nature portfolio

Peer Review File

A Generalized Platform for Artificial Intelligence-powered Autonomous Enzyme Engineering

Corresponding Author: Professor Huimin Zhao

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)
This manuscript presents a significant advance in automated enzyme engineering by integrating AI-guided design with in-house laboratory automation. Unlike previous approaches relying on commercial cloud laboratories, this system provides complete transparency and control over the experimental process, successfully improving catalytic activity and substrate specificity by 16-fold and 26-fold for two different enzymes within four rounds of evolution. The platform effectively combines multiple AI components with automated experimental workflows, featuring a high-accuracy HiFi assembly-based mutagenesis method and modular automation design that enables error recovery. The comprehensive kinetic characterization of variants provides strong validation of the improvements.

However, the manuscript would benefit from addressing several key points:

The title and content should be revised to specifically focus on "enzyme engineering" rather than "protein engineering," as this better reflects the scope and validated capabilities of the work. Similarly, claims about broad applicability ("virtually any protein") should be moderated. Many proteins lack such convenient expression systems or function assays as the two enzymes studied here.

Claims about system autonomy should be tempered. While the platform successfully automates many steps and incorporates AI guidance, true autonomy would require additional capabilities such as automated protocol generation, dynamic optimization of biofoundry workflows, and automated data analysis and decision-making. For example, were the worklist Python scripts written by human or an llm agent?

The manuscript states: "Variants with multiple mutations can be generated through SDM of a template plasmid containing one fewer mutation and, as such, there is no need to order new primers for each iterative cycle." While this approach simplifies primer design, it has important limitations. The HiFi assembly method cannot efficiently generate variants with multiple mutations in adjacent residues, which is particularly limiting for engineering active sites. The authors' own results support this limitation - when their model suggested triple mutants that ignored the previously beneficial S99T/V140T combination, these novel combinations outperformed the intuitive stepwise approach, highlighting potential benefits of exploring more diverse mutation combinations.

The supervised low-N ML model shows relatively weak correlation with experimental results (ρ = 0.1-0.34). While the system still succeeded in identifying improved variants, this disconnect between predictive accuracy and practical effectiveness deserves deeper investigation.

More rigorous quantitative comparisons with existing automated protein engineering systems (ref 17, 40) in tabular form, including automation systems, prediction and design algorithms, library construction, expression and screening, improvement, iterative rounds, would help readers better understand the advances and trade-offs of this approach.

Minor point:
In Line 230, it seems that reference 28 was misquoted in "we developed a natural language-based user interface utilizing LLMs (28)".

(Remarks on code availability)
The results of the paper are reproducible and the code is a usable resource for the community

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)
Dear Editor,
I am writing to provide my feedback on the manuscript under consideration. While I believe the paper shows promise, I would like to outline several major points that need clarification or elaboration before publication.

General Impressions:
The authors have made a commendable effort to enhance the readability of their manuscript for non-computational readers, which broadens the accessibility of their work. However, this approach comes at the expense of critical technical details, especially concerning the computational methods employed. Methods and supplementary sections could have been extended and enriched. This lack of detail diminishes the ability of readers to fully appreciate the contributions of the work and assess its novelty.
Key Concerns:

1)Novelty:
- While the computational methods and the automated platform themselves are not new, the novelty appears to lie in their combination. Additionally, the use of a GPT-based frontend is presented as a key feature to allow researchers without coding expertise to access the platform. However, the level of real automation achieved is unclear.
- Specifically, the manuscript suggests that researchers must still make case-specific decisions (e.g., model selection and design choices tailored to the specific proteins being studied). It remains uncertain whether these decisions would generalize effectively to other proteins, particularly in cases with different evolutionary constraints or mutational landscapes. For instance, the unsupervised models employed (EV-mutation model and ESM-2) may not generalize to all proteins due to specific limitations:
EV-mutation is MSA-based and depends on the availability of homologous sequences.
ESM-2 as a zero-shot predictor has been shown to correlate poorly with experimental data in certain benchmarks (e.g., ProteinGym [1]).
- For these reasons, the starting library of single mutants with a low number of mutants could not always be sufficient as a starting point for optimization, and for certain proteins they might fail.
- The authors should clarify the extent to which the platform is automated and whether the approach would work universally or only under specific conditions. Claims such as “virtually any protein” should be tempered or justified.

2) Lack of Benchmarking and Contextualization:
The manuscript lacks benchmarks for the computational results, making it difficult to evaluate the significance of the proposed method.
A state-of-the-art section discussing the evolution of the specific proteins studied in this work (AtHMT and YmPhytase) is missing. Without this, readers cannot fully understand the improvement offered by the proposed method relative to prior approaches.

3) Details on Computational Methods:
The choice to exclude probability features in the low-N model from [2] is only briefly justified. While the authors argue that these features negatively correlate with experimental results, they could still offer value in exploring novel positions or mutations. Using just the one-hot encoding representation of the sequence has as a consequence that the only mutations and positions that can be explored are those that were already present in previous rounds. This seems very restrictive, especially when ESM-2 and EV-mutation are not good at correctly approximating the probability distribution of the protein, and are making a poor starting library of mutants.

4) Concerns Regarding Epistasis:
The approach starts with a library of single-point mutants. However, if strong epistatic effects are present, this may result in suboptimal predictions and limited optimization of protein function. The manuscript does not adequately address this possibility.
The two examples presented in the manuscript should be explicitly shown to not be cherry-picked cases where epistasis is low. The authors should provide evidence or arguments to demonstrate the generalizability of their approach across proteins with varying degrees of epistasis.

5) Figures and Supplementary Material:
Figures 3 and 4 are underdeveloped and lack sufficient detail. These figures are central to understanding the results and should be expanded upon.
Additional methodological details could be included in the supplementary material to address the concerns raised above.

[1] Notin, Pascal, et al. "Proteingym: Large-scale benchmarks for protein fitness prediction and design." Advances in Neural Information Processing Systems 36 (2024).
[2] Hsu, Chloe, et al. "Learning protein fitness models from evolutionary and assay-labeled data." Nature biotechnology 40.7 (2022): 1114-1122.

(Remarks on code availability)
Computational methods use in the paper are standard in the field and in my opinion do not require revision

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have adequately addressed all my comments.

(Remarks on code availability)
The results of the paper are reproducible and the code is a usable resource for the community

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)
I would like to thank the authors for addressing the previous concerns, clearly outlining the limitations of their approach, and enhancing the clarity of the manuscript—particularly through improved figure readability. However, I remain cautious about the significance of the experimental validation, as the number of improved proteins (n = 2) remains too low to draw robust conclusions. Moreover, since these proteins were specifically selected to avoid overlap with machine learning-optimized cases, this makes the comparison more challenging for alternative methods. While this selection criterion is understandable and provides a useful use case, it limits generalizability. For example, including additional proteins - as those mentioned in the rebuttal - would significantly strengthen the experimental section and provide a more convincing demonstration of the method’s utility.

(Remarks on code availability)

Reviewer #1 (Remarks to the Author):

This manuscript presents a significant advance in automated enzyme engineering by integrating AI-guided design with in-house laboratory automation. Unlike previous approaches relying on commercial cloud laboratories, this system provides complete transparency and control over the experimental process, successfully improving catalytic activity and substrate specificity by 16-fold and 26-fold for two different enzymes within four rounds of evolution. The platform effectively combines multiple AI components with automated experimental workflows, featuring a high-accuracy HiFi assembly-based mutagenesis method and modular automation design that enables error recovery. The comprehensive kinetic characterization of variants provides strong validation of the improvements.

Response: We express our gratitude to the reviewer for the positive comments and valuable feedback.

However, the manuscript would benefit from addressing several key points: The title and content should be revised to specifically focus on "enzyme engineering" rather than "protein engineering," as this better reflects the scope and validated capabilities of the work. Similarly, claims about broad applicability ("virtually any protein") should be moderated. Many proteins lack such convenient expression systems or function assays as the two enzymes studied here.

Response: Based on the reviewer's comments, we have modified the title of the manuscript and made changes throughout the text to better reflect the scope of our work. Also, we have modified the usage of "virtually any protein" to "a wide array of proteins."

In addition, we have added an additional paragraph to the discussion section to address the limitations of this work:
"Nonetheless, our platform still contains several limitations. On the computational side, while we use ESM-2 and EVmutation for initial variant predictions, ESM-2 has performed poorly in some benchmarks and may not be suitable for all proteins[42] and EVmutation depends on the availability of homologous sequences[23]. Applying our platform to proteins that are not well-suited for these models may result in poorly-performing initial variant libraries. Moreover, the supervised low-N model used for iterative predictions showed relatively weak correlation with experimental results, suggesting the possibility that certain proteins may perform particularly poorly using such a predictive model. Our site-directed mutagenesis method employing PCR and HiFi assembly to accumulate mutations in a stepwise manner also contains some limitations. Recycling of mutagenesis primers limits the implementation of simultaneous multiple mutations in adjacent residues. As such, implementing multiple mutations in adjacent residues may require the design and purchase of new mutagenesis primers which contain desired mutations. High GC-content sequences may also lead to difficulty with PCR amplification and HiFi assembly. Additionally, our platform requires a functional assay that is compatible with high throughput automation. Here, we focused specifically on enzymes with established high throughput quantitative assays that are functional in crude cell lysates. The suitability of our platform for transporters, chaperones, transcription factors, and other types of proteins remains unexplored at this time and would require the development of suitable functional assays."

Claims about system autonomy should be tempered. While the platform successfully automates many steps and incorporates AI guidance, true autonomy would require additional capabilities such as automated protocol generation, dynamic optimization of biofoundry workflows, and automated data analysis and decision-making. For example, were the worklist Python scripts written by human or an llm agent?

Response: We agree with the reviewer that true autonomy is a difficult task for biological experiments and complex workflows. Complex experimentation and random errors arising from numerous unpredictable issues make it challenging to design a truly autonomous experimentation platform. We would also like to

mention that previous manuscripts attempting automated protein engineering on a much smaller scale have described their system as “fully autonomous protein engineering”. (Rapp et al., *Nature Chemical Engineering*, 2024)

Additionally, following the suggestion by the reviewer, we have categorized our level of automation using one of the standards established in a recent paper (Physical Laboratory Automation in Synthetic Biology, Stephenson *et al., ACS Synth. Biol.* 2023). Based on our assessment, the platform’s level of automation falls between level 3 (conditional automation) and 4 (high automation). We have added following text in the discussion section:
“Following one of the frameworks developed to evaluate levels of automation[20], we assess that our current autonomous enzyme engineering platform falls between level 3 (conditional automation) and level 4 (high automation), as it can perform end-to-end protein engineering with minimal human intervention and predict variants for screening in subsequent cycles.”

Finally, we have occasionally replaced “autonomous” with “automated” to moderate the overall tone of the manuscript. Our platform, though designed by humans, is capable of running continuous automated experiments within each module of enzyme engineering. Additionally, the recommended variants for each iterative round of engineering are provided by AI/ML models, which do not require human decision making. The Python scripts, though written by humans, do not need to be updated for each run and thus help create a continuous workflow with minimal human intervention.

The manuscript states: "Variants with multiple mutations can be generated through SDM of a template plasmid containing one fewer mutation and, as such, there is no need to order new primers for each iterative cycle." While this approach simplifies primer design, it has important limitations. The HiFi assembly method cannot efficiently generate variants with multiple mutations in adjacent residues, which is particularly limiting for engineering active sites. The authors' own results support this limitation - when their model suggested triple mutants that ignored the previously beneficial S99T/V140T combination, these novel combinations outperformed the intuitive stepwise approach, highlighting potential benefits of exploring more diverse mutation combinations.

Response: We agree with the reviewer that a HiFi-based site-directed mutagenesis approach has its inherent limitations. However, we aim to recycle primers as much as possible to reduce both cost and time. Since we introduce only one mutation per cycle, multiple adjacent mutations over multiple screening cycles can be addressed using a Python script to design primers for such specific cases. Nonetheless, there could still be a few edge cases that will require a more customized approach or the HiFi assembly may fail entirely. The supervised learning algorithm can tolerate an error of 5-10% and thus a few errors in mutagenesis should not affect the efficiency of the workflow. A Python script to find such cases and design the primers will address this issue.

While we acknowledge that the S99T/V140T results highlight the need to explore diverse mutation combinations, all triple mutants predicted by the model were compatible with our stepwise approach. In other words, all predicted triple mutants could be created using the existing primers along with a double mutant from the previous round. Any potential issues with multiple adjacent mutations could be resolved using an additional Python script, new primers, and updated worklists. This approach helps streamline the experimental workflow for generating mutations via HiFi assembly.

Following reviewer’s suggestion, we have acknowledged potential issues with PCR and HiFi assembly based SDM in the discussion section.
“Our site-directed mutagenesis method employing PCR and HiFi assembly to accumulate mutations in a stepwise manner also contains some limitations. Recycling of mutagenesis primers limits the implementation of simultaneous multiple mutations in adjacent residues. As such, implementing multiple

mutations in adjacent residues may require the design and purchase of new mutagenesis primers which contain desired mutations. High GC-content sequences may also lead to difficulty with PCR amplification and HiFi assembly."

The supervised low-N ML model shows relatively weak correlation with experimental results (ρ = 0.1-0.34). While the system still succeeded in identifying improved variants, this disconnect between predictive accuracy and practical effectiveness deserves deeper investigation.

More rigorous quantitative comparisons with existing automated protein engineering systems (ref 17, 40) in tabular form, including automation systems, prediction and design algorithms, library construction, expression and screening, improvement, iterative rounds, would help readers better understand the advances and trade-offs of this approach.

Response: The weak correlation is also observed in other computation-focused studies, such as 'Learning protein fitness models from evolutionary and assay-labeled data' by Hsu et. al. (Hsu, et al. "Learning protein fitness models from evolutionary and assay-labeled data." *Nature Biotechnology* 40, 1114-1122 (2022)) and ProteinGym by Notin et. al. (Notin, et al. "Proteingym: Large-scale benchmarks for protein fitness prediction and design." *Advances in Neural Information Processing Systems* 36 (2024)). Since we used a greedy acquisition approach to minimize the number of samples per round as well as the total number of rounds, the most important metric for us is not the correlation but the ranking accuracy of top regime, which is demonstrated by the observed practical effectiveness.

Additionally, it is beyond the scope of this paper to have a quantitative comparison with refs 17, 40 (original numbering, current refs: 11, 40). Since the acquisition method is different, comparing the dataset we obtained in our paper to those prior references would not be fair or rigorous. The prior report of 'Benchmarking uncertainty quantification for protein engineering' by Greenman et. al. (*PLoS Comput Biol* 21: e1012639 (2025)) demonstrates greedy is a very strong baseline.

Following the reviewer's suggestion, we have created a table summarizing refs. 17, 19, and 40 (original numbering, current refs: 11, 13, and 40) and qualitatively comparing it with our work in the Supplementary Information (Supplementary Table 7). It has also been referenced in discussion section.
"Supplementary Table 7 provides a concise comparison with prior automated protein engineering studies."

Minor point:

In Line 230, it seems that reference 28 was misquoted in "we developed a natural language-based user interface utilizing LLMs (28)".

Response: Thank you for pointing it out, we have corrected the misquote.

Reviewer #1 (Remarks on code availability):

The results of the paper are reproducible and the code is a usable resource for the community

Response: We express our gratitude to the reviewer for the positive comments and valuable feedback.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We express our gratitude to the reviewer for the positive comments and valuable feedback.

Reviewer #2 (Remarks on code availability):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We express our gratitude to the reviewer for the positive comments and valuable feedback.

Reviewer #3 (Remarks to the Author):

Dear Editor and the authors,

I am writing to provide my feedback on the manuscript under consideration. While I believe the paper shows promise, I would like to outline several major points that need clarification or elaboration before publication.

General Impressions:
The authors have made a commendable effort to enhance the readability of their manuscript for non-computational readers, which broadens the accessibility of their work. However, this approach comes at the expense of critical technical details, especially concerning the computational methods employed. Methods and supplementary sections could have been extended and enriched. This lack of detail diminishes the ability of readers to fully appreciate the contributions of the work and assess its novelty.

Response: We appreciate the positive comments regarding the accessibility of our report. The computational methods used in this study for mutation prediction (ESM-2, EVmutation, and low-N supervised learning) are previously published methods with detailed technical documentation available with the original publications. We have referenced these methods in the manuscript and interested readers can access the original publications for further reading. Any modification of the original code is clearly documented in the indicated GitHub page and the exact python commands were included in our manuscript as well as in the Github readme file.

In this work, we aimed to establish a complete and integrated platform for automating the protein engineering process using protein predictive models and robotic biofoundries. Since we have not developed new computational methods for mutation prediction, we would refer readers to the original publications for critical technical details such as model benchmarking and implementation.

Following the reviewer's suggestions, we have revisited our experimental methods section and ensured that all technical details are well articulated and provided in sufficient details in the Supplementary Information. We have also added additional text to figure legends to ensure these contain critical technical details or refer readers to relevant sections of the Supplementary Information for more information. The automation workflow is comprehensively documented in Supplementary Fig. 4-8. The Python scripts we developed for this work are available from GitHub with detailed implementation instructions. We have re-organized Figure 2 to convey more details about the workflow. Previous supplementary Figures 2 & 3 have been combined with main Figure 2 and re-organized to create the new Figure 2 for more comprehensive description of the system.

Key Concerns:

1) Novelty: - While the computational methods and the automated platform themselves are not new, the novelty appears to lie in their combination. Additionally, the use of a GPT-based frontend is presented as a key feature to allow researchers without coding expertise to access the platform. However, the level of real automation achieved is unclear.

Response: Following the suggestion by the reviewer, we have added more details in the results section to make the level of automation better explained.
"Seven automated modules comprise the end-to-end workflow for each cycle of protein engineering. Experimental steps within a module are completely automated, requiring little, if any, human intervention (Supplementary Figs. 2-6). Each module was individually programmed onto the iBioFAB platform and meticulously refined to ensure robust and reliable operation during continuous automated execution. Such examples include preparation of mutagenesis PCR, DpnI digestion, 96-well microbial transformations, plating on 8-well omnitray LB plates, crude cell lysate removal from 96-well plates, and functional enzyme assays. For automated execution, the instruments were scheduled via Thermo Momentum software and fully integrated by a central robotic arm."

We have also re-organized Figure 2 to convey more details about the workflow. Previous supplementary Figures 2 & 3 have been combined with main Figure 2 and re-organized to create the new Figure 2 for more comprehensive description of the system.

We have also followed the criteria in a recent paper (Physical Laboratory Automation in Synthetic Biology, Stephenson *et al., ACS Synth. Biol.* 2023) to add additional text to clarify the level of automation of this platform. Overall, our platform, though designed by humans, can run continuous automated experiments for enzyme engineering. From our assessment, our platform's level of automation falls somewhere between level 3 (conditional automation) and 4 (high automation). This is mainly determined by the fact that individual modules run fully automated but may require human intervention between modules. We have also added the following text to the discussion section to address the reviewer's comment.
"Following one of the frameworks developed to evaluate levels of automation[20], we assess that our current autonomous enzyme engineering platform falls between level 3 (conditional automation) and level 4 (high automation), as it can perform end-to-end protein engineering with minimal human intervention and predict variants for screening in subsequent cycles."

- Specifically, the manuscript suggests that researchers must still make case-specific decisions (e.g., model selection and design choices tailored to the specific proteins being studied). It remains uncertain whether these decisions would generalize effectively to other proteins, particularly in cases with different evolutionary constraints or mutational landscapes. For instance, the unsupervised models employed (EV-mutation model and ESM-2) may not generalize to all proteins due to specific limitations: EV-mutation is MSA-based and depends on the availability of homologous sequences. ESM-2 as a zero-shot predictor has been shown to correlate poorly with experimental data in certain benchmarks (e.g., ProteinGym [1]).

Response: We agree with the reviewer, we have added in discussion about limitations of our platform. Please see our response to Reviewer 1 about additional limitations.

The only case-specific decision for the two enzymes in our work was the choice of functional assay. All protein engineering efforts will require a method to quantify the performance of the protein through choice of a functional assay and no single functional assay will be applicable to all enzymes. However, with our workflow, the steps before functional assay remain same and consistent for a wide array of proteins: initial predictions, site-directed mutagenesis, stepwise accumulation of mutations, and supervised learning for variant prediction for subsequent cycles.

Although there is no single protein model that is suitable for all conceivable proteins, our workflow is not tied to a specific model. As new and improved models emerge (e.g. the new protein large language model Evo 2 developed by the Arc Institute), they can be integrated in the workflow for improved sequence-based prediction. Similarly, improved supervised machine learning models can be implemented as they are developed and released.

- For these reasons, the starting library of single mutants with a low number of mutants could not always be sufficient as a starting point for optimization, and for certain proteins they might fail.

Response: We agree that the success rate of the chosen initial prediction model will likely vary depending on protein of choice. However, in our opinion, 200 is not a low number for an initial variant library. Beyond the two proteins presented in this study, we have as of now employed this workflow to three additional enzymes in our lab with considerable success, which will be presented in future publications. In all cases, we observed significant improvement of at least one predicted mutant for each enzyme, which was then used for subsequent iterations.

Additionally, in this work, for both HMT and phytase, the initial library of ~200 variants found enough variants with good improvements in the targeted activity as mentioned in the manuscript.
"We found that 59.6% of *At*HMT and 55% of *Ym*Phytase variants performed above the wild type baseline (Fig. 3c, Supplementary Figs. 7 and 10), with 50% and 23% being significantly better, respectively (two-tailed Student's t-test $p < 0.05$, Supplementary Figs. 7 and 10)."

- The authors should clarify the extent to which the platform is automated and whether the approach would work universally or only under specific conditions. Claims such as "virtually any protein" should be tempered or justified.

Response: We appreciate the reviewer's comment and opportunity to clarify our report. We have added more explicit language about the level of automation employed in our pipeline by referencing a framework proposed by Stephenson et al., (*ACS Syn Bio*, 2023) to qualify the extent of automation. From our assessment the level of automation of this platform falls somewhere between level 3 (conditional automation) and 4 (high automation). We have also added the following text to the discussion section (page 9) to address the reviewer's comment.
"Following one of the frameworks developed to evaluate levels of automation[20], we assess that our current autonomous enzyme engineering platform falls between level 3 (conditional automation) and level 4 (high automation), as it can perform end-to-end protein engineering with minimal human intervention and predict variants for screening in subsequent cycles."

The variant library creation workflow using our high fidelity site-directed mutagenesis method should be more or less universally applicable for a large number of proteins. The functional assay will have to be protein/enzyme specific, which is the case for most protein engineering studies.

We have modified the usage of "virtually any protein" to "a wide array of proteins". We have added an additional paragraph about limitations of our platform. Please see our response to Reviewer 1 about additional limitations.

2) Lack of Benchmarking and Contextualization: The manuscript lacks benchmarks for the computational results, making it difficult to evaluate the significance of the proposed method.

Response: There is no new computational result in this manuscript that requires benchmarking. We are using ESM-2 and EVmutation for predicting an initial variant library, both of which have been extensively benchmarked previously. The low-N supervised learning model has also been previously benchmarked. All

relevant publications have been referenced in the manuscript for interested readers to find details about their development and benchmarking. Our goal in this report was not to develop a new computational tool for mutation prediction, but instead to integrate existing predictive machine learning models for protein engineering with automation and a high-fidelity site-directed mutagenesis method. The integration of these distinct parts leads to a reliable and efficient platform for protein engineering that is compatible with automation.

A state-of-the-art section discussing the evolution of the specific proteins studied in this work (AtHMT and YmPhytase) is missing. Without this, readers cannot fully understand the improvement offered by the proposed method relative to prior approaches.

Response: We have discussed the results of directed evolution for *At*HMT and *Ym*Phytase in the section "**New insights through ML-guided protein engineering**". There are two paragraphs that describe these findings in broader context for both enzymes. The Figure 3A also shows the mutations landscape for both proteins as they evolve through the 4 rounds. We have discussed the structural details of mutations found in most active variants as well as provided a movie showing the best mutations in 3-D structure of these enzymes for better clarity.

For both *At*HMT and *Ym*Phytase, the best variants identified in this study showed remarkable improvement compared to previous reports utilizing rational design approaches [Refs. 1-3, see below]. We did not find any reports using AI/ML models to engineer *At*HMT and *Ym*Phytase enzymes. For *At*HMT, the V140T mutation was identified in a previous report too [1], but none of their other double and triple mutation combinations performed better than V140T alone. In our approach, we have numerous variants that perform better than V140T, which includes majority of triple and quadruple mutants in rounds 3 and 4. For *Ym*Phytase, we identified a new residue V141M, which was not identified by rational design [2,3]. Interestingly, our approach didn't identify the T44/K45 sites, identified in the previous studies [2,3].

We have added the following to the discussion section (Page 9).
"We assayed the previously identified best mutants for *At*HMT (V140T)[25] and *Ym*Phytase (T44V/K45E)[26] alongside the top variants identified in this study. For both *At*HMT and *Ym*Phytase, the most active variants discovered in this study demonstrate a remarkable improvement compared to the previously reported enzymes[25-27]."

1. Tang, Q. *et al.* Directed Evolution of a Halide Methyltransferase Enables Biocatalytic Synthesis of Diverse SAM Analogs. *Angew. Chem. Int. Ed.* **60**, 1524–1527 (2021).
2. Shivange, A. V. *et al.* Directed evolution of a highly active Yersinia mollaretii phytase. *Appl Microbiol Biotechnol* **95**, 405–18 (2012).
3. Körfer, G. *et al.* Directed evolution of an acid Yersinia mollaretii phytase for broadened activity at neutral pH. *Appl. Microbiol. Biotechnol.* **102**, 9607–9620 (2018).

3) Details on Computational Methods: The choice to exclude probability features in the low-N model from [2] is only briefly justified. While the authors argue that these features negatively correlate with experimental results, they could still offer value in exploring novel positions or mutations. Using just the one-hot encoding representation of the sequence has as a consequence that the only mutations and positions that can be explored are those that were already present in previous rounds. This seems very restrictive, especially when ESM-2 and EV-mutation are not good at correctly approximating the probability distribution of the protein, and are making a poor starting library of mutants.

Response: We appreciate the comments but would disagree that our initial variant library is "poor" based on results from the first round of screening. As shown in Figure 3C, the initial library is in fact quite high quality, which is also mentioned in the manuscript.

"We found that 59.6% of *At*HMT and 55% of *Ym*Phytase variants performed above the wild type baseline (Fig. 3c, Supplementary Figs. 7 and 10), with 50% and 23% being significantly better, respectively (two-tailed Student's t-test $p < 0.05$, Supplementary Figs. 7 and 10)."

In addition, two major reasons made us exclude probability features: (1) These features not only negatively correlate with experimental results but adding them also makes the prediction correlation worse. (2) We did in fact explore new variants with these features in the second round of HMT engineering, but none of the newly explored ones showed higher activity than WT (Fig. 3d).

4) Concerns Regarding Epistasis: The approach starts with a library of single-point mutants. However, if strong epistatic effects are present, this may result in suboptimal predictions and limited optimization of protein function. The manuscript does not adequately address this possibility.

Response: We appreciate the thoughtful consideration of the reviewer and hope our response provides some clarifying information. First, EVMutation is a statistical method based on epistasis, thus residues with strong epistatic effects should be captured in the initial library. Secondly, the data in Figure 3e shows that synergetic effects can be captured to some extent. Most importantly, however, is that our manuscript primarily focuses on the integration of biofoundry automation with predictive machine learning models for protein engineering. While our results are surely largely dependent and limited by model performance, the choice of a predictive model will evolve as new methods are developed. In other words, we have strong reasons to anticipate that our automated platform for protein engineering will improve in performance as stronger and more accurate predictive models are released in the coming years.

The two examples presented in the manuscript should be explicitly shown to not be cherry-picked cases where epistasis is low. The authors should provide evidence or arguments to demonstrate the generalizability of their approach across proteins with varying degrees of epistasis.

Response: The criteria we used for picking HMT and phytase for this study were: (1) We wanted to explore enzymes that have sparse reports of being studied using machine learning models, (2) we hoped for our improved variants to have potential for practical applications, and (3) we selected proteins where the performance could be quantified in a high throughput manner using a plate reader assay. We have added following on Page 5:
"Both proteins have rarely been studied using AI/ML tools, offer potential for industrial applications, and are suitable for automation-friendly high-throughput quantification."

These proteins were not pre-selected because of any considerations to their epistasis, and as we are not aware of any publications that have investigated the level of epistasis of HMT and phytase, we could not have pre-selected these for low epistasis. We selected these proteins for their novelty, usefulness, and potential for generating new knowledge. Additionally, we are currently using this platform to successfully improve three more proteins, which will be published in the near future.

The machine learning models used in this study ESM-2, EVmutation, and Low-N supervised learning have been well documented and benchmarked previously, demonstrating their limitations and generalizability. Additionally, as more universal protein LLM models are developed, they can be incorporated into our workflow for a more comprehensive and generalized approach.

5) Figures and Supplementary Material: Figures 3 and 4 are underdeveloped and lack sufficient detail. These figures are central to understanding the results and should be expanded upon. Additional methodological details could be included in the supplementary material to address the concerns raised above.

Response: We appreciate the comment from the reviewer and the opportunity to improve the clarity of our figures. We have added details into Figures 2, 3, and 4 as well as revised the methods section and Supplementary Information to ensure that complete technical details are available. For additional automation details, Supplementary Figures 2-5 provide comprehensive information about each step.

For Figure 3, all the original plot data is provided in the source file. For each subfigure, complete raw data are provided as source data. Additionally, the methods sections describe all protocols that were used to perform the experiments.

For Figure 4, all raw data is provided in the source file. Detailed graphs for Fig. 4c & 4d are provided in Supplementary Figure 7 & 10. For Figure 4e & 4f, all the replicates and the data are provided in source file. Additionally Supplementary Figures 11-26 list detailed and comprehensive information that expand upon the main figures while providing additional depth. Supplementary Tables 1-6 provides all measurements for these figures.

We have also expanded upon the site-mutagenesis method and added the following descriptions:
"We observed that NEB Q5 polymerase can tolerate a wide range of theoretical annealing temperatures, thus all mutagenic PCRs were annealed at 65 °C for ease of automation. Ensuring correct amount of input template DNA was critical to successful PCRs and efficient SDM, thus the DNA concentration was measured using Invitrogen Qubit Fluorometer with dsDNA Quantitation kit (Invitrogen #Q32853) and further verified by agarose gel electrophoresis. The primers were ordered from IDT in 96-well plate at 2 µM. For PCRs, 12.5 µL of primers were added to 96-well PCR plates on Tecan Fluent."

We are confident that the manuscript contains all information needed to repeat or implement our protein engineering platform and that all underlying data is easily accessible by readers.

[1] Notin, Pascal, et al. "Proteingym: Large-scale benchmarks for protein fitness prediction and design." Advances in Neural Information Processing Systems 36 (2024).
[2] Hsu, Chloe, et al. "Learning protein fitness models from evolutionary and assay-labeled data." Nature biotechnology 40.7 (2022): 1114-1122.

Reviewer #3 (Remarks on code availability):
Computational methods use in the paper are standard in the field and in my opinion do not require revision

Response: We express our gratitude to the reviewer for the positive comments and valuable feedback.

Reviewer #4 (Remarks to the Author):
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We express our gratitude to the reviewer for the positive comments and valuable feedback.

Reviewer #1 (Remarks to the Author):
The authors have adequately addressed all my comments.
Response: We thank the reviewer again for his/her constructive comments.

Reviewer #1 (Remarks on code availability):
The results of the paper are reproducible and the code is a usable resource for the community

Reviewer #2 (Remarks to the Author):
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #2 (Remarks on code availability):
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks to the Author):
I would like to thank the authors for addressing the previous concerns, clearly outlining the limitations of their approach, and enhancing the clarity of the manuscript—particularly through improved figure readability. However, I remain cautious about the significance of the experimental validation, as the number of improved proteins (n = 2) remains too low to draw robust conclusions. Moreover, since these proteins were specifically selected to avoid overlap with machine learning-optimized cases, this makes the comparison more challenging for alternative methods. While this selection criterion is understandable and provides a useful use case, it limits generalizability. For example, including additional proteins - as those mentioned in the rebuttal - would significantly strengthen the experimental section and provide a more convincing demonstration of the method's utility.
Response: For the variant prediction using unsupervised to supervised AI/ML models, we have used state-of-the-art and extensively validated models including ESM-2, EVmutation, and low-N supervised learning. These methods have been extensively benchmarked against many proteins. Therefore, we do not believe that doing experiments on already benchmarked proteins would add anything further to existing literature and create new knowledge. The wide applicability of the platform arises from the seamless and robust integration of AI/ML tools and robotic experimentation for enzyme engineering. We have successfully used HMT and Phytase as test cases to demonstrate that the autonomous platform resulted in significant targeted activity improvements for both enzymes. We are currently working on applying this platform to engineer additional proteins, which are their own projects and combining them into this manuscript would not be ideal. In addition, we will offer this autonomous enzyme engineering platform as a service to the broad research community through the newly established NSF iBioFoundry (https://ibiofoundry.illinois.edu), which should further demonstrate its general applicability.