Informed Training Set Design Enables Efficient Machine Learning-Assisted Directed Protein Evolution

Bruce J. Wittmann, Yisong Yue, & Frances H. Arnold

Summary

Initial Submission: Received Dec. 15, 2020
Deposited on bioRxiv, Dec. 4, 2020
Scientific editor: Quincey Justman, Ph.D.

First round of review: Number of reviewers: Three
Three confidential, zero signed
Revision invited Feb. 2, 2021
Major changes anticipated
Revision received May 6, 2021

Second round of review: Number of reviewers: Three
Three original, zero new
Three confidential, zero signed
Accepted Jul. 26, 2021

Data freely available: Yes
Code freely available: Yes

This Transparent Peer Review Record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor’s letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.
Editorial decision letter with reviewers’ comments, first round of review

Dear Dr. Arnold,

I hope this email finds you well. The reviews of your manuscript are back and I’ve appended them below. On balance, the reviewers appreciate the goals of the work presented here; they’ve provided constructive comments that are aligned with our hopes for the paper. Accordingly, I’m happy to invite a revision.

To help guide this revision, I’ve highlighted points raised by the reviewers that seem to warrant special attention in blue. In addition, you’ll note a general call for more pedagogical and helpful text, which I think is important for maximizing this manuscript's impact. As you revise your text, please feel free to disregard our standard manuscript length limits in favor of expanded text that is both more penetrable and more complete. When you do this, consider shifting the balance of the paper towards the conceptually important points raised by Reviewer 3 and highlighted in yellow.

I’d also like to be explicit about an almost philosophical stance that we take at Cell Systems. We believe that understanding how approaches fail is fundamentally interesting: it provides critical insight into understanding how they work. We also believe that all approaches do fail and that it’s unreasonable, even misleading, to expect otherwise. Accordingly, when papers are transparent and forthright about the limitations and crucial contingencies of their approaches, we consider that to be a great strength, not a weakness. Please keep this in mind as you revise your manuscript.

I hope you find this feedback helpful. If you have any questions or concerns, I’m always happy to talk, either over Zoom or on the phone. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,
Quincey

Quincey Justman, Ph.D.
Editor-in-Chief, Cell Systems

Reviewers' comments:

Reviewer #1: The paper "Machine Learning-Assisted Directed Evolution Navigates a Combinatorial Epistatic Fitness Landscape with Minimal Screening Burden" by Wittmann et al. benchmarks a number of design choices when using machine learning to recommend mutations for experimental acquisition, e.g., as part of a directed evolution effort. Overall, I found the paper very informative and transparent and therefore recommend this paper for publication. The methods are sound, this kind of benchmarking study is greatly needed in this area of research, and the results will be useful to practitioners in the emerging
field of machine-guided biological design. The authors could, however, work to improve the clarity and the accessibility of the manuscript, for which more specific comments are provided below.

Specific comments:

- I was particularly struck by the transparency and pragmatism of the paper - the authors simply set out to find the best method choices without any particular methodological agenda. From a communication perspective, however, I think the paper would really benefit from a more concise and direct summary of the pipeline choices that worked the best and recommendations that the authors would make to practitioners based on their results. Most computational benchmarking papers have a concise "topline" summary of the best performing configurations in both the introduction and the conclusion. The authors should consider having such a summary in their paper as well, e.g., "we found X search space restriction with X training samples and X embedding strategy performed the best according to our experimental configuration" (or something similar to that).

- Related to communication, I am worried that the paper in many places jumps directly into technical jargon without adequate explanation for the broader biological audience of Cell Systems. While the Cell Systems audience includes machine learning practitioners, this audience also includes those with more limited computational expertise, and the main text explicitly mentions that the pipeline is designed "for use by non-computational and non-ML experts." In addition to in-text definitions, perhaps one way to improve accessibility is to include a glossary of common technical terms alongside the manuscript.

- Additional methodological details are needed to explain exactly how the regressors at the final stage of the pipeline are ensembled together. I could not find a clear description of this in the main text and the "MLDE Programmatic Implementation" part of the Methods section gives only very sparse detail. I also had trouble finding the exact values for "k" and "N" referred to in this section. The way the authors describe how results are aggregated across CV folds and regressors lacks detail -- the authors should probably just provide an equation.

- Related to the regressor ensembling, the authors use a rather eclectic combination of methods that includes various neural architectures, decision tree-based methods, and the collection of regressors available in scikit-learn. The authors seem to treat this ensemble more or less like a black box. The authors should at least comment on the individual regressors that were most useful for prediction, e.g., which ones are consistently included in the top "N" of the ensemble. This would be very useful information for a practitioner.

- The authors do a good job at noting that their framework is only tested on a single protein, GB1, and that this necessarily limits the breadth of conclusions that they can draw from this study. Still, I do think the authors can and should make specific recommendations for practitioners based on this specific experimental setup.

Reviewer #2: Summary:
This is a well-written paper with interesting simulation studies, providing useful insights into directed evolution and fitness landscape exploration. Overall this is a nice contribution, but there are a few aspects of the study that are underdeveloped. Concrete suggestions for improving the manuscript are provided below.

Major Comments:

1. Given the recent results showing the models trained from Rives et al 2020 and Elnaggar et al 2020, it would be very interesting to rerun results using these larger transformers. In particular, ESM-1b (Meier et al. 2020) and ProtBERT-BFD (Elnaggar et al 2020) should hopefully be reasonably sized to allow for updated results without excessive computational demands. Since there is continued effort by these teams and others to scale up pretraining, results indicating whether current efforts to scale up result in positive outcomes for this application would be widely useful to the representation learning for proteins community.

2. This work proposes to use priors induced by ΔΔG or by evolutionary sequences for sampling in place of a uniform prior. A third interesting prior would be one induced by pretrained language models. For each variant within hamming distance 4 of the wild type, one could compute the log probability of that sequence under the TAPE LSTM by adding the log probability under the forward and backward LSTMs. This is the exact loss used to train the model, so comparing this distribution over variants to the one given by ΔΔG would yield insight into whether pretraining produces useful local information. In particular, this would help ground the claim "These results may align, however, with the recent observations of Biswas et al., who proposed that embeddings generated from unsupervised models guide a subsequent supervised search away from sequences the unsupervised model deems to be 'unnatural' (Biswas et al., 2020)" made in reference to Biswas et al. 2020.

3. Oddly only physicochemical representation (Georgiev 2009) is considered in the studies described on pages 5-7. Embeddings from language models should also be evaluated in these studies.

4. The results described in Table 2 are rather strange. It would be helpful if the authors could explain more thoroughly the pattern of "Max Achieved" dependence on the Sampling Limit. In particular, why is there such a sharp transition in "Max Achieved" when Sampling Limit is increased from 1600 to 3200? It would be helpful to compare the fitness distribution of the set from which training data are sampled, for different Sampling Limits.

Minor Comments:

1. It would help if the motivation for using Normalized Discounted Cumulative Gain (NDCG) in the case of directed evolution could be made more explicit.

2. The authors show that ΔΔG is a reasonable prior for choosing sequences by referencing the high Spearman's ρ of ΔΔG with measured GB1 fitness. It would strengthen the author's claim that
unsupervised models (EVMutation, language models) give poor priors over sequences if a similar Spearman's $\rho$ for those models was computed (e.g. with EVMutation likelihood or LSTM likelihood). The result that unsupervised models do not give priors appropriate for sequence design is a fascinating result by itself, and this work is well-positioned to show that!

3. Page 5: In the sentence: "To demonstrate the concept and test the effectiveness… a series of classifiers that, with 50% accuracy...", "accuracy" should be changed to "True Discovery Proportion"

Reviewer #3: The manuscript by Wittman et al. entitled "Machine Learning-Assisted Directed Evolution Navigates a Combinatorial Epistatic Fitness Landscape with Minimal Screening Burden " describes a protein engineering approach in which a ML-model is developed that allows in silico screening of combinatorial libraries constructed using saturation mutagenesis. While this approach was previously published in Wu et al. 2019, many of the design decisions involved were not explored and this paper addresses these omissions. The authors claim that their machine learning guided approach MLDE avoids the path-dependence inherent in existing directed evolution approaches that use successive rounds of single-site mutagenesis to explore a protein fitness landscape. The resulting protocol obtained the global fitness maximum for a four-site combinatorial fitness landscape many times more frequently than the baseline path-dependent approach.

The ultimate goals of the work were two-fold:

i. To demonstrate that specific encoding strategies, training and modeling procedures and training set design strategies enable the improved performance of a machine learning-assisted approach to directed evolution.

ii. To demonstrate that the three major paradigms of protein engineering can be combined to yield a highly efficient engineering pipeline.

This is an interesting study, where the authors successfully highlight a major issue with navigating protein fitness landscapes (the ubiquity of holes) and demonstrate that careful training data design can be leveraged to significantly improve model performance. Overall, I am not convinced by the value added by specific encoding strategies, and the inclusion of 1D CNN models and XGB models seems to repair an oversight rather than be innovative. However, I think the insights about holes and how to navigate fitness landscapes that contain many holes are novel and of significant interest to this field.

Major comments:
In Figure 2, it is interesting that the one-hot encoding generally returns a slightly lower NDCG than the more informative encodings, but it isn't clear that a higher NDCG implies a better campaign outcome. The fact that NDCG doesn't correlate with the max fitness over the top 96 prediction suggests that NDCG isn't the most useful metric for this context. My concern would be that the NDCG is swamped by the very large...
number of low fitness variants shown in Fig. 3A, which presumably cannot be usefully ranked. Given this distribution, it might be more relevant to consider the NDCG over the top 1536 or similar number, especially as this is roughly the number of model-designed variants that reasonably might be tested in a typical experimental campaign.

Overall, the differences in performance between these encodings seem small. The strong performance of the one-hot encoding is striking, and is perhaps the most interesting result particularly as it is far simpler for users to implement than many of the others. I don't understand why 2000 simulations were used - is there some reason? Otherwise it would make sense to plot these metrics as a function of the number of simulations carried out, since individuals are unlikely to run 2000 versions of the campaign in the lab so understanding the behavior with smaller numbers is relevant. Given that the simulations have already been run, it should be difficult to compute these statistics, potentially also for different orderings. As it stands, I don't think that the evidence strongly supports the claim of the paper that more informative encodings improve MLDE outcomes.

The last full paragraph on page four is highly speculative. I would suggest that the authors either substantiate the hypotheses contained therein, or simply remove the paragraph. Given the size of the space and the large number of low fitness variants and the relatively small dynamic range of the fitness I see no reason to expect that a high NDCG would lead to a higher maximum fitness achieved - these metrics measure different things. If the authors do wish to substantiate these hypotheses, they could retrain the models that generate the learned embeddings using additional GB1 sequences, and ask whether the resulting embeddings improve the maximum fitness achieved. In addition they could retrain on many more sequences and ask the same question. Given the large numbers of sequences used to train these models, I am not confident that an improved result will follow in either case, but if it did it would be interesting.

The section regarding the challenge of holes in combinatorial fitness landscapes is really interesting, and could usefully be expanded. In particular the second paragraph of this section is really dense and hard to follow. I suggest that the authors allow more paragraphs, and explain the results stated rapidly at the end of this paragraph using more words and more detail, to make them easier to understand. These initial results have real potential ramifications for how users collect their training data (even in cases unlike GB1 where e.g. it is possible to build reasonable sequence alignments), so I would suggest allocating them more space in the manuscript.

Minor comments:

At the bottom of page three of the main text, please could the authors provide a brief explanation of how normalized discounted cumulative gain is defined. I do not believe that many readers will be familiar with this metric, and it would be useful to have it briefly explained here, rather than ask readers to refer to a different section.

In the section 'More informative encodings improve MLDE outcome' the authors state that they perform 2000 rounds of simulated MLDE, using 384 randomly drawn GB1 variants as training data in each simulation'. It would be clearer if the authors simply said that they perform 2000 MLDE simulations, where
each uses a random draw of 384 GB1 variants as training data. I don't see how the terminology 'round' adds anything here - rather it is confusing, because I imagine there would be several rounds within a single MLDE simulation, and as far as I can tell this isn't what the authors are doing?

Methods, encoding simulations. It would be better to use a random seed for model training, to avoid the introduction of variance by models that rely on randomness for training. Please could the authors set the random seed, or quantify the variance that may result from not setting it.

In the section on Models/Training Procedures more tailored for combinatorial fitness landscapes it might be helpful to readers to note that 1D CNNs effectively compute properties using a sliding window over the sequence, a technique that has been used in bioinformatics for decades. It would be nice to add some references to this significant body of work, rather than focusing on the recent literature where it has been largely ignored.

Authors’ response to the reviewers’ first round comments

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear Frances and Bruce,

I'm very pleased to let you know that the reviews of your revised manuscript are back, the peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication.

In addition to the final comments from the reviewers, I've made some suggestions about your manuscript within the “Editorial Notes” section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager. We hope to receive your files within 5 business days, but we recognize that the COVID-19 pandemic may challenge and limit what you can do. Please email us directly if this timing is a problem or you're facing extenuating circumstances.

I'm looking forward to going through these last steps with you. Although we ask that our editorially-guided changes be your primary focus for the moment, you may wish to consult our FAQ (final formatting checks tab) to make the final steps to publication go more smoothly. More technical information can be found below my signature, and please let us know if you have any questions.

All the best,
Quincey Justman, Ph.D.
Editor-in-Chief, Cell Systems

Editorial Notes

*Manuscript Text:* Your manuscript text is superlative. Please leave it essentially as is, making only the minor changes requested in this letter. These minor changes are necessary to adhere to our “house style” as follows:

- House style disallows editorializing within the text (e.g. strikingly, surprisingly, importantly, etc.). These terms are a distraction and they aren't needed—your excellent observations are certainly impactful enough to stand on their own. Please remove these words and others like them. “Notably” is suitably neutral to use once or twice if absolutely necessary.
- Please only use the word "significantly" in the statistical sense.

*Figures and tables:* Articles can have up to 7 (Figures + Tables) in the main text. Please make changes accordingly.

Thank you!

**Reviewer comments:**

Reviewer #1: The authors have done an excellent job of responding to reviewer comments. I find the expanded benchmarking of different zero-shot prediction methods particularly useful. The exploration of pipeline design choices in the paper will be a valuable resource to the community.

Reviewer #2: I would like to commend the authors for doing an outstanding job of revising their manuscript. I much enjoyed reading this revised version. The newly added comparative study of state-of-the-art protein embeddings and zero-shot predictors in the context of ftMLDE is fascinating, and I am sure that the community will appreciate the authors' effort. All in all, this is really a great paper.

All my previous comments have been satisfactorily addressed. I have just a couple of very minor comments on this revision:

Page 8, last paragraph: I found the phrase "describing each combination with 8192 features" a bit confusing at first. Perhaps "describing each combination of 4 amino acids with 8192 features" would be clearer?
Page 12: I find it counterintuitive that using "naïve" probability led to better predictions than using conditional probability. Could the authors provide an explanation for this behavior?

Reviewer #3: I thank the authors for their constructive responses to the comments that I made, and I have no further comments to add.
Response to Reviewers

Reviewers' comments:
Reviewer #1: The paper "Machine Learning-Assisted Directed Evolution Navigates a Combinatorial Epistatic Fitness Landscape with Minimal Screening Burden" by Wittmann et al. benchmarks a number of design choices when using machine learning to recommend mutations for experimental acquisition, e.g., as part of a directed evolution effort. Overall, I found the paper very informative and transparent and therefore recommend this paper for publication. The methods are sound, this kind of benchmarking study is greatly needed in this area of research, and the results will be useful to practitioners in the emerging field of machine-guided biological design. The authors could, however, work to improve the clarity and the accessibility of the manuscript, for which more specific comments are provided below.

Specific comments:
1) I was particularly struck by the transparency and pragmatism of the paper - the authors simply set out to find the best method choices without any particular methodological agenda. From a communication perspective, however, I think the paper would really benefit from a more concise and direct summary of the pipeline choices that worked the best and recommendations that the authors would make to practitioners based on their results. Most computational benchmarking papers have a concise "topline" summary of the best performing configurations in both the introduction and the conclusion. The authors should consider having such a summary in their paper as well, e.g., "we found X search space restriction with X training samples and X embedding strategy performed the best according to our experimental configuration" (or something similar to that).

We agree that the addition of a topline summary would be beneficial. We have added a sentence to that effect in the second-to-last paragraph of the introduction as well as within the section discussing the different effects of training set design strategy and encoding on ftMLDE ("Zero-Shot Predictions for Training Set Design Enable Highly Effective ftMLDE on the GB1 Landscape").

2) Related to communication, I am worried that the paper in many places jumps directly into technical jargon without adequate explanation for the broader biological audience of Cell Systems. While the Cell Systems audience includes machine learning practitioners, this audience also includes those with more limited computational expertise, and the main text explicitly mentions that the pipeline is designed "for use by non-computational and non-ML experts." In addition to in-text definitions, perhaps one way to improve accessibility is to include a glossary of common technical terms alongside the manuscript.

Thank you for the feedback. We have added in-line definitions of ML jargon as well as added a glossary of common technical terms used in the manuscript. We have also rewritten much of the manuscript to be more accessible to a biological audience, specifically by expanding the amount of detail we go into when first introducing ML concepts as well as, where possible, framing ML concepts within the scope of protein engineering by providing relevant examples.

3) Additional methodological details are needed to explain exactly how the regressors at the final stage of the pipeline are ensembled together. I could not find a clear description of this in the main text and the "MLDE Programmatic Implementation" part of the Methods section gives only very sparse detail. I also had trouble finding the exact values for "k" and "N" referred to in this section. The way the authors describe how results are aggregated across CV folds and regressors lacks detail -- the authors should probably just provide an equation.

We agree that this initial description was not clear. The nature of the procedure was not really amenable to description by an equation; however, we have expanded the section “MLDE Programmatic Implementation” to include specific details and examples of the values we used throughout our study. We have also gone into greater detail regarding each step of the programmatic implementation of MLDE.

4) Related to the regressor ensembling, the authors use a rather eclectic combination of methods that includes various neural architectures, decision tree-based methods, and the collection of regressors
available in scikit-learn. The authors seem to treat this ensemble more or less like a black box. The authors should at least comment on the individual regressors that were most useful for prediction, e.g., which ones are consistently included in the top "N" of the ensemble. This would be very useful information for a practitioner.

Thank you for the feedback. You are correct that we treat the ensemble of models as a black box. We use a variety of model architectures and treat the ensemble as a black box as it is impossible to know, a priori what models will be most beneficial for a given landscape. To make this clearer, we have added a sentence to the section “MLDE Procedure, Simulated MLDE, and Evaluation Metrics” that states “A variety of models are trained because the shape of the fitness landscape is not known a priori...”. Additionally, we now provide summary statistics (NDCG, mean fitness achieved, and max fitness achieved) for all model architectures for all simulations performed as a downloadable supplementary file.

5) The authors do a good job at noting that their framework is only tested on a single protein, GB1, and that this necessarily limits the breadth of conclusions that they can draw from this study. Still, I do think the authors can and should make specific recommendations for practitioners based on this specific experimental setup.

We believe that this comment is addressed along with our additions that address comment 1.

Reviewer #2: Summary:

This is a well-written paper with interesting simulation studies, providing useful insights into directed evolution and fitness landscape exploration. Overall this is a nice contribution, but there are a few aspects of the study that are underdeveloped. Concrete suggestions for improving the manuscript are provided below.

Major Comments:

1. Given the recent results showing the models trained from Rives et al 2020 and Elnaggar et al 2020, it would be very interesting to rerun results using these larger transformers. In particular, ESM-1b (Meier et al. 2020) and ProtBERT-BFD (Elnaggar et al 2020) should hopefully be reasonably sized to allow for updated results without excessive computational demands. Since there is continued effort by these teams and others to scale up pretraining, results indicating whether current efforts to scale up result in positive outcomes for this application would be widely useful to the representation learning for proteins community.

Thank you for the suggestion. We agree that inclusion of simulations run using embeddings from these state-of-the-art models would be beneficial. As such, we have included the results from MLDE simulations run using encodings from ESM-1b, ProtBERT-BFD, and the MSA Transformer recently published by Rao et al. (Rao et al., 2021). To test the effects of these encodings in line with recent works using them in benchmarking studies, we have also included the results of simulations run in a low-N setting.

2. This work proposes to use priors induced by ΔΔG or by evolutionary sequences for sampling in place of a uniform prior. A third interesting prior would be one induced by pretrained language models. For each variant within hamming distance 4 of the wild type, one could compute the log probability of that sequence under the TAPE LSTM by adding the log probability under the forward and backward LSTMs. This is the exact loss used to train the model, so comparing this distribution over variants to the one given by ΔΔG would yield insight into whether pretraining produces useful local information. In particular, this would help ground the claim “These results may align, however, with the recent observations of Biswas et al., who proposed that embeddings generated from unsupervised models guide a subsequent supervised search away from sequences the unsupervised model deems to be ‘unnatural’ (Biswas et al., 2020)” made in reference to Biswas et al. 2020.
Thank you for the excellent suggestion. We agree that exploration of additional priors would improve the work. We have additionally evaluated the ability of EVmutation and DeepSequence to act as zero-shot predictors for ftMLDE with GB1. We have also incorporated your suggestion of using priors from pretrained language models. We decided to test mask-filling protocols with all models from the ESM GitHub repository as well as ProtBERT and ProtBERT-BFD from the ProtTrans repository. We chose these models instead of an LSTM as they allowed additional comparison of the effects of model size (ESM provides many models of many different sizes) and training corpus (e.g., UniRef50 vs UniRef100 vs BFD). While our results generally support our original hypothesis that learned embeddings may guide a supervised model away from predicting unnatural variants as high in fitness, we have decided to remove this claim and associated paragraph from the work—in agreement with other reviewer comments we did not feel there was enough evidence provided by our experiments to support such a claim.

3. Oddly only physicochemical representation (Georgiev 2009) is considered in the studies described on pages 5-7. Embeddings from language models should also be evaluated in these studies.

Physicochemical parameters were chosen as they were the most effective "more informative encoding" with low representation dimensionality. Computational time expands rapidly with encoding dimension, and so inclusion of all encodings (which were typically much higher dimensional representations than physicochemical) for all experiments performed would be extremely expensive. Additionally, simulation results take up a large amount of disk space in aggregate, further limiting how many simulations can be performed and stored. Still, we agree that investigation of the effects of encoding in combination with the ftMLDE protocol would be valuable. As such, we have additionally performed all ftMLDE simulations using one-hot encoding and embeddings from the MSA Transformer and have expanded discussion on the effect of encoding on ftMLDE and MLDE outcome. These computations took tens of thousands of compute-hours to complete and their results take up >13 TB of disk space. We feel that, while it would be informative to include all other encodings tested in the other experiments, the added value would not outweigh the computational cost.

4. The results described in Table 2 are rather strange. It would be helpful if the authors could explain more thoroughly the pattern of "Max Achieved" dependence on the Sampling Limit. In particular, why is there such a sharp transition in "Max Achieved" when Sampling Limit is increased from 1600 to 3200? It would be helpful to compare the fitness distribution of the set from which training data are sampled, for different Sampling Limits.

Thank you for the suggestion. We provide ECDFs giving the distribution of fitness and training set diversity in the different training sets tested for the ftMLDE results in question, which show that the likely reason for the jump between 1600 and 3200 is a result of having too little sequence diversity in the training data. We have expanded discussion on the effects of training data and encoding on MLDE outcome and have devoted half a paragraph to discussion of the observed effect of increasing the sampling limit. Because the goal of this work was to investigate which design considerations could lead to improved MLDE, however, we feel that detailed investigation into specific observations goes beyond the scope of this work. We do include a note now, however, that our results show that there appears to be an interplay between the source of training data, encoding used, and ftMLDE outcome.

Minor Comments:

----------------

1. It would help if the motivation for using Normalized Discounted Cumulative Gain (NDCG) in the case of directed evolution could be made more explicit.

Thank you for the feedback. We have added a paragraph to the section “MLDE Procedure, Simulated MLDE, and Evaluation Metrics” discussing the motivation for using NDCG as an evaluation metric in our work.
2. The authors show that ΔΔG is a reasonable prior for choosing sequences by referencing the high Spearman's ρ of ΔΔG with measured GB1 fitness. It would strengthen the author's claim that unsupervised models (EVMutation, language models) give poor priors over sequences if a similar Spearman's ρ for those models was computed (e.g. with EVMutation likelihood or LSTM likelihood). The result that unsupervised models do not give priors appropriate for sequence design is a fascinating result by itself, and this work is well-positioned to show that!

Thank you for the excellent suggestion. We agree that exploration of additional priors would improve the work and further support our claim. As such, we have additionally evaluated the ability of EVMutation and DeepSequence to act as zero-shot predictors for fitMLDE with GB1. We also decided to test mask-filling protocols with all models from the ESM GitHub repository as well as ProtBERT and ProtBERT-BFD from the ProtTrans repository. While our results generally support our original hypothesis that learned embeddings may guide a supervised model away from predicting unnatural variants as high in fitness, we have decided to remove this claim and associated paragraph from the work.

3. Page 5: In the sentence: "To demonstrate the concept and test the effectiveness... a series of classifiers that, with 50% accuracy...", "accuracy" should be changed to "True Discovery Proportion"

Thank you for feedback. We have reworked the section introducing designed training data to completely avoid using the term "simulated classifier", and instead simply define how we designed our training data. We felt that this change made the manuscript more accessible to a larger audience. As a result, we no longer include the use of the word “accuracy” as you describe.

Reviewer #3: The manuscript by Wittman et al. entitled "Machine Learning-Assisted Directed Evolution Navigates a Combinatorial Epistatic Fitness Landscape with Minimal Screening Burden " describes a protein engineering approach in which a ML-model is developed that allows in silico screening of combinatorial libraries constructed using saturation mutagenesis. While this approach was previously published in Wu et al. 2019, many of the design decisions involved were not explored and this paper addresses these omissions. The authors claim that their machine learning guided approach MLDE avoids the path-dependence inherent in existing directed evolution approaches that use successive rounds of single-site mutagenesis to explore a protein fitness landscape. The resulting protocol obtained the global fitness maximum for a four-site combinatorial fitness landscape many times more frequently than the baseline path-dependent approach.

The ultimate goals of the work were two-fold:

i. To demonstrate that specific encoding strategies, training and modeling procedures and training set design strategies enable the improved performance of a machine learning-assisted approach to directed evolution.

ii. To demonstrate that the three major paradigms of protein engineering can be combined to yield a highly efficient engineering pipeline.

This is an interesting study, where the authors successfully highlight a major issue with navigating protein fitness landscapes (the ubiquity of holes) and demonstrate that careful training data design can be leveraged to significantly improve model performance. Overall, I am not convinced by the value added by specific encoding strategies, and the inclusion of 1D CNN models and XGB models seems to repair an oversight rather than be innovative. However, I think the insights about holes and how to navigate fitness landscapes that contain many holes are novel and of significant interest to this field.

Major comments:
In Figure 2, it is interesting that the one-hot encoding generally returns a slightly lower NDCG than the more informative encodings, but it isn't clear that a higher NDCG implies a better campaign outcome. The
Thank you for your comment. We agree that low-fitness values cannot be reasonably ranked and that ranking them appropriately is an unimportant objective for models used in MLDE (or protein engineering in general). However, this was precisely why we chose NDCG over others: it inherently weighs correct identification and ranking of high-fitness variants as more important than correct identification and ranking of low-fitness variants.Importantly, NDCG does not require us to make an arbitrary decision of what constitutes variants that are important to rank correctly, as the importance of different variants for determining the ideal NDCG naturally diminishes with their fitness. We further expand on this logic in a paragraph that has been added to the section “MLDE Procedure, Simulated MLDE, and Evaluation Metrics”. We feel that this addition justifies and makes clearer our choice of NDCG as an evaluation metric for MLDE. Finally, we have included the requested calculations of NDCG using the top 384 or 1536 samples as supplementary evaluation metrics in data available for download. We do not discuss them in the main body of the work, as the trends are more or less the same as for NDCG calculated using all samples.

The applicability of NDCG aside, we do agree that a greater NDCG does not necessarily mean a better campaign outcome. We now much more explicitly state in our introduction of the different evaluation metrics that they all measure related, yet different outcomes of the procedure. We have adapted our discussion of the results to include all evaluation metrics tested, which yields the conclusion that, while more informative embeddings can improve MLDE, it depends on what metric is used for measuring “improvement” and the specific “more informative embedding” chosen. This conclusion is now what we report in our discussion of the effects of encoding strategy, which amounts to a qualification of our original claim.

Overall, the differences in performance between these encodings seem small. The strong performance of the one-hot encoding is striking, and is perhaps the most interesting result particularly as it is far simpler for users to implement than many of the others. I don’t understand why 2000 simulations were used – is there some reason? Otherwise it would make sense to plot these metrics as a function of the number of simulations carried out, since individuals are unlikely to run 2000 versions of the campaign in the lab so understanding the behavior with smaller numbers is relevant. Given that the simulations have already been run, it should be difficult to compute these statistics, potentially also for different orderings. As it stands, I don’t think that the evidence strongly supports the claim of the paper that more informative encodings improve MLDE outcomes.

2000 simulations were chosen to fit within computational limitations while still enabling construction of an informative distribution of results. Thank you for the suggestion. We note that the order in which simulations are performed will have no effect on the distribution, as all simulations are independently performed with no input from previous ones. As such, we do not believe that it is necessary to plot the results as a function of the number of simulations carried out. However, we agree that presenting the results as a distribution would be far more informative than having them summarized in a table, allowing readers to see the extremes of simulation results. In conjunction with our response to your previous comment, we have thus included the full distribution of max and mean fitness achieved for all simulations (plotted as violin plots).

The last full paragraph on page four is highly speculative. I would suggest that the authors either substantiate the hypotheses contained therein, or simply remove the paragraph. Given the size of the space and the large number of low fitness variants and the relatively small dynamic range of the fitness I see no reason to expect that a high NDCG would lead to a higher maximum fitness achieved - these metrics measure different things. If the authors do wish to substantiate these hypotheses, they could
retrain the models that generate the learned embeddings using additional GB1 sequences, and ask whether the resulting embeddings improve the maximum fitness achieved. In addition they could retrain on many more sequences and ask the same question. Given the large numbers of sequences used to train these models, I am not confident that an improved result will follow in either case, but if it did it would be interesting.

Thank you for the feedback. We agree that this paragraph was highly speculative but feel that further investigation of the observations would be beyond the scope of this work. As such, we have removed this paragraph from the manuscript.

The section regarding the challenge of holes in combinatorial fitness landscapes is really interesting, and could usefully be expanded. In particular the second paragraph of this section is really dense and hard to follow. I suggest that the authors allow more paragraphs, and explain the results stated rapidly at the end of this paragraph using more words and more detail, to make them easier to understand. These initial results have real potential ramifications for how users collect their training data (even in cases unlike GB1 where e.g. it is possible to build reasonable sequence alignments), so I would suggest allocating them more space in the manuscript.

We agree that expansion on the topic of the challenge of holes would be valuable. We have expanded the dense paragraph into multiple to make the discussion clearer. Additionally, we have expanded the number of zero-shot predictors tested to include sequence-based strategies, including EVMutation, DeepSequence, and mask-filling using language models. We hope that investigation into these different zero-shot strategies will be beneficial for the community. We also believe that investigation of training set design strategies beyond predicted $\Delta\Delta G$ further supports our claim of the value of designing training sets to avoid holes. Discussion of various training set design strategies now constitutes the bulk of our manuscript, and we identify it as the single most important consideration in improving the effectiveness of our MLDE protocol.

Minor comments:

At the bottom of page three of the main text, please could the authors provide a brief explanation of how normalized discounted cumulative gain is defined. I do not believe that many readers will be familiar with this metric, and it would be useful to have it briefly explained here, rather than ask readers to refer to a different section.

Thank you for the feedback. We have expanded our discussion of NDCG upon its introduction to motivate its use as an evaluation metric. We have included the equation defining it in this section as well.

In the section 'More informative encodings improve MLDE outcome' the authors state that they perform 2000 rounds of simulated MLDE, using 384 randomly drawn GB1 variants as training data in each simulation'. It would be clearer if the authors simply said that they perform 2000 MLDE simulations, where each uses a random draw of 384 GB1 variants as training data. I don't see how the terminology 'round' adds anything here - rather it is confusing, because I imagine there would be several rounds within a single MLDE simulation, and as far as I can tell this isn't what the authors are doing?

We agree that this terminology could be confusing and have eliminated its use in the manuscript. We instead refer to either “MLDE simulations” or “simulated MLDE experiments” to be clear that the MLDE simulations are not iterative.

Methods, encoding simulations. It would be better to use a random seed for model training, to avoid the introduction of variance by models that rely on randomness for training. Please could the authors set the random seed, or quantify the variance that may result from not setting it.

Thank you for the feedback. We have rerun all calculations including random seeds.
In the section on Models/Training Procedures more tailored for combinatorial fitness landscapes it might be helpful to readers to note that 1D CNNs effectively compute properties using a sliding window over the sequence, a technique that has been used in bioinformatics for decades. It would be nice to add some references to this significant body of work, rather than focusing on the recent literature where it has been largely ignored.

Thank you for this comment. We agree that inclusion of the use of sliding window analysis improves the interpretability of how CNNs work. We have additionally included references to works using sliding window analysis to highlight the related body of work.