlinear interactions, we found that it is able avoid path dependencies and reduce the screening burden."*

- L247-248: "The evaluated augmented models outperformed the ridge regression model alone." Looking at Fig. S15, I would argue that this is not consistently the case and depends on the reaction and performance metric. E.g. for Moc, the improvements over the simple model are e.g. minor. For Meto, the simple model outperforms all augmented ones in terms of Spearman

ranking. The authors should put this claim into perspective and provide error bars for the performance metrics in Fig. S15 for the different training/validation splits performed. The latter would allow to evaluate whether the improvements are within error or not.

We thank you for the suggestion to put our statements into perspective. In the revised manuscript, we now describe how the training/validations splits were performed. Specifically, to clarify Figure S15, we have adjusted the caption to state that:

> *"Model performance was evaluated by training a model on single mutant data and testing on the withheld combinatorial data (double, triple, and quadruple mutants) from iterative site saturation mutagenesis."*

The training set was initially split into train/validation sets for hyperparameter tuning, but ultimately the model was retrained on the entire single mutant dataset after tuning. This prevents error bars on performance metrics as the test and train sets are held constant. To additionally clarify the manuscript, we revised the results to state on p. 10:

> *"The augmented models outperformed the ridge regression model alone when evaluated with NDCG."*

A similar observation can be made here (https://www.nature.com/articles/s41587-021-01146-5); occasionally, a simple model with no zero shot predictor outperforms the augmented model when scoring by spearman ranking. We hypothesize that in our case, spearman correlation is much higher than NDCG due to the nature of the test set containing double, triple, and quadruple mutants. While the quadruple mutants are enriched in activity, they are also the most likely to experience epistasis (something we notice in several instances). Presumably, the simple model may be able to readily rank the lower order double and triple mutants, but struggle with the quadruple mutants. Given that NDCG biases models that can score the top predictions accurately, this could cause the simple model to relatively underperform in NDCG.

- Design of new variants:

- the improvements over top variants from ISM ("M4") are rather small in Fig. 4 C+D. This is surprising because it implies that the sequence space of 20^4 variants does not contain any significantly better variants than the ones found in ISM (which covers <1% of total space). For Moc, only three variants with miniscule improvements (probably within error) were found. For meto, most of the designed variants are inactive. Crucially, comparing the ranks in Fig. 4C-E with actual performance of the best-predicted variants, there seems to be little rank correlation. The authors are encouraged to report the ranking correlations also for these designed variants and put the claims made in terms of model performance in perspective.

We thank you for giving us the chance to clarify our claims. Our strategy for implementing a ML-guided approach was not to be able to accurately predict the entire fitness landscape, but to simply find mutants with superior activity over wildtype without having to comprehensively

sample the fitness landscape. This rationale is why we opted to use NDCG to select a model instead of spearman correlation, which may have given predictions that correlate better with experimental results. Our results in Fig. 4C-E were meant to convey this, but we agree that additionally adding spearman correlation is important in understanding this. In the revised manuscript, we put this into context in the revised manuscript by stating on p. 11:

> *"However, the rank of experimentally tested predictions correlates poorly with predicted rank (average Spearman's rank correlation of 0.21 ± 0.15), indicating the models may not accurately capture the entire sequence-fitness landscape."*

We think of this in a similar way to traditional directed evolution campaigns. Most directed evolution campaigns are not necessarily capable of finding the global maximum, but just want an improved variant quickly and economically, and our goal was to provide another route to do so, but without the significant time and cost associated with comprehensive sampling. The fact that our framework could predict more active mutants with a lower screening burden than iterative site saturation mutagenesis for 9 small-molecule pharmaceuticals shows how our goal was achieved.

- Generalizability of the model across the sequence space: This is strongly related to the point before. Looking more closely at top predictions, one finds that the variants have a very high sequence similarity, i.e. are very close in sequence space to each other. Examples (see Tab. S1):
--> Moc: M4 = SSFL, and the three slightly better variants are SSFV, SSFI, SSFT.
--> Meto: M4 = SCFT, and the only better variant is SCFS (SAFS and SSFT are close as well)
In other words, the predictions deviate often only by one mutation from the ISM path. The found replacements are often conservative ("trivial") (see above, L to I or T to S; the later is highlighted as "surprising finding" in the main text). Further, there is a strong tendency of the model to suggest amino acids found to be best/good in the single-site screens. These are strong indications that there may be strong overfitting to small regions of the entire sequence space close to the training data/ISM path. A way to test this would be to construct a set of variants that is further away and uniformly distributed in sequence space, and then test the models' performance on these variants.

We agree with you that it is hard to make claims about generalizability across sequence space. We hypothesize that the high sequence similarity is a combination of suggesting amino acids that performed well in the training data as well as EVmutation limiting the search space around what is evolutionarily acceptable. In the latter, it makes sense that mutations would be limited to more conservative mutations, as those evolutionary substitutions are often similar in sequence space.

You make an excellent suggestion to construct variants further away in sequence space, but it is difficult to make an informed choice on these without *a priori* information on the entire combinatorial space. Based on our comprehensive ISM data with moclobemide, metoclopramide, and cinchocaine (Figures 3, S7, S13, and S14), it is likely that most of these

mutants farther in sequence space are inactive or low activity, and it would be hard to reasonably explore 20^4 in a meaningful manner. For example, of the 1,216 moclobemide mutants, we found that 1,155 were less active than wildtype, and this is only considering single mutations. We think that avoiding any claims about generalizability across the entire sequence space and focusing on using the models to predict high activity (not highest activity) mutants is in line with the data we present.

In support of our claim that an A to L mutation was a surprising finding (see: “The best predicted variant for cinchocaine had significantly higher activity than the best single mutation on its own and surprisingly contained a mutation (A205L) that decreased activity compared to wt-McbA in the HSS”, we would like to point out that our data in the hot spot screen indicated that A205L as a single mutant was detrimental to activity. Even more so, A205F was shown to have a higher activity than wild type. It was therefore surprising that in the context of three other mutations that A205L now becomes beneficial for activity instead of detrimental. While we agree that differences between alanine and leucine are small on an amino acid level, the data we present suggest otherwise on a functional protein level.

In the revised manuscript, we removed the claim that we “comprehensively mapped” the sequence-fitness landscape so as to avoid any exaggeration and replaced it with the following statement on p. 13:

> *“Through efforts to build ML models and all ISM rounds, we explored the sequence-function landscape of McbA by assessing 2,856 variants of McbA, 1,100 possible amide products, and 12,584 substrate pair-mutant reactions.”*

In addition, we directly added the following statement in the discussion on p. 14 about possible constraints of generalizability of the model across the sequence space:

> *“There may be instances where more complex models (e.g., random forests, support vector machines, neural networks, etc.) that can generalize better to the entire sequence space will be necessary to navigate complex fitness landscapes, especially where more data is collected for training.”*

With the presented results at hand, I would argue a justified claim would be that the ML models aid the search around hypothetical ISM trajectories and thus might help to reduce effort and increases success rates. But good generalizability across the entire 20^4 space it is not tested for and seems highly unlikely given the high sequence similarity amongst predicted top variants.

We agree with you and thank you for the suggestion. We removed any claims that would indicate good generalizability across the entire combinatorial space in favor of searching around hypothetical ISM trajectories. See statements above. We also added the following statement in the revised manuscript on p. 14:

"…the ML-guided framework used here aided in the search around hypothetical ISM trajectories, reducing effort and increasing success rates in enzyme engineering campaigns."

- Designed variants for six additional compounds: some of the improvements found are nice, but a comparison to ISM is missing, which unfortunately does not allow to make any conclusions on the "effectiveness" of the ML-guided search over the conventional approach in these cases. Moreover, there is a high degree of sequence similarity amongst predicted top variants also in these cases.

We thank you for raising this point of clarification, which helped us realize we needed to do a better job explaining our objectives in predicting six additional compounds. The key idea was that we wanted to show we could get higher mutants by simply performing experiments with 80 sequences per compound using the ML model and a lower screening burden.

In the paper, we first set out to develop a suitable model framework and demonstrate the efficacy of the method. For this, we carried out three comprehensive ISM campaigns on moclobemide, metoclopramide, and cinchocaine. Then, we wished to apply our validated method to diversify McbA mutants for an additional 6 compounds as a test to demonstrate that our method can be used for parallelized engineering of multiple reactions simultaneously to expedite enzyme engineering campaigns. In both cases, ML predictions were able to capture some of the top performing mutants in the ISM path and some predicted mutants that outperformed them, giving us confidence in the ability of our framework to robustly accelerate enzyme engineering campaigns. To clarify this rationale for including the six additional compounds without comparison to ISM, we now state in the discussion on p. 13:

"This framework uniquely integrates a cell-free gene expression and mutagenesis method, ML to expedite directed evolution campaigns, and divergent evolution to convert a generalist enzyme into multiple specialists. We showcased this framework by rapidly navigating nine protein engineering campaigns for the amide synthetase McbA, six of which were performed simultaneously."

On the effectiveness of the ML-guided approach, our ISM campaigns for moclobemide and metoclopramide required multiple additional engineering rounds and testing 646 and 475 unique mutants (adding months of time), whereas our ML-guided approach only required one additional round and testing 25 unique mutants each. Additionally, our cinchocaine ISM campaigns resulted in two "dead-ends", meaning we could not find a beneficial mutation after the second round of ISM. Our ML-guided approach enabled us to overcome these dead-ends and find high activity mutants.

- Discussion:
- L341-344: The claim "…we comprehensively mapped the sequence-function landscape of McbA" is unjustified, since only four residues of McbA were looked at and even within the 20^4

subspace generalizability remains undemonstrated (see comment above on generalizability and overfitting).
- L352-356: The claims made in this part are in my opinion a strong exaggeration of the actual findings. It is merely speculative and likely not true that the entire space of 20^N was comprehensively mapped. All tested variants are very close in sequence space and generalizability is not tested for.

We agree with you and have moderated comments made in the discussion. Specifically, the manuscript now states on p. 13:

> *"Through efforts to build ML models and all ISM rounds, we explored the sequence-function landscape of McbA by assessing 2,856 variants of McbA, 1,100 possible amide products, and 12,584 substrate pair-mutant reactions."*

We further apologize that our claims came off as an exaggeration and agree that the entire space of 20^4 was not comprehensively mapped (nor can the models accurately predict all 20^4). We only intend to claim that our approach lowers the experimental burden and time required to find a mutant with improved activity over wild type. Towards this, we edited the manuscript to remove the reference to the speculative 20^N statement.

• Minor points
- Generally, the figures/captions lack critical information. Few examples are: what the training and test data were for models; quantitative information on activity metrics; what the displayed different activity metrics area and whether/how they are normalized (e.g. Fig. 4A/B); for which encoding strategy the model performance metrics are shown in Fig. 3A etc.). This should be carefully revised to facilitate understanding for the readers.

- Typo in Fig. 1: "accesible"

We thank you for noticing this typo and have corrected it.

- It would be helpful to add the ATP concentration in Fig. 1A

We agree with you and have added the correct concentration.

- Fig. 2C: It is not specified how high the conversion was for these compounds (only a range is given in the main text). This info should be added to the figure.

We thank you for this suggestion. In the revised manuscript we now state that:

> *"We observed that McbA was able to synthesize 11 pharmaceutical compounds as well as dozens of hybrid molecules (Fig. 2C), ranging from trace amounts detectable only by mass spectrometry (MS) to approximately 12% conversion."*

When reactions are only detectable by mass spec it is not possible to estimate % conversion and we are not able to generate standard curves for many of these hybrid pharmaceutical molecules. We intend for this data to be qualitative and guide future engineering efforts of McbA.

- Fig. 3a: Should be 64 x 19 = 1216 mutants (instead of 1280) for the HSS?

We thank you for noticing this and have corrected it to say 1,216 single mutants. We initially stated 1,280 as that is the number of unique reactions. Wild type is overrepresented 63 times but is included and ran as an internal control.

- Fig. 3a: typo “extrpolate”

We have corrected the typo.

- Fig. 3b-d: the colour scale has no label, ticks etc. specifying what the activity metric is here (caption says it should be % conv.)

We thank you for requesting this clarification and have edited the figure caption to clarify.

- It was not clear from Fig. 3 or main text which were the top4 residues selected as training data (n=77) for the initial models. I believe this is somewhere in the SI, but it would be helpful to give this information in the main manuscript.

We have changed the coloring in Fig. 3B-D to represent the top 4 residues that were selected for training and have clarified this in the figure caption.

- L259-263: “the augmented EVmutation model”: upon reading it was not entirely clear whether it is really one model trained for the two reactions (multi-objective learning) or two models trained separately for each reaction. According to Fig. 3A and after further reading, I’m quite sure it is the later, but it would help understanding to make this clear early on.

We thank you for requesting this clarification. In the updated manuscript, we added on p. 10:

> *“Using our trained ML models (one single-objective model per compound), we screened 20^4 combinatorial enzyme variants for the synthesis of moclobemide, metoclopramide, and cinchocaine in silico and selected the top 25 predictions to subsequently build and test.”*

Reviewer #1 (Remarks to the Author):

1. The authors have still not fully described how the HSS (particularly the active site selection process) is performed. Please describe the methodology used to identify active sites, including any criteria or algorithms employed.

There were no criteria or algorithms employed to identify active sites, we relied on previous structural characterization of McbA that had determined the active site. The solved crystal structure of McbA contains the native substrates that McbA utilizes (thus capturing the active site). We used this information to identify the residues around those substrates.

On page 7 we state, *"Guided by the crystal structure of McbA (PDB: 6SQ8), we selected 64 residues that completely enclosed the active site and putative substrate tunnels (e.g., residues within 10 Å of the docked native substrates)."*

2. The GitHub page lacks a comprehensive Readme file to guide readers through the project. Although the authors have provided all necessary raw data, source code, and an environment.yml file, there is no overview of the entire workflow.

We have updated the readme file of the GitHub repository to explain the contents of the repository and how they can be used. In combination with the Jupyter notebooks that guide users through every step of our process (including model selection and use) we have provided all resources necessary to recreate the results shown in the manuscript.

Reviewer #1 (Remarks on code availability):

The GitHub page lacks a comprehensive Readme file to guide readers through the project. Although the authors have provided all necessary raw data, source code, and an environment.yml file, there is no overview of the entire workflow.

We have updated the readme file of the GitHub repository to explain the contents of the repository and how they can be used. In combination with the Jupyter notebooks that guide users through every step of our process (including model selection and use) we have provided all resources necessary to recreate the results shown in the manuscript.

Reviewer #2 (Remarks to the Author):

The authors have addressed all of my concerns.

Reviewer #3 (Remarks to the Author):

The authors have added critical information and analyses during the revision of their manuscript. Further, they have addressed my previous concerns by discussing some of the main claims and/or putting them into perspective. Thus, I am supportive of publication.