nature research

Peer Review Information

Journal: Nature Computational Science
Manuscript Title: Machine Learning Ensemble Directed Engineering of Genetically Encoded Fluorescent Calcium Indicators
Corresponding author name(s): Dr Andre Berndt

Editorial Notes:

Reviewer Comments & Decisions:

Decision Letter, initial version:

Date: 11th September 23 23:27:52
Last Sent: 11th September 23 23:27:52
Triggered By: Jie Pan
From: jie.pan@us.nature.com
To: berndtuw@uw.edu
BCC: jie.pan@us.nature.com
Subject: Decision on Nature Computational Science manuscript NATCOMPUTSCI-23-0833A
Message: ** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Dear Dr Berndt,

Your manuscript "Machine Learning Ensemble Directed Engineering of Genetically Encoded Fluorescent Calcium Indicators" has now been seen by 4 referees, whose comments are appended below. You will see that while they find your work of interest, they have raised points that need to be addressed before we can make a decision on publication.

The referees' reports seem to be quite clear. Naturally, we will need you to address *all* of the points raised.

While we ask you to address all of the points raised, the following points need to be substantially worked on:

- Please follow the referees' suggestions to add the required comparison data.
- Please follow referees' suggestion to better discuss the limitation in your Discussion section.
- Please address those technical concerns raised by all referees.

Please use the following link to submit your revised manuscript and a point-by-point response to the referees' comments (which should be in a separate document to any cover letter):

[REDACTED]

** This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first. **

To aid in the review process, we would appreciate it if you could also provide a copy of your manuscript files that indicates your revisions by making use of Track Changes or similar mark-up tools. Please also ensure that all correspondence is marked with your Nature Computational Science reference number in the subject line.

In addition, please make sure to upload a Word Document or LaTeX version of your text, to assist us in the editorial stage.

If you have any issues when updating your Code Ocean capsule during the revision process, please email the Code Ocean support team Cc'ing me.

To improve transparency in authorship, we request that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We hope to receive your revised paper within three weeks. If you cannot send it within this time, please let us know.

We look forward to hearing from you soon.

Best regards,

Jie Pan, Ph.D.
Senior Editor
Nature Computational Science

Reviewers comments:

Reviewer #1 (Remarks to the Author):

Berndt and coworkers describe their efforts to apply machine learning methods to develop improved versions of the GCaMP calcium ion indicator. To train their machine learning algorithm, they used previously reported data for 1078 variants that were screened during the development of GCaMP6 and jGCaMP7. The two key properties they focussed on were the fluorescent response to one action potential and the fluorescent decay time. The model was also trained on amino acid properties such as size and polarity, which seemed to improve the predictive properties of the models. Sequences that were predicted by the model to have either the biggest changes in the fluorescent response and kinetics were tested experimentally.

The authors clearly explain how the L317 variants had the opposite effect from the prediction. While they go on to rationalize why this is the case. While I don't expect any sort of protein sequence prediction algorithm to be perfect, the whole point of this work is to try to demonstrate the utility of the machine learning algorithms for predicting mutations that would improve the performance. With the L317 mutations, it is clear that the algorithm correctly identified an important "hot spot" in the protein sequence where mutations were likely to impact the performance. I am certain that a well-trained biochemist could have made the same prediction. Indeed, the fact that Dana et al previously tested mutations at this position, demonstrates that this is the case.

Based on the machine learning predictions, and subsequent in vitro testing, the authors identified jGCaMP7s L317H as the most promising variant. As noted above, this machine learning algorithm had predicted that this variant would have decreased fluorescent response. To continue to improve this variant, the authors resorted to conventional (that is, empirical) protein engineering. They transplanted the L317H mutation to jGCaMP8f, and tested various combinations of promising mutations in the jGCaMP7s scaffold, ultimately leading to the identification of 3 improved variants (eGCaMP, eGCaMP+, and eGCaMP2+). These improved variants also showed improved performance in primary neurons.

Overall, I found this work to be interesting and innovative. I greatly appreciate the ambition and goals of the work, and I am excited about the potential for machine learning algorithms to accelerate protein engineering and ultimately provide better GEFIs than would otherwise be attainable. However, I also found myself unconvinced that the machine learning provided any valuable insight, beyond what an appropriately trained protein chemist could have gained by studying the literature and the crystal structures. Though, I also appreciate that maybe the goal here is to match the ability of expert, and not necessarily exceed it? If the goal is to match the insight of an expert (who still might make imperfect predictions, like the L317 mutations), then I would consider this work a success. If the goal is to exceed the abilities of an expert, then I don't consider this work to be a success. Either way, I feel that there needs to be discussion that frames this work in this context, which is truly fundamental to this and all other machine learning and AI efforts.

In addition to the fundamental concern described above, I have three other major concerns related to the protein engineering aspects of this work. I will leave it to other reviewers to comment on the appropriateness of the computational methods.

1. The data set used for training is highly biased. As far as I understand, the residues

to be mutated were chosen based on inspection of the crystal structure to identify mutations that were most likely to have an impact on the GCaMP function. I assume that this bias must propagate through every aspect of this work and the every prediction that the algorithm makes. The authors will need to more clearly address this concern, and explain why (or why not) it is a limitation or this work. I feel that a useful discussion point would be to provide some perspective on what the ideal training set for such efforts would be, as this may inspire other workers to collect the appropriate data.

2. In the section on in vitro testing of predicted mutations, the authors explain how some mutations gave the expected response, yet others gave the opposite response. Somehow, the authors will need to quantify the overall accuracy of their predictions and convince the reader that their predictions, in sum, are better than would be expected based on random chance.

3. Finally, I feel that there needs to be a bit more nuance applied to the comparisons between the three new variants and the previously reported GCaMP variants. I believe that previous workers settled on their final variants (GCaMP6s, jGCaMP7s, etc), as the best possible compromises, taking all of the properties into account. During their screening, the previous workers certainly found variants that were improved in one property or another (and are thus in the training sets), but these are not necessarily the final variants that they settled on. The authors will need to acknowledge this and write their comparisons and discussion accordingly.

Reviewer #2 (Remarks to the Author):

This paper reports a successful case study about improving the ability of GEFI. It is clearly written and the result is basically worth publishing. However, I have the following comments.

1) The ensemble used here is based on relatively old machine learning techniques. Please clarify why modern deep learning methods are not preferred here.

2) The search range from the wild type is limited. Single mutations and a few double mutations are considered, but not more. Please discuss if this technique can be used to discover more distant mutations.

3) Not being an expert in this field, I could not really understand how the performance of the new variants compares to the best GEFI available. Please discuss about it.

Reviewer #3 (Remarks to the Author):

Wait et al. present a new approach for machine-learning-based engineering of genetically-encoded calcium (or more generally, fluorescence) indicators. Their approach depends on the existence of a large database of standardized experiments that explore the effect of multiple mutations on a protein sensor of interest. Then, they implemented a combination of three regression algorithms to learn the

important parameters that affected fluorescence fluctuations and kinetics when different amino acids were mutated. The model is then used to predict the performance of new mutations, which were never tested before, and finally, the predicted hits were tested in vitro to validate the model accuracy. The main advantage of the presented approach is its ability to saturate the tested mutation space, i.e. to computationally test all the possible combinations of amino acids in multiple specific positions, which is expensive, time- and labor-consuming, but is also experimentally inefficient – as many of the variants are expected to show low performance levels.
The manuscript is well-written, and the concept is exciting. The main limitation to its application to additional types of sensors is the availability of experimental databases that are required to train the regression algorithms. However, such databases exist for few sensors and with the current trends of large-scale projects and data sharing, may be generated for additional sensors. In addition, the authors show that even in a "mature" protein like GCaMP they can identify regions that were neglected in previous optimization cycles, which further demonstrated the power of their approach and emphasize the potential for less-explored proteins.
Some weaknesses are noted below, first some general comments and then more specific questions or concerned. Overall, upon a satisfactory revision of the manuscript, this paper can be a good fit for the journal, as it highlights a new path to engineer protein sensor and overcome a major experimental hurdle of screening through a huge amount of candidates.

General comments:
1. It seems the model is currently limited to predicting the effect of one mutation at a time, or at least, this is how it is implemented in this manuscript. I suggest discussing this topic and how it may develop in the future in the Discussion section.
2. The authors used two databases that were derived from experiments with cultured neurons but implemented their model's predictions on HEK cells. The section where they tested the new variants with cultured neurons doesn't include a systematic comparison of the model prediction as in Fig. 3. This seems like a flaw in the logic of the manuscript and is not addressed. The authors should include these comparisons (if exist) and discuss changes between the two assays and how they affect the model's prediction accuracy.
3. There is no in vivo data included in the paper to validate that the new eGCaMPs sensors work as good as expected (see comments below regarding different values in screening experiments compared to previously reported values). Since the main novelty of this work is in the way it implemented ML-based approach, this is not a fundamental issue.
4. The DF/F0 amplitudes and decay times the authors report (Fig. 5A-E) are substantially different than previously reported values for the GCaMP6s and mainly the jGCaMP7s sensors. The 1AP response amplitudes are very low, the decay times are very slow. Do the authors have any explanation for these changes? This should be referred in the manuscript as well.
5. The paragraph in page 5 lines 11-28 is an excellent demonstration of the power of the presented work to explore the mutation space in a way that is hard to do experimentally. I think it can be further emphasized in the discussion.

Specific comments:
1. Page 4 line 7: The encoding of the amino acid properties, and the nature of these properties should be better explained. The current explanation is vague (including the

Methods and supplementary information parts) and doesn't generate a coherent picture of why the authors picked this model for describing the AA, why 5 parameters were chosen, whether some parameters were more "predictive" than others across AAs, etc.
2. Page 4 lines 17-20: the last sentence of the paragraph is not clear.
3. The authors use the term "fluorescence" to describe the DF/F0 changes (Fig. 1A). Although this term is defined in the paper, it is quite confusing in respect to the way it is used in the literature. Fluorescence will generally describe the raw signal, and not the DF/F0, and aspects that relate to the fluorescence will also include the baseline fluorescence levels, maximal fluorescence level, and bleaching rate. Since the authors limit themselves to consider DF/F0, then maybe they should use this term explicitly.
4. Page 6 line 5: I think the authors should better clarify the "retrospective analysis", which is mentioned here and in other places across the manuscript. What exactly was done?
5. Page 6 lines 11-12: can you add quantification to the multiple examples that are mentioned?
6. Page 8 Lines 1-2 (Fig 4): Where these changes significant?
7. Page 8 lines 4-21: Missing – quantification of the accuracy of the models' prediction for the performance of GCaMP variants in cultured neurons (similar to the section that studied the performance in the HEK cells). Missing – what is the agreement between HEK cells and cultured neurons assays? The authors should also explain why the HEK cells assay is required (is that throughput?).
8. Fig. 1C : It can help to add labels (Calmodulin, GFP, CBD).
9. Fig. 2A: The bubble plot is not clear. What is presented there?
10. Fig. 3B-D: What are the dotted lines in panels B-D?
11. Fig. 3D: Why were the variants arranged in that order?

Reviewer #3 (Remarks on code availability):

The code provides detailed information for the users to install the software and reproduce the data, including a readme file, a demo movie with step-by-step installation guide, and the input data used for generating the data in the paper.

Reviewer #4 (Remarks to the Author):

Genetically encoded fluorescent sensors play a crucial role in monitoring neural activity and neurochemicals. To achieve optimal in vivo performance, the iterative optimization of these sensors often entails extensive mutagenesis and screening. To enhance the efficiency of this optimization process, the authors employed a machine learning ensemble to predict potential beneficial mutations. By integrating these identified mutations, the authors successfully improved calcium sensors with faster decay kinetics and a high fluorescent response. This study introduces a valuable strategy with the potential to be adopted to optimize various other sensors. There are several minor concerns that I hope the authors will address.
1.The authors get improved eGCaMP series sensors. However, the photophysical properties of these sensor, including affinity, extinction coefficient and quantum yield have not been fully characterized. This information holds significance for end-users as they consider the utility of these sensors.

2.In figure 4F, it is recommended to clearly label which variants are eCaMP+ and eCaMP2+. Besides, it will be clearer to align the bar color in figure4F with the trace color in figure4G.
3.In Figure 5, it would be beneficial for the authors to include a comparison of the signal-to-noise ratios between different sensors.

Author Rebuttal to Initial comments

nature research

Reviews: Nature Computational Science

S.Wait et al.

Table of Contents:

Dear Dr. Pan and Reviewers,

Thank you for taking the time to provide detailed feedback on our manuscript. Your comments have helped us to refine our work, improve clarity, and ensure completeness. We sincerely appreciate your overall positive and constructive response. For example, reviewer #1 remarked on our "interesting and innovative" work. Reviewers #2 and #3 stated that the manuscript is "clearly written and the result is basically worth publishing", and that "the concept is exciting". Reviewer #4 found that the study "introduces a valuable strategy with the potential to be adopted to optimize various other sensors".

We have addressed each comment and made the necessary modifications to the manuscript. For example, we characterized the photophysical properties of our engineered eGCaMPs together with existing alternatives (Suppl. Table 6). We also conducted *in vivo* experiments to validate the enhanced performance of eGCaMPs in behaving animals using fiber photometry (Suppl. Figure 8). Other comments concerned the accuracy of the machine learning approach, how generalizable our results are, the cross-compatibility of different biological preparations, and the choice of specific ML models and parameters. As a result, we extensively revised the manuscript, highlighting our approach and rationales in much more detail (see Suppl. Figure 9, Results, Discussion, and Methods).

The comments helped tremendously to further increase the impact of this study and its relevance for the readers of *Nature Computational Science*. Together with our responses below, we submitted a revised and annotated version of our manuscript, highlighting all changes in blue fonts. In summary, this work underscores the power of data-driven approaches to accelerate complex projects in biological sciences, such as mutational sensor engineering, while greatly reducing cost and resource commitments.

We look forward to your feedback and hope the revised manuscript meets the high standards of *Nature Computational Science*. Thank you once again for your valuable input that helped improve our work.

Sincerely,

Andre Berndt and Sarah Wait

S.Wait, 2023

Reviewer #1 (Remarks to the Author):

Berndt and coworkers describe their efforts to apply machine learning methods to develop improved versions of the GCaMP calcium ion indicator. To train their machine learning algorithm, they used previously reported data for 1078 variants that were screened during the development of GCaMP6 and jGCaMP7. The two key properties they focussed on were the fluorescent response to one action potential and the fluorescent decay time. The model was also trained on amino acid properties such as size and polarity, which seemed to improve the predictive properties of the models. Sequences that were predicted by the model to have either the biggest changes in the fluorescent response and kinetics were tested experimentally.

The authors clearly explain how the L317 variants had the opposite effect from the prediction. While they go on to rationalize why this is the case. While I don't expect any sort of protein sequence prediction algorithm to be perfect, the whole point of this work is to try to demonstrate the utility of the machine learning algorithms for predicting mutations that would improve the performance. With the L317 mutations, it is clear that the algorithm correctly identified an important "hot spot" in the protein sequence where mutations were likely to impact the performance.I am certain that a well-trained biochemist could have made the same prediction. Indeed, the fact that Dana et al previously tested mutations at this position, demonstrates that this is the case.

Based on the machine learning predictions, and subsequent in vitro testing, the authors identified jGCaMP7s L317H as the most promising variant. As noted above, this machine learning algorithm had predicted that this variant would have decreased fluorescent response. To continue to improve this variant, the authors resorted to conventional (that is, empirical) protein engineering. They transplanted the L317H mutation to jGCaMP8f, and tested various combinations of promising mutations in the jGCaMP7s scaffold, ultimately leading to the identification of 3 improved variants (eGCaMP, eGCaMP+, and eGCaMP2+). These improved variants also showed improved performance in primary neurons.

Overall, I found this work to be interesting and innovative. We greatly appreciate the ambition and goals of the work, and I am excited about the potential for machine learning algorithms to accelerate protein engineering and ultimately provide better GEFIs than would otherwise be attainable. However, I also found myself unconvinced that the machine learning provided any valuable insight, beyond what an appropriately trained protein chemist could have gained by studying the literature and the crystal structures. Though, I also appreciate that maybe the goal here is to match the ability of expert, and not necessarily exceed it? If the goal is to match the insight of an expert (who still might make imperfect predictions, like the L317 mutations), then I would consider this work a success. If the goal is to exceed the abilities of an expert, then I don't consider this work to be a success. Either way, I feel that there needs to be discussion that frames this work in this context, which is truly fundamental to this and all other machine learning and AI efforts.

S.Wait, 2023

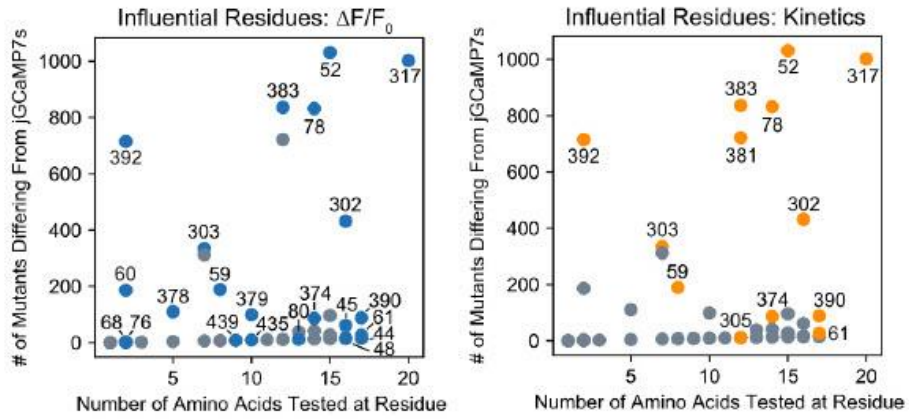
In addition to the fundamental concern described above, I have three other major concerns related to the protein engineering aspects of this work. I will leave it to other reviewers to comment on the appropriateness of the computational methods.

We are grateful for the reviewer's comments on the innovativeness of our work and additional thoughts that reflect many of the fundamental points we discussed when developing the framework for this study. Fundamentally, we aimed to demonstrate that machine learning, a powerful tool for pattern recognition and predictions, can be used to guide the engineering of sensor proteins. In this regard, our study undoubtedly succeeded. It is possible that the identified mutations could, eventually, have been identified by an experienced expert. However, our approach showed that an alternative, ML-based approach can be successful at a much faster speed, with less prior knowledge, and by using fewer resources. Our goal was to identify trends within large mutational datasets and to test a select smaller number of variants in biological host systems. For example, instead of screening the 1492 novel mutations individually by laborious, time-intensive benchwork, our approach succeeded by doing most of the screening *in silico*. Eventually, we tested only about 20 selected variants *in vitro*. Because the training data is derived from an expert, the ML models can learn from it to provide insightful analysis. In other words, the current configuration of the model is best suited for extrapolation as opposed to exploration. In the future, we anticipate the underlying training data can be created specifically with machine learning in mind, i.e., data that is free from empirical biases inherent in many ML datasets. Specifically, we plan to utilize this pipeline together with our own or other high-throughput screening platforms capable of generating large mutational datasets. We will address the additional comments individually below.

1. The data set used for training is highly biased. As far as I understand, the residues to be mutated were chosen based on inspection of the crystal structure to identify mutations that were most likely to have an impact on the GCaMP function. I assume that this bias must propagate through every aspect of this work and the every prediction that the algorithm makes. The authors will need to more clearly address this concern, and explain why (or why not) it is a limitation or this work. I feel that a useful discussion point would be to provide some perspective on what the ideal training set for such efforts would be, as this may inspire other workers to collect the appropriate data.

This is a very astute observation that we have thought about extensively in our current work. We agree that the dataset itself is inherently biased due to the mutated residues being chosen through crystal structure analysis and, as mentioned, expert insight. Whether this is truly a limitation is difficult to ascertain. We did observe that highly mutated positions tend to come to the forefront of final predictions; however, this does not mean that they are not influential or that the mutations that the ensemble suggests cannot be exploited further. Even with the biases inherent to the mutation library, it did not preclude less explored residues from being chosen as influential in sensor performance. For example, in the graph below, we see that the residues with large numbers of amino acids tested (317, 52, 390) get pulled out as important in both the kinetics and $\Delta F/F$ predictions. However, we observed that mutations such as 392, 303, and 59 had relatively

few amino acids tested at each position but were still pulled out as impactful on both biophysical properties.



In creating an ideal dataset for machine learning, we recommend several key steps. First, users should define the sequence space and dimensionality for the acquired data. Smaller dimensionality offers more in-depth analysis and comprehension of combinatorial mutations. At the same time, larger numbers of residue positions will begin to limit the number of mutations that should be tested in combination. Future data acquisition should characterize equal numbers of mutations per residue to avoid any potential biases that may arise due to prevalence. Furthermore, identifying 'loss-of-function' mutations is as vital to proper training as 'gain-of-function'. Iterative model training is an ideal application of this technology, but testing only promising variants should be avoided, as this may introduce bias into the dataset.
We ultimately benefitted more from the well-characterized mutation library that was already published than what was detracted via biased mutations. Although the GCaMP library was not formed with the ML application in mind, the information found within was still incredibly valuable and the driving force behind this work. In addition to inspiring others to collect data appropriate for machine learning, we also want to impress upon the readers and the broader scientific community that a wealth of data exists and can be analyzed in this way.

We have additionally added the below paragraphs to the discussion:

"Lastly, we acknowledge that the dataset used to train the ensemble was more biased toward influential residues, as the mutated residues were chosen through crystal structure analysis and previous empirical insight. Whether this is truly a limitation is difficult to ascertain. We observed that highly mutated positions tend to come to the forefront of final predictions; however, this does not mean that they are not influential or that the mutations that the ensemble suggests cannot be exploited further. Likewise, even with biases in the mutation library, it did not preclude less explored residues from being chosen as influential in sensor performance. We believe that we ultimately benefitted more from the well-characterized mutation library that was already published than what was detracted via biased predictions. Although the mutational dataset was not

intentionally formed with machine learning in mind, the information found within was invaluable and capable of training machine learning models.

As machine learning studies become more prevalent, several considerations for data acquisition may help generate datasets better suited for machine learning extrapolation. First, sequence space and dimensionality have to be well-defined. Smaller dimensionality offers more in-depth analysis and comprehension of combinatorial mutations. At the same time, larger numbers of residue positions will span a much greater sequence space but limit the study to small iterations from the starting sequence. Data acquisition should have equal numbers of mutations per residue in their characterization to avoid any potential biases that may arise due to unbalanced prevalence. Furthermore, identifying 'bad' mutations is as vital to proper training as 'good' mutations. The use case of iterative model training, in which the user is informed by machine learning and then retrains the model with additional information, is an ideal application of this technology. However, testing only promising variants should be avoided, as this may introduce bias into the dataset during retraining. Testing mutations at sites where the ensemble shows significant variability in the predictions can increase understanding."

2. In the section on in vitro testing of predicted mutations, the authors explain how some mutations gave the expected response, yet others gave the opposite response. Somehow, the authors will need to quantify the overall accuracy of their predictions and convince the reader that their predictions, in sum, are better than would be expected based on random chance.

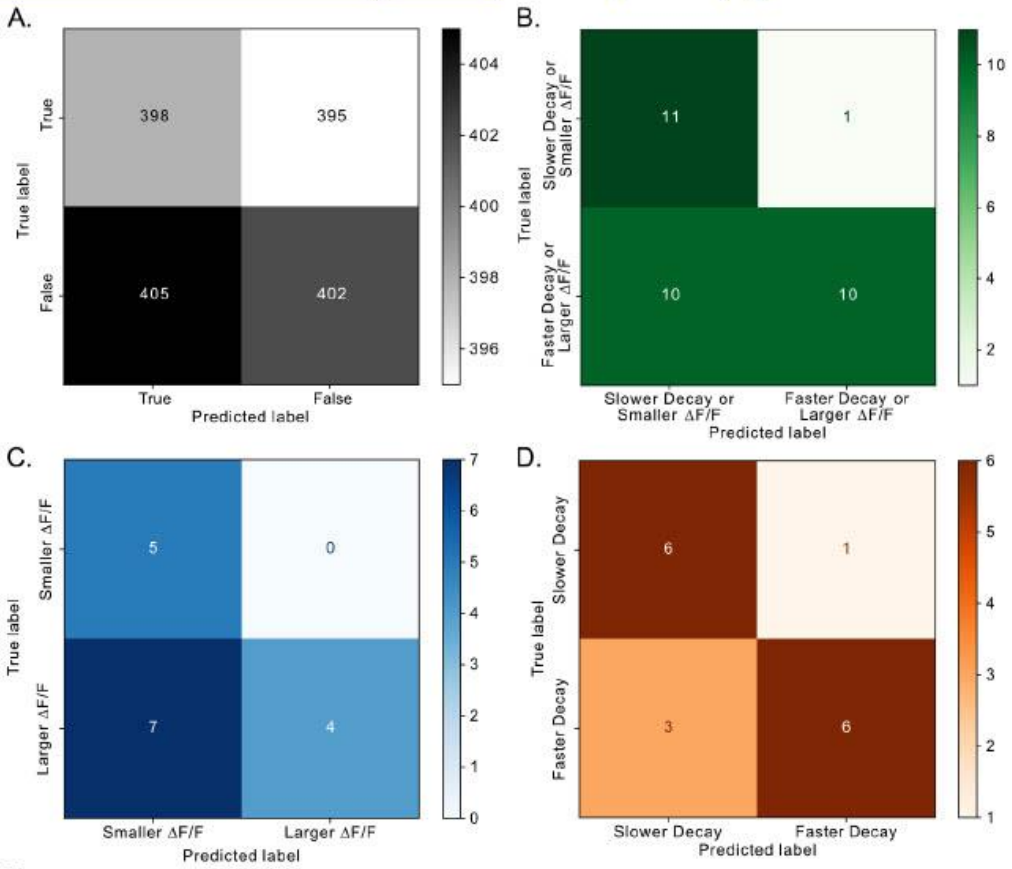
We appreciate the comment and the chance to discuss the accuracy in more detail. To address this comment, we have included confusion matrices for random chance (ie, 50/50 chance of success vs. failure), the ensemble (both models), the ΔF/F model, and the kinetics model (see below).

We categorize our predictions as binary outcomes, classifying kinetics predictions into variants that are either faster or slower than jGCaMP7s, and ΔF/F predictions into variants containing a larger or smaller ΔF/F than jGCaMP7s. To evaluate our model's performance, we computed an accuracy score using the empirical data, which is the ratio of true positives and true negatives to the total number of predictions.

To gauge whether our accuracy scores surpassed what could be expected by random chance, we compared them to the accuracy score of an experiment with an equal probability of success/failure. A standard method for testing equiprobability in binary outcomes involves repeatedly tossing a fair, unbiased coin (Reimers, Donkin, and Le Pelley 2018). We simulated a coin toss experiment computationally, where we recorded 1600 consecutive coin tosses (16 trials x 100 times) to represent **predicted** values for true/false and 1600 consecutive coin tosses to represent **observed** true/false. The accuracy for the coin toss experiment was 0.5, mirroring a genuinely random change in binary outcomes.

In our ensemble quantifications, ΔF/F model quantifications, and kinetics model quantifications, we see that, in every instance, the precision and accuracy are greater than 0.5. This means that the probability of the ensemble prediction matching empirical observation is greater than what would be expected from the coin toss and, thus, random chance. Additionally, it is worth noting that the "random chance" comparison for protein engineering, in actuality, falls far below a 0.5

chance of success and has a much larger probability of failure (Smith 1970; Yang, Wu, and Arnold 2019). Where appropriate, we have included the model accuracies in the results: *In Vitro Performance of Ensemble Predictions*, *Discussion*, and as Supplementary Figure 9.



E.

Model	Accuracy	Precision
Random Chance (50/50)	0.5	0.5
Fluorescence + Kinetics	0.66	0.91
Fluorescence	0.56	1
Kinetics	0.75	0.86

Estimation of Model Accuracy with Acetylcholine Results. (A.) Confusion Matrix of 100 simulated coin flip experiments (each containing 16 samples). (B.) Confusion matrix of all both models predictions and acetylcholine performance. (C.) Confusion matrix of all fluorescence model's predictions and acetylcholine performance. (D.) Confusion matrix of all kinetics model's predictions and acetylcholine performance. (E.) Quantification of each confusion matrices accuracy and precision.

3. Finally, I feel that there needs to be a bit more nuance applied to the comparisons between the three new variants and the previously reported GCaMP variants. I believe that previous workers settled on their final variants (GCaMP6s, jGCaMP7s, etc), as the best possible compromises, taking all of the properties into account. During their screening, the previous workers certainly found variants that were improved in one property or another (and are thus in the training sets), but these are not necessarily the final variants that they settled on. The authors will need to acknowledge this and write their comparisons and discussion accordingly.

We agree with this statement, and in response to this comment, we have conducted additional experiments. Our results now include the photophysical properties of our variants, including their Kds, hill coefficients, extinction coefficients, and quantum yield, which should further benchmark their capabilities compared to previously published variants (Suppl. Table 6).
In discussions with the previous authors, namely Dr. Hod Dana, he mentioned that they decided on their published variants based primarily on kinetic capability and maximal fluorescence. He mentioned that they split the variant performances into kinetic "regimes," i.e., slow, medium, and fast, and then sorted them based on "maximal fluorescence," which is the baseline fluorescence + $\Delta F/F_o$. The variants they published had the highest maximal fluorescence ranking within each regime. Depending on where they set their kinetic cutoffs, some variants may rank better than those in other regimes but not outcompete the best within their kinetic capability. Similarly, the $\Delta F/F_o$ of one variant may be greater than another within the same regime but be considered lower based on a diminished baseline fluorescence. Our approach does most of the screening in silico and selects variants with the same favorable biophysical properties, i.e. response amplitude $\Delta F/F_o$ or off kinetics. As a result, we screened fewer variants in vitro but achieved similar results while reducing the experimental burden.

We have included the below paragraph in the discussion to address this comment and provide further insight for future studies:

"In previous studies, generating the GCaMP libraries, the authors employed a volume approach, in which they tested over a thousand variants iteratively and chose to fully characterize those determined to have optimal kinetic and maximal fluorescence capabilities. Because of the sheer number of experiments, they split their variants into kinetic regimes and determined the best possible variant within each regime based on multiple biophysical properties[4,5]. The approach we employ here allowed the vast majority of the screening to occur *in silico*, which significantly reduced the experimental burden. Importantly, we trained and selected variants incorporating one of two favorable biophysical properties: $\Delta F/F_o$ or off-kinetics. As a result, we tested fewer variants *in vitro* but achieved similar results while greatly reducing time and resource commitments. The selected eGCaMP variants displayed compensation within favorable biophysical characteristics, such as a lower baseline fluorescence (**Supp. Fig. 7**). However, the lower baseline did not impact the performance of eGCaMPs in neuron cultures or *in vivo* fiber photometry (**Supp. Fig. 8**). Hence, it would be an acceptable tradeoff in many use scenarios. As a consideration for future

studies, metrics for baseline fluorescence or other favorable biophysical characteristics could be included in ensemble training to preserve them within the final variants."

Reviewer #2 (Remarks to the Author):

This paper reports a successful case study about improving the ability of GEFI. It is clearly written and the result is basically worth publishing. However, I have the following comments.

1) The ensemble used here is based on relatively old machine-learning techniques. Please clarify why modern deep learning methods are not preferred here.

We appreciate the reviewer's recommendation for publications and are happy to address the concerns. It is true that some underlying concepts and algorithms, such as K-NN and neural networks, have been used for a longer time than deep learning methods. However, the contemporary application of these techniques in modern machine learning settings, along with advancements in computing power and data availability, makes them relevant and modern approaches specifically to regression problems.
It is important to acknowledge that there are numerous approaches to solving the same task. We do not doubt that an approach undertaken with the resources to support modern deep learning algorithms would yield comparable outcomes. Our choice of the ensemble models was influenced by several factors. First, their simplicity makes them less prone to overfitting, particularly when dealing with limited data. These models are also easy to interpret and implement, which allows us to comprehend better the interprotein factors influencing the predictions made by the models. Lastly, our ensemble models have lower computational demands compared to deep learning techniques such as Convolutional Neural Networks (CNNs) or transformers. The reduction in computational demand makes this ensemble more accessible to potential users.

2) The search range from the wild type is limited. Single mutations and a few double mutations are considered, but not more. Please discuss if this technique can be used to discover more distant mutations.

That is a very critical point and a good observation. Therefore, we have added the paragraph below in the main text to discuss the choice of point mutations.

"One of the major hurdles of protein engineering is the susceptibility of proteins to experience epistasis, in which combinations of mutations non-additively influence the phenotypic characteristics of a protein[38]. Though the mutation library we worked with had >1000 well-characterized variants, the large number of mutated residues renders the dimensionality incredibly large. For a library such as this, there are 1.18e+91 possible combinations of residues over 70 mutated residue positions, meaning that the variant library is only a small sampling of the theoretical mutation space. We felt that the risk of epistasis upon combinatorial mutation was too great and that the relatively limited size of the library in comparison to its dimensionality rendered this application better suited to single-point mutation testing. Though investigation of a

combinatorial library was not used in this study, others have shown promise using machine learning to engineer protein combinatorial libraries[15]."

3) Not being an expert in this field, I could not really understand how the performance of the new variants compares to the best GEFI available. Please discuss about it.

We are happy to discuss the benchmarking of eGCaMPs against previous state-of-the-art calcium sensors in more detail. Our in vitro assessment of the eGCaMP suite of sensors shows that they have larger amplitudes of $\Delta F/F_0$ responses and faster decay kinetics than the GCaMP6, jGCaMP7, and jGCaMP8 families of calcium sensors. It is difficult to define which sensor from any of these suites is the "best." Usually, users choose which sensor best fits their specific experimental needs based on tradeoffs such as faster kinetics for lower $\Delta F/F_0$ or vice versa. However, we can conclude that eGCaMP$^+$ is faster than jGCaMP8f (previously the fastest published variant) and has larger $\Delta F/F_0$ under most conditions. The dynamic range of eGCaMP^{2+} under saturating conditions (max. $\Delta F/F_0$) is larger than GCaMP6s (which in our hands was the next best-published variant) but has fast kinetics. Therefore, the eGCaMP variants impressively avoid some of the tradeoffs of previous generations and form a class of their own. Nevertheless, eGCaMPs also have lower baseline fluorescence (Supp Fig. 7), which could impact some applications but may be acceptable in others such as in vivo fiber photometry (see new Supp. Fig. 8). Additionally, a subsequent reviewer asked for each sensor's photophysical properties, and in our revised manuscript, we provide a much more detailed and side-by-side comparison of how our variants compare to previous instances (now included in the text and new supplementary table 6). This complete benchmarking of the sensors should give readers and potential users a better overview of our sensor's capabilities and help them make informed decisions.

Reviewer #3 (Remarks to the Author):

Wait et al. present a new approach for machine-learning-based engineering of genetically-encoded calcium (or more generally, fluorescence) indicators. Their approach depends on the existence of a large database of standardized experiments that explore the effect of multiple mutations on a protein sensor of interest. Then, they implemented a combination of three regression algorithms to learn the important parameters that affected fluorescence fluctuations and kinetics when different amino acids were mutated. The model is then used to predict the performance of new mutations, which were never tested before, and finally, the predicted hits were tested in vitro to validate the model accuracy. The main advantage of the presented approach is its ability to saturate the tested mutation space, i.e. to computationally test all the possible combinations of amino acids in multiple specific positions, which is expensive, time- and labor-consuming, but is also experimentally inefficient – as many of the variants are expected to show low performance levels.
The manuscript is well-written, and the concept is exciting. The main limitation to its application to additional types of sensors is the availability of experimental databases that are required to train the regression algorithms. However, such databases exist for few sensors and with the

current trends of large-scale projects and data sharing, may be generated for additional sensors. In addition, the authors show that even in a "mature" protein like GCaMP they can identify regions that were neglected in previous optimization cycles, which further demonstrated the power of their approach and emphasize the potential for less-explored proteins.
Some weaknesses are noted below, first some general comments and then more specific questions or concerned. Overall, upon a satisfactory revision of the manuscript, this paper can be a good fit for the journal, as it highlights a new path to engineer protein sensor and overcome a major experimental hurdle of screening through a huge amount of candidates.

We are thrilled to read that the reviewer found the concept "exciting." We are happy to discuss the fundamentals and future applications in more detail. For example, as the reviewer noted, larger-scale approaches are becoming more common in biology, which justify or even require machine learning for data analysis. The applications are still limited in protein engineering, but we and others are actively developing approaches that can generate mutational sensor libraries specifically for analysis by machine learning. We added generalizable requirements for future screens into the Discussion section, while this manuscript provides a roadmap for future sensor engineering incorporating ML.

General comments:
1. It seems the model is currently limited to predicting the effect of one mutation at a time, or at least, this is how it is implemented in this manuscript. We suggest discussing this topic and how it may develop in the future in the Discussion section.

We added the paragraph below into the Discussion explaining why we restricted this study to point mutations.

"One of the major hurdles of protein engineering is the susceptibility of proteins to experience epistasis, in which combinations of mutations non-additively influence the phenotypic characteristics of a protein[38]. Though the mutation library we worked with had >1000 well-characterized variants, the large number of mutated residues renders the dimensionality incredibly large. For a library such as this, there are 1.18e+91 possible combinations of residues over 70 mutated residue positions, meaning that the variant library is only a small sampling of the theoretical mutation space. We felt that the risk of epistasis upon combinatorial mutation was too great and that the relatively limited size of the library in comparison to its dimensionality rendered this application better suited to single-point mutation testing. Though investigation of a combinatorial library was not used in this study, others have shown promise using machine learning to engineer protein combinatorial libraries[15]."

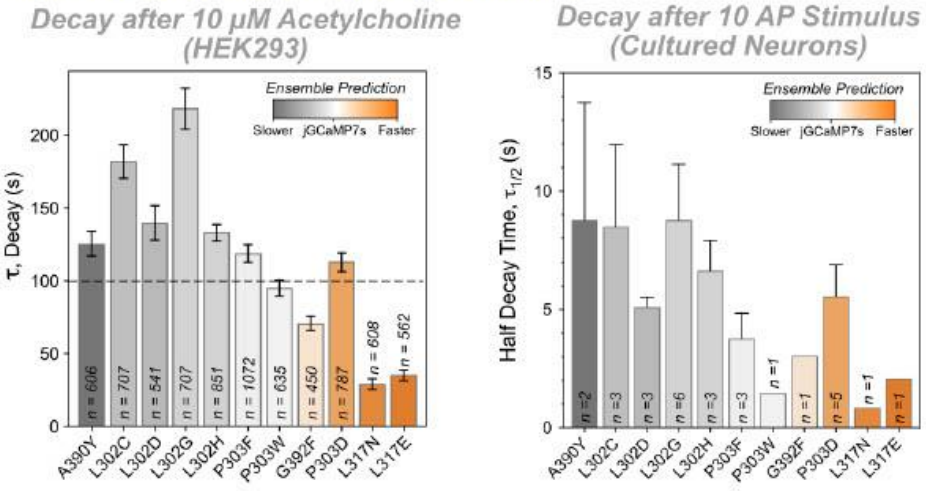
2. The authors used two databases that were derived from experiments with cultured neurons but implemented their model's predictions on HEK cells. The section where they tested the new variants with cultured neurons doesn't include a systematic comparison of the model prediction as in Fig. 3. This seems like a flaw in the logic of the manuscript and is not addressed. The authors should include these comparisons (if exist) and discuss changes between the two assays and how they affect the model's prediction accuracy.

S.Wait, 2023

We agree that the ideal configuration to measure our model's capability should be against data acquired in the same manner as the training data. While our field stimulus protocol was similar to that of previous publications, there were some differences in the final implementation, as noted in our response to comment #4 (below). However, we observed comparable response patterns, where the response to 1 AP from GCaMP7s was larger than either 6s or 6f, and the half decay times of GCaMP7s were larger than 6s, which in turn were larger than 6f. The primary limitation was throughput and sample size acquisition within neuron culture screening.

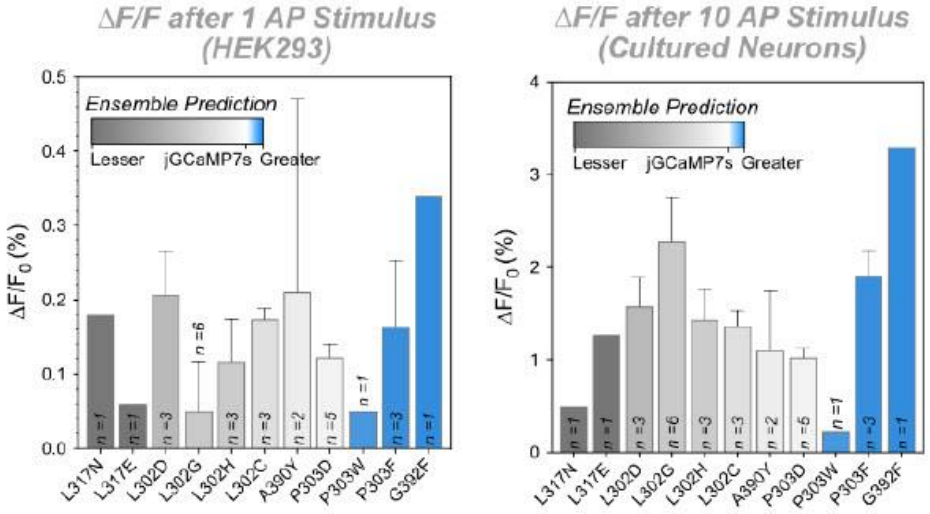
Due to the low throughput nature of the culture neuron screens, we opted to perform an intermediate acetylcholine assay step in HEK293 cells. The primary benefit of the acetylcholine assay is our ability to screen thousands of cells for any given variant efficiently. Despite being a different host system, the acetylcholine assay provided a close approximation of sensor capabilities prior to transitioning promising variants into cultured neuron assays.

In response to the comment, we conducted an additional neuron culture screen with the single-point mutation variants and compared the responses to those obtained using acetylcholine. Despite a limited sample size, the acetylcholine screen largely reflected what we observed in cultured neurons. Specifically, the kinetics assay results from the acetylcholine assay closely matched our observations in cultured neurons, even with a low sample size. This is further compounded by an excellent accuracy score of 0.75 for the kinetics model's predictions in HEK cells *in vitro* (new Supp. Fig. 9). Based on the similarity between the two assays; we expect that the neuron culture screen accuracy score would be similar.



Within the ΔF/F testing, we did see some dissimilarities in the acetylcholine assay and the neuron responses; however, the differences aligned better with the machine learning predictions. This is illustrated in Figure 5A, where the 1AP response from eGCaMP (L317H) is diminished compared to jGCaMP7s, which is consistent with the model's predictions. In the neuron culture screen, we begin to see this same effect where the model predictions appear to reflect what was observed in neurons, particularly with the L317 variants.

S.Wait, 2023

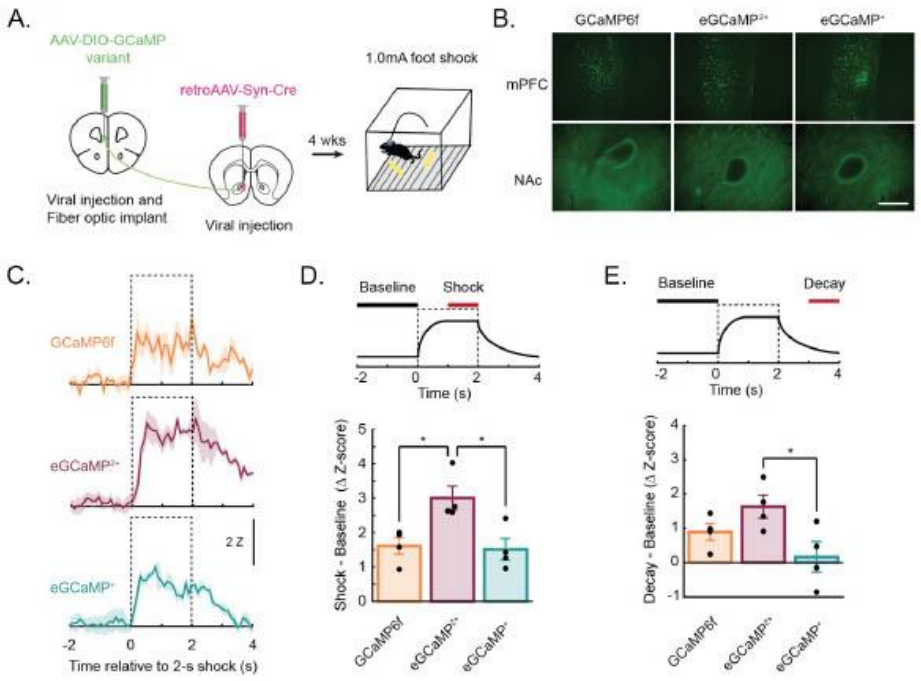


Overall, the acetylcholine assay in HEK cells proved to be a close approximation for assessing sensor capabilities in neurons, offering significantly higher throughput. We acknowledge differences in experimental approaches between labs but emphasize that these did not compromise the ensemble's ability to guide our engineering efforts. Our approach yielded multiple variants of interest and highlighted potential mutations for future exploration by other researchers and developers.

3. There is no in vivo data included in the paper to validate that the new eGCaMPs sensors work as good as expected (see comments below regarding different values in screening experiments compared to previously reported values). Since the main novelty of this work is in the way it implemented ML-based approach, this is not a fundamental issue.

We agree that the main focus of the manuscript is on the ML part. However, many potential users of eGCaMPs would be interested in confirming the in vitro results *in vivo*. Therefore, we added expression and stimulus results of our eGCaMP^{2+} and eGCaMP^{+} sensors (currently available on addgene) from the mouse prefrontal cortex, measured by fiber photometry *in vivo* (New Supp. Fig. 8). Confirming their *in vitro* performance, we found that eGCaMP^{2+} displayed a significantly larger $\Delta F/F_o$ compared to GCaMP6f. Similarly, eGCaMP^{+} displayed the fastest decay after activation. We have included these findings as Supplementary Figure 8 and included results/discussion within the main text!

Supplementary Figure 8: In Vivo Performance of eGCaMP+ and eGCaMP2+ expressed in mPFC



4. The DF/F0 amplitudes and decay times the authors report (Fig. 5A-E) are substantially different than previously reported values for the GCaMP6s and mainly the jGCaMP7s sensors. The 1AP response amplitudes are very low, the decay times are very slow. Do the authors have any explanation for these changes? This should be referred in the manuscript as well.

While we developed a neuron culture screen similar to that of the previous authors, it is difficult to replicate the experimental setup entirely. Therefore, we developed our paradigms as the closest approximation. Then, we thoroughly benchmarked the new eGCaMPs against the previously published variants in various biological host systems under identical conditions. Thus, even when the magnitude or speed of response of individual variants differed, we reproduced the relative trends between different sensors, which reflected what previous authors had published. Importantly, we showed how our sensors performed under identical experimental conditions against other published variants. The relative performance of GcaMP variants was largely recapitulated in HEK cells and *in vivo (*new Supplementary Figure 8*)*.

To shed light on the observed variations, we highlight some key differences in our experimental setup compared to previous studies (as reported in the Methods section), which may contribute to the disparities in the reported $\Delta F/F_o$s and kinetics.

Experimental Setup Differences:

- Different Neuron Sources:
  - The source of the previous work's neurons were from the hippocampus (Wardill et al. 2013), whereas we chose to derive our neurons from the cortex, where we could acquire many more healthy cells.
- Different Promoters:
  - In our experiment, we used variants driven by the CAG promoter, whereas previous authors used the Syn promoter (Wardill et al. 2013). It is possible that the differences in strengths of these promoters led to differences in both expression speed and levels (which could affect the baseline, and thus $\Delta F/F$)
- Different Transfection Methods:
  - We used calcium phosphate as a method to transfect our neuron cultures, while the other studies used viral transfection. This could affect transfection efficiencies and expression rates.
- Different Wires
  - We used silver electrodes, whereas the previous authors reported using platinum electrodes.
- Different imaging media
  - Previous Authors Imaging Solution: 145 mM NaCl, 2.5 mM KCl, 10 mM glucose, 10 mM HEPES pH 7.4, 2 mM CaCl2, 1 mM MgCl2
  - Our Imaging media: 150 mM NaCl, 4 mM KCl, 10 mM glucose, 10 mM HEPES pH 7.33, 3 mM CaCl2, 1 mM MgCl2
- Inhibitor Usage:
  - We observed these calcium transients without using any inhibitors, whereas the previous authors used certain neuronal receptor antagonists.

In response to your suggestion, have incorporated the following into the manuscript, discussing these variations and their potential impact on the reported results.

"We included these previously published variants to benchmark the responses from our sensors under identical experimental conditions."

5. The paragraph in page 5 lines 11-28 is an excellent demonstration of the power of the presented work to explore the mutation space in a way that is hard to do experimentally. I think it can be further emphasized in the discussion.

We appreciate the comment and have added a paragraph to the discussion highlighting the residue interactions and their promise in future studies.

"With the functional predictions gathered from the model, we were not only able to gather mutations that directed sensor engineering but also able to observe the learning and predictive patterns to better understand the protein function. For example, when we mapped the residues the ensemble predicted would be influential back onto the GCaMP crystal structure, we found that the highlighted residues were in structurally significant parts of the GCaMP protein and often faced inward toward one other. This phenomenon may indicate that the ensemble is learning which residue interactions are important for protein function and govern the given biophysical property. As such, these residue interactions constitute a promising basis for further mutational studies and may even be used to influence future mutation library generation."

Specific comments:
1. Page 4 line 7: The encoding of the amino acid properties, and the nature of these properties should be better explained. The current explanation is vague (including the Methods and supplementary information parts) and doesn't generate a coherent picture of why the authors picked this model for describing the AA, why 5 parameters were chosen, whether some parameters were more "predictive" than others across AAs, etc.

We agree that the readers would benefit from more detailed information and rationals. We have edited the text and added more information where appropriate. The responses to individual points and edits to manuscript sections are listed below.

The encoding of the amino acid properties
To perform the encoding, we replace each amino acid in the sequence with the corresponding value from the property dataset, i.e., the float type value that exists for that amino acid's position in the property dataset list. *[see Methods: Data Preprocessing]*

nature of these properties should be better explained.
AAINDEX consists of 554 complete matrices that each describe a different AA property, such as size, polarity, or hydrophobicity. The general shape and composition of each one of the property datasets is a list of 20 float type values, in which the order is linked to the amino acid, and the float type value is a quantitative value that is dependent on the property in question. *[see Methods: Data Preprocessing, and Results: Description of Variant Library, Computational Approach, and Predictions on Novel Sequences].* The specific AA properties that were selected for each ensemble are listed in Supplementary tables 1 and 2. We also added a more in depth discussion on the AA properties and their impact in the "Discussion" section.

why the authors picked this model for describing the AA

We needed to choose an encoding method in order to discretize our sequence and make it available for interpretation by the models. There are previous examples in which authors performed both one-hot encoding (Bedbrook et al. 2019) and amino acid property encoding (Saito et al. 2018) that both led to successful machine learning-guided engineering of protein properties,

and both are valid methodologies (Yang, Wu, and Arnold 2019). We evaluated each of these encoding methods alongside label encoding and found that our encoding using 5 datasets led to the best predictive outcomes in our hands (Supplementary Figure 2C). *[Results: Description of Variant Library, Computational Approach, and Predictions on Novel Sequences]*

why 5 parameters were chosen,

The reason we chose to include five property datasets in the final ensemble prediction was twofold. The first was that, during our training, we found that the top-performing datasets often achieved R^2 values that were remarkably similar (**Supp. Table 1&2; Supp. Fig. 2**). Given the marginal superiority of the top-performing dataset over its counterparts, a strategic choice was made to include additional matrices. The selection of five datasets was made semi-arbitrarily, as it afforded the desired additional insights without dramatically impacting computational demands, processing time, and storage requirements. The second reason we chose to include more was due to the type of ensemble we were performing. Within the stacked ensemble, each final ensemble prediction was determined through unweighted averaging. This method is not free from outlier corruption, meaning that if one model's prediction is vastly different from the others, it will influence how that prediction is considered in the final ensemble predictions. The addition of more datasets/models enables some buffering to happen and for a large sample size to determine our ensemble's mean predictions. *[see Methods: Ensemble Training]*

whether some parameters were more "predictive" than others across AAs, etc.

This is the discussion we aim to promote in Supplementary Figure 2 by describing the top-performing properties in Supplementary Tables 1 and 2.

Within supplemental figure 2, we aimed to provide some insight into the top-performing datasets and whether or not there was a pattern in each model's learning. In our analysis, we took all 30 top-performing datasets (15 from each model) and performed PCA clustering to garner an idea of the degree of similarity between each dataset.

We evaluated the contents of each cluster (Supp. Fig. 2D) to identify a shared parameter within them that is driving predictive capabilities. For example, we found that the parameter 'Interactivity scale obtained from the contact matrix' clustered in group 3, alongside parameters such as 'Average surrounding hydrophobicity,' 'Modified Kyte-Doolittle hydrophobicity scale,' and 'Normalized hydrophobicity scales for alpha/beta-proteins.' This gave us some insight into the fact that even though that parameter was associated with high predictive capabilities, it may be due in part to its similarity with hydrophobicity matrices, which may be the driving interaction within the protein. This enabled us to ascertain the underlying properties that lead to our model decisions (i.e., altering amino acids' hydrophobicity at key residues).

We found that parameters that described hydrophobicity were commonly associated with higher-performing predictive capabilities in the ΔF/F model, meaning that some of the protein behavior modifications may be partly due to key hydrophobic interactions. In comparison, parameters

associated with protein folding and energetics were common amongst the higher-performing predictive capabilities in the kinetics model.

While the information in this figure does not drive the predictions that we test downstream, it is an exciting exploration into how the model was able to learn and served as a lens into the interprotein interactions we may be missing. *[Discussion]*

Text included in the discussion:
"Analysis of the top performing datasets within each model additionally provides insight into how the model was able to learn and served as a lens into the interprotein interactions. For instance, we found that AA property datasets that described hydrophobicity were commonly associated with higher-performing predictive capabilities in the ΔF/F model (**Supp. Fig 2B-D; Supp. Table 1**), meaning that some of the modifications in protein behavior may be due in part to key hydrophobic interactions. In comparison, AA property datasets associated with protein folding and energetics were common amongst the higher-performing predictive capabilities in the kinetics model (**Supp. Fig 2B-D; Supp. Table 2**)."

2. Page 4 lines 17-20: The paragraph's last sentence is unclear.
We agree that this was a run-on sentence and not clear in the way that it was written. We removed some of the quantification from the text, disambiguating the message we aimed to present.

3. The authors use the term "fluorescence" to describe the DF/F0 changes (Fig. 1A). Although this term is defined in the paper, it is quite confusing in respect to the way it is used in the literature. Fluorescence will generally describe the raw signal, and not the DF/F0, and aspects that relate to the fluorescence will also include the baseline fluorescence levels, maximal fluorescence level, and bleaching rate. Since the authors limit themselves to consider DF/F0, then maybe they should use this term explicitly.
This is a fair critique, and we understand how the term can be too ambiguous for the context in which we meant it. We have replaced all instances of fluorescence with $\Delta F/F_0$ (where appropriate). Thank you for your insight!

4. Page 6 line 5: We think the authors should better clarify the "retrospective analysis", which is mentioned here and in other places across the manuscript. What exactly was done?
When we refer to retrospective analysis in the text, it is mostly phrased for deeper analysis of the mutant library. In the context in which we use the phrase, we are isolating all of the variants that exist within the characterized library that harbor a 317 mutation and observing their performance as compared to jGCaMP7s. This was not clear in the way that it was written, so we have rephrased these sentences to describe the analysis performed better.

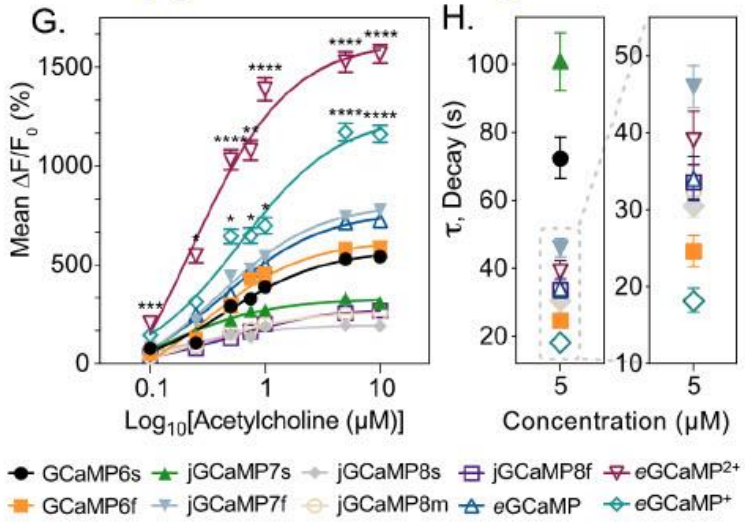
5. Page 6 lines 11-12: can you add quantification to the multiple examples that are mentioned?
We absolutely agree and have added quantification for this line! It now includes that we observed 4 variants that displayed the predicted increase in ΔF/F and five that displayed the predicted decrease in ΔF/F!

S.Wait, 2023

Text changed to: "Regardless, we identified four mutations (P303W, P303F, G392F, G392W) that displayed their predicted increase in ΔF/F0 as well as five mutations (A390Y, L302C, L302H, L302G, L302R) that displayed the predicted decrease in ΔF/F0."

6. Page 8 Lines 1-2 (Fig 4): Where these changes significant?

In the case of eGCaMP^{2+}, the changes in ΔF/F at all concentrations were significant (by student's t-test, tested against jGCaMP7f (the next highest performing published variant). Not all of eGCaMP^{+}'s changes in ΔF/F were significant compared to jGCaMP7f with the same analysis, but We have indicated in Figure 4G levels of significance and listed the p-values in the figure legend. We have also slightly modified the text to delineate significance!

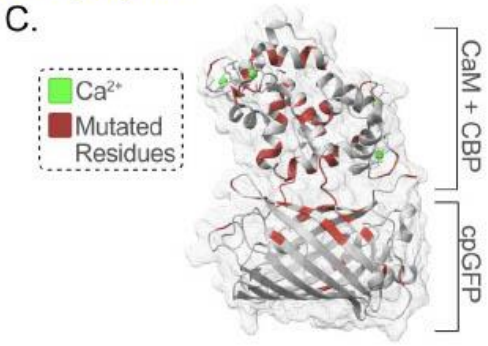


7. Page 8 lines 4-21: Missing – quantification of the accuracy of the models' prediction for the performance of GCaMP variants in cultured neurons (similar to the section that studied the performance in the HEK cells). Missing – what is the agreement between HEK cells and cultured neurons assays? The authors should also explain why the HEK cells assay is required (is that throughput?).

This is a valid point. We addressed many of these concerns regarding comparing HEK cells and neurons in comment #2 above. We believe the information provided there will also answer this point here.

8. Fig. 1C: It can help to add labels (Calmodulin, GFP, CBD).

We added bracket labels of where the domains are in relation to the structure! We additionally provide a full characterization of all of the mutation positions and their location on the protein in Supplementary Figure 1A, as well as a crystal structure that is color-labeled for the domains in Supplementary Figure 3A. (The updated Figure 1C is provided below). We appreciate the comment on the clarity of our figure panels.



9. Fig. 2A: The bubble plot is not clear. What is presented there?
The specific bubble plot in 2A is meant to be a representative graph to demonstrate how we derive the bubble plot in 2D. We want to show that we take the mutations that the ensemble predicted would have the greatest effect on sensor function (IE, those that occur at the bottom and top of the ranked predictions) and count the number of times that each residue appears. This becomes what we describe as the influential residues for each biophysical characteristic. To address any ambiguity, we have bolstered the description of this process both in the main text and figure caption.

10. Fig. 3B-D: What are the dotted lines in panels B-D?
They indicate the mean of the base construct, jGCaMP7s, and were intended to be visual aids to determine the change compared to the baseline. We will better indicate the identity of these dotted lines in the corresponding figure legends.

11. Fig. 3D: Why were the variants arranged in that order?
The variants in B/C were ordered based on the ranked prediction from the ensemble, whereas we did not form predictions for SNR or performance score and instead listed them in ascending order with the base construct as the first to appear. We have included this in the figure legend to improve clarity.

Reviewer #3 (Remarks on code availability):

The code provides detailed information for the users to install the software and reproduce the data, including a readme file, a demo movie with step-by-step installation guide, and the input data used for generating the data in the paper.

S.Wait, 2023

Reviewer #4 (Remarks to the Author):

Genetically encoded fluorescent sensors play a crucial role in monitoring neural activity and neurochemicals. To achieve optimal in vivo performance, the iterative optimization of these sensors often entails extensive mutagenesis and screening. To enhance the efficiency of this optimization process, the authors employed a machine learning ensemble to predict potential beneficial mutations. By integrating these identified mutations, the authors successfully improved calcium sensors with faster decay kinetics and a high fluorescent response. This study introduces a valuable strategy with the potential to be adopted to optimize various other sensors. There are several minor concerns that I hope the authors will address.

1.The authors get improved eGCaMP series sensors. However, the photophysical properties of these sensor, including affinity, extinction coefficient, and quantum yield, have not been fully characterized. This information holds significance for end-users as they consider the utility of these sensors.

We appreciate the reviewer's valuable insights and comments. We agree that the manuscript would benefit from additional biophysical characterization as these key features interest many readers and potential users. Therefore, we conducted additional experiments on purified proteins. Below are the affinities, hill coefficients, extinction coefficients in the saturated state, and quantum yields in the saturated states of our sensors compared to previously published variants. We have included this data as a supplementary table 6 and discussed the results in the text.

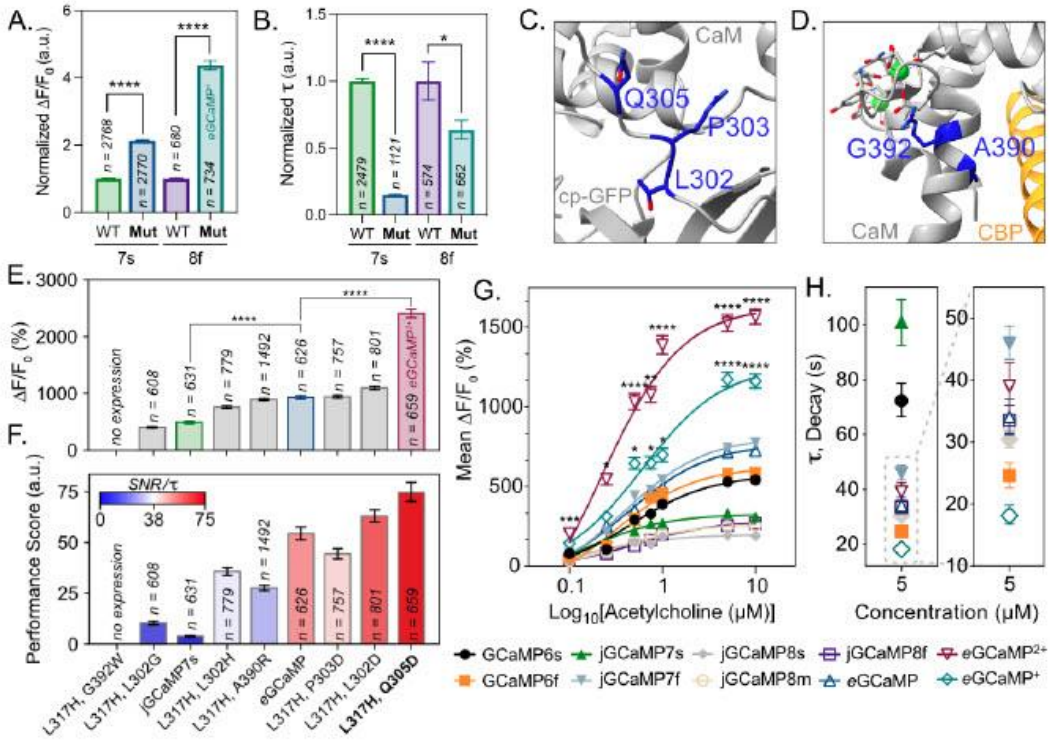
Sensor	Kd (nM)	Hill Coefficient	$\varepsilon_{Saturated}$ (x1000) (M^{-1} cm^{-1})	$\varphi_{Saturated}$
GCaMP6s	120.8 [110.6, 132.2]	2.014 [1.716, 2.387]	N/A	N/A
GCaMP6f	291.3 [256.4, 333.1]	1.857 [1.544, 2.261]	65.276	0.6
jGCaMP7s	46.2 [39.3, 53.7]	2.138 [1.596, 1.918]	N/A	N/A
eGCaMP	354.8 [262.8, 516.4]	1.761 [1.087,3.339]	62.726	0.68
eGCaMP2+	358.7 [310.4, 418.8]	1.925 [1.540, 2.461]	60.070	0.72
eGCaMP+	1885 [1.082, 34.02]	0.9976 [0.4875, 1.871]	58.988	0.63

2.In figure 4F, it is recommended to clearly label which variants are eCaMP+ and eCaMP2+. Besides, it will be clearer to align the bar color in figure4F with the trace color in figure4G.

We have added the color of eGCaMP^{2+} to Figure 4E as well as bold the label at the bottom to indicate its importance. eGCaMP^{+} was discussed in Figure 4A/B, and we attempted to match the colors in these graphs to their corresponding color in 4G, but we can see how the color is slightly off. We will change the color of these bars to match it better and additionally bold the text similar to what is done in 4F.

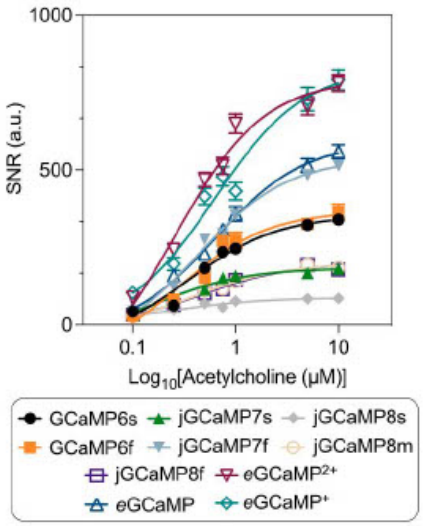
S.Wait, 2023

Figure 4: Mutation Transfer and Combinatorial Mutation For The Identification of eGCaMP+ and eGCaMP2+



3.In Figure 5, it would be beneficial for the authors to include a comparison of the signal-to-noise ratios between different sensors.

We had included this data in Supplementary Figure 7 as a ratiometric comparison originally; however, we additionally included SNR calculations (using Eq.2) for the data collected during the acetylcholine concentration curve so that the side-by-side comparison can be appreciated. This data can now be found in Supplementary Figure 5C. We observed that the trend was consistent, and the eGCaMPs again outperformed the previously published variants.



Citations:

Bedbrook, Claire N., Kevin K. Yang, J. Elliott Robinson, Elisha D. Mackey, Viviana Gradinaru, and Frances H. Arnold. 2019. "Machine Learning-Guided Channelrhodopsin Engineering Enables Minimally Invasive Optogenetics." *Nature Methods* 16 (11): 1176–84.

Reimers, Stian, Chris Donkin, and Mike E. Le Pelley. 2018. "Perceptions of Randomness in Binary Sequences: Normative, Heuristic, or Both?" *Cognition* 172 (March): 11–25.

Saito, Yutaka, Misaki Oikawa, Hikaru Nakazawa, Teppei Niide, Tomoshi Kameda, Koji Tsuda, and Mitsuo Umetsu. 2018. "Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins." *ACS Synthetic Biology* 7 (9): 2014–22.

Smith, J. M. 1970. "Natural Selection and the Concept of a Protein Space." *Nature* 225 (5232): 563–64.

Wardill, Trevor J., Tsai-Wen Chen, Eric R. Schreiter, Jeremy P. Hasseman, Getahun Tsegaye, Benjamin F. Fosque, Reza Behnam, et al. 2013. "A Neuron-Based Screening Platform for Optimizing Genetically-Encoded Calcium Indicators." *PloS One* 8 (10): e77728.

Yang, Kevin K., Zachary Wu, and Frances H. Arnold. 2019. "Machine-Learning-Guided Directed Evolution for Protein Engineering." *Nature Methods* 16 (8): 687–94.

Decision Letter, first revision:

Date: 25th January 24 01:42:59
Last Sent: 25th January 24 01:42:59
Triggered By: Jie Pan
From: jie.pan@us.nature.com
To: berndtuw@uw.edu
CC: computationalscience@nature.com
BCC: jie.pan@us.nature.com
Subject: AIP Decision on Manuscript NATCOMPUTSCI-23-0833B
Message: Our ref: NATCOMPUTSCI-23-0833B

25th January 2024

Dear Dr. Berndt,

Thank you for submitting your revised manuscript "Machine Learning Ensemble Directed Engineering of Genetically Encoded Fluorescent Calcium Indicators" (NATCOMPUTSCI-23-0833B). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Computational Science, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about 10 days. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Computational Science offers a transparent peer review option for original research manuscripts. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please remember to choose, using the manuscript system, whether or not you want to participate in transparent peer review.**
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Computational Science. Please do not hesitate to contact me if you have any questions.

Sincerely,

Jie Pan, Ph.D.
Senior Editor
Nature Computational Science

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #1 (Remarks to the Author):

All reviewers were fundamentally supportive of the work, and the authors have done an exceptionally thorough job of addressing all of the comments. I recommend that the manuscript be accepted in its current form.

Reviewer #2 (Remarks to the Author):

The authors addressed all of my comments properly.

Reviewer #3 (Remarks to the Author):

Wait et al., has done an excellent work in responding thoroughly to all reviewers, including to my comments. I think the corrected manuscript has significantly improved as a result of that and should be accepted to publication with some minor corrections that are noted below:
1. The term GEFI is defined twice.
2. The term AA (amino acid) is not defined in the paper.
3. The terms “accuracy” and precision” that are used to characterize the model (main text and Supp. Fig. 9) are not defined.
4. In the Methods/Animals section, the term ad-lithium should be changed to ad-libitum.
5. In the Methods “fiber photometry recording” and “fiber photometry analysis” sections, it is not clear what is the role of the 410nm line, what is the linear scaling that is mentioned, how it is done, and why. It would be helpful to add a brief description and a reference to a detailed protocol.

Reviewer #4 (Remarks to the Author):

I wanted to express my satisfaction upon reviewing the revised manuscript, which

has been meticulously organized by the authors. They have diligently addressed most concerns raised during the review process. However, I would like to draw your attention to a specific point mentioned in the general comments 2 provided by reviewer #3.

In their review comments, reviewer #3 highlighted the need for additional descriptions regarding the inconsistency of model organisms used in the database used for training (based on HEK293) and the test results (based on cultured neurons) within the main text. While the authors have made clear explanation for the issue raised in the response letter with experimental data, these results do not seem to be incorporated in the main text. To ensure a comprehensive understanding of the study, it would be beneficial for the authors to incorporate this information. Additionally, it is recommended that the relevant data be included in the supplementary figures to provide a more complete picture.

Once these additions have been made, I wholeheartedly support the publication of this manuscript. The authors' commitment to addressing the reviewers' comments and incorporating the necessary revisions demonstrates their dedication to producing a high-quality piece of work.

Author Rebuttal, first revision:

S.Wait, 2024

Secondary Reviews:

Reviewer #1 (Remarks to the Author):

All reviewers were fundamentally supportive of the work, and the authors have done an exceptionally thorough job of addressing all of the comments. I recommend that the manuscript be accepted in its current form.

Reviewer #2 (Remarks to the Author):

The authors addressed all of my comments properly.

Reviewer #3 (Remarks to the Author):

Wait et al., has done an excellent work in responding thoroughly to all reviewers, including to my comments. I think the corrected manuscript has significantly improved as a result of that and should be accepted to publication with some minor corrections that are noted below:

1. The term GEFI is defined twice.
We have removed the second instance, thank you for noticing!

2. The term AA (amino acid) is not defined in the paper.
We have added the definition to *Results: Description of Variant Library, Computational Approach, and Predictions on Novel Sequences*, once again, thank you for noticing!

3. The terms "accuracy" and precision" that are used to characterize the model (main text and Supp. Fig. 9) are not defined.
We have added the equations to the methods sections and referenced the equations in the text as well as in the figure caption.

Results: In Vitro Performance of Ensemble Predictions
The overall accuracy (*Eq.6*) of the $\Delta F/F_0$ model is 0.56 (**Supp. Fig. 9C,E**).

Methods:

To calculate the accuracy of our model, we classified kinetics predictions into variants that are either faster or slower than jGCaMP7s, and ΔF/F predictions into variants containing a larger or smaller ΔF/F than jGCaMP7s. To evaluate our model's performance, we computed an accuracy score *(Eq.6)* using the empirical data, which is the ratio of sum of the true positives (TP) and true negatives (TN) to the total number of predictions.

$$Accuracy\ Score\ = \frac{TP+TN}{n_{predictions}} \qquad (Eq.6)$$

To calculate the precision of our model, we classified kinetics predictions into variants that are either faster or slower than jGCaMP7s, and ΔF/F predictions into variants containing a larger or smaller ΔF/F than jGCaMP7s. To evaluate our model's performance, we calculated the precision *(Eq.7)* of our models using the empirical data, which is the ratio of number of TP over the number of TP and false positive (FP).

$$Precision\ = \frac{TP}{TP+FP} \qquad (Eq.7)$$

Supp Fig 9. Caption:

Supplementary Figure 9: *Estimation of Model Accuracy with Acetylcholine Results*.

- **A.** Confusion Matrix of 100 simulated coin flip experiments (each containing 16 samples).
- **B.** Confusion matrix of all both models' predictions and acetylcholine performance.
- **C.** Confusion matrix of all fluorescence model's predictions and acetylcholine performance.
- **D.** Confusion matrix of all kinetics model's predictions and acetylcholine performance.
- **E.** Quantification of each confusion matrices accuracy (*Eq.6)* and precision *(Eq.7)*.

4. In the Methods/Animals section, the term ad-lithium should be changed to ad-libitum.
We changed the term.

5. In the Methods "fiber photometry recording" and "fiber photometry analysis" sections, it is not clear what is the role of the 410nm line, what is the linear scaling that is mentioned, how it is done, and why. It would be helpful to add a brief description and a reference to a detailed protocol.

We added the following paragraph to the Methods section "Fiber photometry recording"
"Imaging GCaMP with 410 nm wavelength excitation light represents the isosbestic wavelength of the sensor. This means the GFP emission when imaging at this wavelength is not dependent on calcium. Measuring the fluorescence signal using 410 nm wavelength allows us to get a control signal that shows non-Ca2+ related signal changes that could be contributing to the measured Ca2+-dependent signal (470 nm signal). The 410 nm signal was linearly scaled to best fit the 470 nm signal using least-squares regression. The scaled 410 nm was then used as a reference trace to obtain the motion-corrected 470 nm signal by subtracting it from the 470 nm signal. Please see the reference for further details (Kim et al. 2016)."

S.Wait, 2024

Reviewer #4 (Remarks to the Author):

I wanted to express my satisfaction upon reviewing the revised manuscript, which has been meticulously organized by the authors. They have diligently addressed most concerns raised during the review process. However, I would like to draw your attention to a specific point mentioned in the general comments 2 provided by reviewer #3.

In their review comments, reviewer #3 highlighted the need for additional descriptions regarding the inconsistency of model organisms used in the database used for training (based on HEK293) and the test results (based on cultured neurons) within the main text. While the authors have made clear explanation for the issue raised in the response letter with experimental data, these results do not seem to be incorporated in the main text. To ensure a comprehensive understanding of the study, it would be beneficial for the authors to incorporate this information. Additionally, it is recommended that the relevant data be included in the supplementary figures to provide a more complete picture.

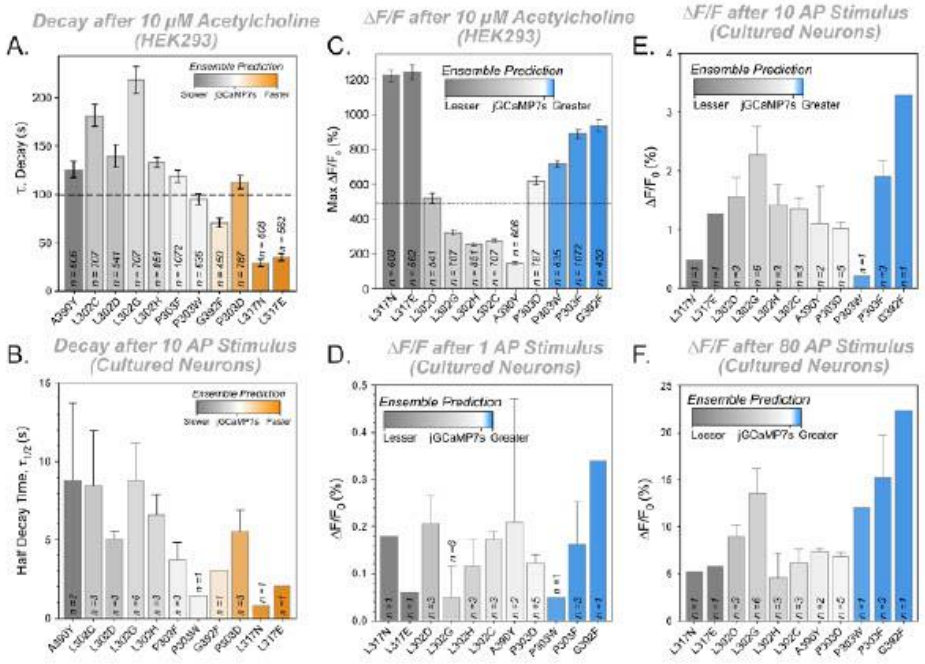
We have included the response to reviewer #3 in the main text as well as an additional supplemental figure! Thank you for your feedback!

We have included the text below in results section: "*In Vitro Performance of Ensemble Predictions*"

"The ideal configuration would be to evaluate them in the same manner as the training data. However, due to the lower throughput of cultured neuron screens, we first performed an intermediate acetylcholine assay step in HEK293 cells. We found the acetylcholine assay approximated variant performances accurately before cultured neuron assays (**Supplementary Figure 4A-F**)."

We have included the following figure as Supplementary Figure 4.

S.Wait, 2024



Once these additions have been made, I wholeheartedly support the publication of this manuscript. The authors' commitment to addressing the reviewers' comments and incorporating the necessary revisions demonstrates their dedication to producing a high-quality piece of work.

Citations:

Bedbrook, Claire N., Kevin K. Yang, J. Elliott Robinson, Elisha D. Mackey, Viviana Gradinaru, and Frances H. Arnold. 2019. "Machine Learning-Guided Channelrhodopsin Engineering Enables Minimally Invasive Optogenetics." *Nature Methods* 16 (11): 1176–84.

Kim, Christina K., Samuel J. Yang, Nandini Pichamoorthy, Noah P. Young, Isaac Kauvar, Joshua H. Jennings, Talia N. Lerner, et al. 2016. "Simultaneous Fast Measurement of Circuit Dynamics at Multiple Sites across the Mammalian Brain." *Nature Methods* 13 (4): 325–28.

Reimers, Stian, Chris Donkin, and Mike E. Le Pelley. 2018. "Perceptions of Randomness in Binary Sequences: Normative, Heuristic, or Both?" *Cognition* 172 (March): 11–25.

Saito, Yutaka, Misaki Oikawa, Hikaru Nakazawa, Teppei Niide, Tomoshi Kameda, Koji Tsuda, and Mitsuo Umetsu. 2018. "Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins." *ACS Synthetic Biology* 7 (9): 2014–22.

Smith, J. M. 1970. "Natural Selection and the Concept of a Protein Space." *Nature* 225 (5232): 563–64.

Wardill, Trevor J., Tsai-Wen Chen, Eric R. Schreiter, Jeremy P. Hasseman, Getahun Tsegaye, Benjamin F. Fosque, Reza Behnam, et al. 2013. "A Neuron-Based Screening Platform for Optimizing Genetically-Encoded Calcium Indicators." *PloS One* 8 (10): e77728.

Yang, Kevin K., Zachary Wu, and Frances H. Arnold. 2019. "Machine-Learning-Guided Directed Evolution for Protein Engineering." *Nature Methods* 16 (8): 687–94.

nature research

Final Decision Letter:

Date: 15th February 24 14:59:20

Last Sent: 15th February 24 14:59:20

Triggered By: Fernando Chirigati

From: fernando.chirigati@us.nature.com

To: berndtuw@uw.edu

CC: jie.pan@us.nature.com

BCC: rjsproduction@springernature.com,computationalscience@nature.com,fernando.chirigati@us.nature.com

Subject: Decision on Nature Computational Science manuscript NATCOMPUTSCI-23-0833C

Message: Dear Dr Berndt,

We are pleased to inform you that your Article "Machine Learning Ensemble Directed Engineering of Genetically Encoded Fluorescent Calcium Indicators" has now been accepted for publication in Nature Computational Science.

Once your manuscript is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

Please note that *Nature Computational Science* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access

requirements, or our legal forms, please contact ASJournals@springernature.com

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see https://www.nature.com/natcomputsci/for-authors). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

Once your manuscript is typeset, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

If you have queries at any point during the production process then please contact the production team at rjsproduction@springernature.com.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: https://www.nature.com/authors/policies/embargo.html

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

We look forward to publishing your paper.

Best,
Fernando (on behalf of Jie Pan)

--
Fernando Chirigati, PhD
Chief Editor, Nature Computational Science
Nature Portfolio

P.S. Click on the following link if you would like to recommend Nature Computational Science to your librarian: https://www.springernature.com/gp/librarians/recommend-to-your-library

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**