nature portfolio

Peer Review File

Deep learning models simultaneously trained on multiple datasets improve base-editing activity prediction

Corresponding Author: Professor Jan Gorodkin

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This study by Sun et al. aims at improving prediction of CRISPR base editing properties by integrating deep learning models trained on combined datasets. By generating over 20,000 guide RNAs (gRNAs) for adenine and cytosine base editors (ABE and CBE), it addresses the limitations of prior datasets. The models accurately predict gRNA efficiency and editing outcomes, outperforming existing tools. Key factors like sequence motifs and editing windows were identified, enhancing prediction reliability. The fused models, accessible via software and web tools, potentially enabling more precise gRNA design.

major

1. The authors use outdated base editors for this study. Nowadays, most people would use ABE8e or ABE8-17/20 and CBE6b (i.e., a TadA-based CBE with mutations N46V Y73P in TadDE).
ABE8e: https://pubmed.ncbi.nlm.nih.gov/32433547/
CBE6b: https://www.nature.com/articles/s41467-024-45969-7

I assume outcomes might be very different for these editors?
It would be important to show if the “old” models predict editing of these new editors as well and/or it would make sense to test the same guides with newer editors to update the tool’s usefulness for current editors.

2. All experiments are based on HEK293T cells. While this is totally fine for an initial run, it is possible that editing properties might differ across different cell types? In BE-HIVE (Arbab et al., Cell 2020), I think there was a difference between HEK and mES outcomes?

minor

3. Interest in BEs has waned a bit, given most people would now aim to use Prime Editing, if possible. Furthermore, many PE prediction ML/DL tools have been published. It would be interesting to discuss at least, if your approach might also be amenable to PE prediction.

4. I would rewrite this sentence: “Cas9 cleaves the double-stranded DNA, which complicates downstream part of the editing process.” To something like this: “ Cas9 DNA cleavage leads to a variety of possible editing outcomes, depending on DNA repair pathways in the cell.”

5. I would rewrite to: These challenges have been met by engineering next-generation CRISPR editors such as base editors (BE).

(Remarks on code availability)
NA

Reviewer #2

(Remarks to the Author)
This study significantly advances base-editing activity prediction by integrating multi-source datasets (SURRO-seq, Song, Arbab) and developing fused deep learning models (CRISPRon-ABE/CBE). Key innovations include:
- Large-scale data generation: ~11,000 novel gRNAs for ABE/CBE, addressing the scarcity of high-quality datasets.
- Multi-dataset fusion training: A novel dataset-aware weighting strategy effectively combines heterogeneous datasets, overcoming the limitations of single-dataset-dependent methods.
- Practical tool development: The online platform and standalone software enable users to adjust dataset weights, enhancing flexibility for real-world applications.
However, there are several issues that need to be addressed to further strengthen the manuscript before it can be considered for publication.

Major Comments:
1.Validation of Feature Engineering Efficacy
The manuscript describes extensive feature engineering involving binding features, dataset-specific features, and indel frequency features. While these efforts are commendable, the empirical evidence supporting their individual contributions to model performance remains insufficient. To substantiate the methodology, we strongly recommend conducting systematic ablation studies that Quantitatively evaluate the predictive power of each feature category (binding/dataset/indel) through sequential removal.
2. Data Partitioning Strategy
While the manuscript emphasizes the importance of leveraging diverse datasets for deep learning model training, the choice of partitioning the data such that sequences within each fold are highly similar and those across folds are highly divergent raises concerns. Traditional cross-validation aims to ensure that each fold is representative of the overall dataset distribution. The proposed method may introduce biases that could:
• Overestimate the model's performance within similar folds (due to redundancy or overfitting).
• Fail to accurately measure the model's generalization capabilities across diverse and unseen data distributions.
Please provide a detailed justification for this partitioning strategy. Specifically:
• What are the theoretical or empirical motivations for prioritizing fold homogeneity over fold representativeness?
• How does this strategy impact the evaluation of the model's robustness in practical scenarios, where the test data may not align with the training data?

Minor comments:
1. In line 263, how many gRNAs supported by over 100,000 total reads were removed? The more reads, the more accurate the efficiency and proportion.
2. The process of creating dataset partitions is an important part of the manuscript. The flowchart can help us to understand the process better.
3. In line 317-319 (step 2), are those sequence highly similar?
4. In line 321, the sequence was randomly distributed into different partitions, which conflicts with the principle predefined in line 310-312.
5. The CIRSPRon model has six validation sets (5 validation + 1 test), all sequence were assigned to the same partition (step 2). But sequence assigned to the test partition of CRISPRon were not allocated to any partitions in line 324-325.
6. In supplementary table 7, each dataset was split into six partitions. Was this done on purpose or is just a coincidence?
7. Are the first four bases, the last three bases and PAM in the 30nt target sequence fixed? If not, the full-length sequence instead of 20bp protospacer should be provided in the Supplementary Data 1.
8. The training and test set data (6 partitions) from SURRO-seq, Song and Arbab in supplementary Table 7 should be provided as supplementary data.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thanks for addressing my concerns. Congrats on this work.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Authors have answered all my questions.

(Remarks on code availability)
Authors have answered all my questions. Thank you.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This study by Sun et al. aims at improving prediction of CRISPR base editing properties by integrating deep learning models trained on combined datasets. By generating over 20,000 guide RNAs (gRNAs) for adenine and cytosine base editors (ABE and CBE), it addresses the limitations of prior datasets. The models accurately predict gRNA efficiency and editing outcomes, outperforming existing tools. Key factors like sequence motifs and editing windows were identified, enhancing prediction reliability. The fused models, accessible via software and web tools, potentially enabling more precise gRNA design.

major

Reviewer 1 comment 1:
1. The authors use outdated base editors for this study. Nowadays, most people would use ABE8e or ABE8-17/20 and CBE6b (i.e., a TadA-based CBE with mutations N46V Y73P in TadDE).
ABE8e: https://pubmed.ncbi.nlm.nih.gov/32433547/
CBE6b: https://www.nature.com/articles/s41467-024-45969-7

I assume outcomes might be very different for these editors?
It would be important to show if the "old" models predict editing of these new editors as well and/or it would make sense to test the same guides with newer editors to update the tool's usefulness for current editors.

Response:
Thanks for the suggestions! Different editors show different outcomes in terms of editing efficiency and bystander editing frequencies. Taken on this great suggestion, now we collect the latest publicly available ABE8e data and one additional ABEmax data (Kissling et al, 2025) and combine the datasets to train our model. Limited by current publicly available datasets, we did not obtain enough data points with outcome frequency from other base editors (including CBE6b) to train or evaluate the performance. Details on the performance is now included in the revised Fig. 3. Trained with dataset from the new ABE8e, our models trained with the same deep-learning network exhibit the best performance compared to other state-of-the-art BE prediction tools. Thus, the base editing deep learning model developed by our study provides an advanced training framework for the development of other base editors in the future.

Reviewer 1 comment 2:
2. All experiments are based on HEK293T cells. While this is totally fine for an initial run, it is possible that editing properties might differ across different cell types? In BE-HIVE (Arbab et al., Cell 2020), I think there was a difference between HEK and mES outcomes?

Response:
Thanks for the suggestion. We agree that base editing properties can vary across different cell lines. We have now evaluated the performance on Arbab data across other cell lines. See the detailed performance evaluation in Supplementary Fig. 12. Our results show that the model trained with data from HEK293T cells can satisfactorily predict the editing properties in other cell types.

Compared with the difference across cell lines, the diverse properties across different platforms are more critical in current application. Thus, in this study we only focus on the dataset generated in HEK293T cells since the observation of diverse properties of base editor datasets from different platforms (or different studies). And we agree that the difference across cell lines should definitely be considered in the further improvements.

minor

Reviewer 1 comment 3:
3. Interest in BEs has waned a bit, given most people would now aim to use Prime Editing, if possible. Furthermore, many PE prediction ML/DL tools have been published. It would be interesting to discuss at least, if your approach might also be amenable to PE prediction.

Response:
Thanks for pointing it out. We agree that there is a strong interests and need in the development of deep-learning models for PE design. Compared to BE, the generation of large-scale and systemic in-cell measured PE efficiency data will be needed to facilitate the development of prediction models. We have now added this in the discussion. Based on this suggestion, the following texts are added in the manuscript accordingly.

(Page 6) "In addition, our dataset integration strategy not only proves effective for combining datasets from multiple base editors but also holds potential for broader applications in other CRISPR technologies. For instance, integrating cell-line information when training models on prime editing datasets with same edit types such as short indels but measured in different cell lines may enhance predictive accuracy."

Reviewer 1 comment 4:
4. I would rewrite this sentence: "Cas9 cleaves the double-stranded DNA, which complicates downstream part of the editing process." To something like this: " Cas9 DNA cleavage leads to a variety of possible editing outcomes, depending on DNA repair pathways in the cell."

Response:
Thanks for this suggestion. We rewrite the sentences in the manuscript.

(Page 2). "Classical CRISPR-Cas9-based gene editing relies on the introduction of double-stranded DNA breaks (DSB) guided by the small guide RNA (gRNA). Repair of the CRISPR-induced DSBs in cells leads to the introduction of a variety of possible insertions or deletions (known as indels), depending on DNA repair pathways."

5. I would rewrite to: These challenges have been met by engineering next-generation CRISPR editors such as base editors (BE).

Response:
Thanks for this suggestion. We have rewritten the sentences in the manuscript.

(Page 1). "These challenges have been met by engineering next-generation CRISPR editors such as base editors (BE) and prime editors (PE)."

Reviewer #1 (Remarks on code availability):

NA

Reviewer #2 (Remarks to the Author):

This study significantly advances base-editing activity prediction by integrating multi-source datasets (SURRO-seq, Song, Arbab) and developing fused deep learning models (CRISPRon-ABE/CBE). Key innovations include:
- Large-scale data generation: ~11,000 novel gRNAs for ABE/CBE, addressing the scarcity of high-quality datasets.
- Multi-dataset fusion training: A novel dataset-aware weighting strategy effectively combines heterogeneous datasets, overcoming the limitations of single-dataset-dependent methods.
- Practical tool development: The online platform and standalone software enable users to adjust dataset weights, enhancing flexibility for real-world applications.
However, there are several issues that need to be addressed to further strengthen the manuscript before it can be considered for publication.

Major Comments:

Reviewer 2 comment 1:
1.Validation of Feature Engineering Efficacy
The manuscript describes extensive feature engineering involving binding features, dataset-specific features, and indel frequency features. While these efforts are commendable, the empirical evidence supporting their individual contributions to model performance remains insufficient. To substantiate the methodology, we strongly recommend conducting systematic ablation studies that Quantitatively evaluate the predictive power of each feature category (binding/dataset/indel) through sequential removal.

Response:
Thanks for suggestion. We added the ablation analysis that evaluates the impact of removing binding energy ΔG_B, dataset labelling and predicted Cas9 indel frequency. The model performances after removing ΔG_B and the predicted Cas9 indel frequency are consistent with the feature importance from SHAP analysis. After removing the dataset labelling led to a substantial decrease in model performance compared to when using the different datasets in the training process. See the detailed description of the performance in Supplementary Fig. 4.

Reviewer 2 comment 2:
2. Data Partitioning Strategy
While the manuscript emphasizes the importance of leveraging diverse datasets for deep learning model training, the choice of partitioning the data such that sequences within each fold are highly similar and those across folds are highly divergent raises concerns. Traditional cross-validation aims to ensure that each fold is representative of the overall dataset distribution. The proposed method may introduce biases that could:
• Overestimate the model's performance within similar folds (due to redundancy or overfitting).

Response:
Thanks for pointing it out. Clustering data points into folds lead to folds that are minimized to overlap, e.g. the distance between the data points inside a folder is smaller that then distance to data points in any other fold. We consider this a conservative approach and use it in spite that potentially can lead to underestimation of the reported performance. When we evaluate,

the evaluation will not be fair if there are data point in the test set that are similar data points used in the training. This is what we aim to avoid.

Considering that a random partition would be acceptable if all data points were dissimilar, we for completeness use this as "baseline" and compared to a random partition of the training, validation and test sets.

We used our SURRO-seq dataset for training, and we included target and outcome sequences and gRNA-target-DNA binding energy ΔG_B without CRISPRon prediction in case it introduces data leakage during the nested cross-validation process. Then we trained several models in two scenarios: (1) based on our current 6-partition independent SURRO-seq dataset by 6-fold nested cross-validation with 10 random initiations; (2) A random 6-fold partition trained the same training strategy with the same hyper-parameter combinations. We applied two random splits using two random seeds to ensure the robustness of the result.

Then we evaluated the performance on the test sets by averaging the results from the best 5 models selected based on their validation performance. Through comparing the performance from the models trained on different dataset splitting strategies, we found using our current data partition does not lead to overestimation.

Dataset splitting	Independent dataset	Random split 1	Random split 2
MSE from test set	118.76	119.08	119.75

Hence, we do not see any sign of overfitting.

Reviewer 2 comment 3:
• Fail to accurately measure the model's generalization capabilities across diverse and unseen data distributions.
Please provide a detailed justification for this partitioning strategy. Specifically:
• What are the theoretical or empirical motivations for prioritizing fold homogeneity over fold representativeness?

Response:
As touch on above we want to avoid data points in the independent test set that are similar to the ones used for training the model as such data leakage will lead to an inflated performance evaluation. We do not really see how a potential lack of fold representatives can lead to overfitting, but acknowledge that it can as mentioned above lead to reporting a lower performance. The more dissimilar the training and testing data are the lower the performance will be, which is actually also what we see when comparing between the more dissimilar test sets from different sources.

It should also be noted that the data points which are most dissimilar to all other data points are distributed randomly in all partition. These data points make up around 99.2% for both ABE and CBE respectively, showing that a substantial fraction of the data is representative as well. See also the methods subsection "Generation of dataset partitions", which includes "4) Remaining similar targets were grouped and assigned to one partition randomly".

In the study by Roberts et al., (https://nsojournals.onlinelibrary.wiley.com/doi/10.1111/ecog.02881) did a simulation to highlight the importance of considering data structures during cross-validation. The random

split can be overfit while block the combinations of dataset may lead to overestimating interpolation errors.

We hope that our intend of avoiding data leakage, while still having good representativeness have been clearly explained.

Reviewer 2 comment 4:
• How does this strategy impact the evaluation of the model's robustness in practical scenarios, where the test data may not align with the training data?

Response:
If the test set is not aligned well with the training set, the performance of the model will drop as shown when evaluating on test sets more distant from the training data (see Fig 3).

Minor comments:

Reviewer 2 comment 5:

1. In line 263, how many gRNAs supported by over 100,000 total reads were removed? The more reads, the more accurate the efficiency and proportion.

Response:
Thanks for pointing it out. 2 gRNAs (TGATGTAGTCCCGAGGTTTC, TGCAATCACCTTGGGTCCAA) from ABE and 3 gRNAs (CAGGCTGGTGGCGATGTTCT, TGCAATCACCTTGGGTCCAA, TGATGTAGTCCCGAGGTTTC) are removed in Song dataset due to the high total reads, whose total reads are 1,156,803, 125,596, 161,517, 135,586 and 134,525, respectively. These 5 gRNAs are removed since they have a much higher total reads than other gRNAs (median ABE: 1,421, CBE: 2,036).

Reviewer 2 comment 6:
2. The process of creating dataset partitions is an important part of the manuscript. The flowchart can help us to understand the process better.

Response:
Thanks for the suggestion. We add flowchart in Supplementary Fig. 14. The core idea of the dataset splitting is to ensure no highly similar data points between training and validation or training and test set to prevent overestimation.

Reviewer 2 comment 7:
3. In line 317-319 (step 2), are those sequence highly similar?

Response:
Yes.

Reviewer 2 comment 8:
4. In line 321, the sequence was randomly distributed into different partitions, which conflicts with the principle predefined in line 310-312.

Response:
Thanks for pointing it out. Our idea for dataset partition is putting the highly similar data points into one partition. The remaining dissimilar ones are randomly put into different partitions. See details in the flowchart.

Reviewer 2 comment 9:
5. The CIRSPRon model has six validation sets (5 validation + 1 test), all sequence were assigned to the same partition (step 2). But sequence assigned to the test partition of CRISPRon were not allocated to any partitions in line 324-325.

Response:
For those gRNAs in Song train set which are assigned to the test partition in CRISPRon are randomly put into 5 validation partitions. See details in the flowchart.

Reviewer 2 comment 10:
6. In supplementary table 7, each dataset was split into six partitions. Was this done on purpose or is just a coincidence?

Response:
We aim for training our model using 5-fold cross-validation on a combined dataset from different sources. Each dataset was split into six partitions for two main reasons. 1) we started with training models on individual dataset, which were trained using same strategy for a fair comparison with 5-fold cross-validation. 2). Considering these datasets are different, we keep the number of gRNAs balanced among different partitions to avoid bias between different partitions during cross-validation.

Reviewer 2 comment 11:
7. Are the first four bases, the last three bases and PAM in the 30nt target sequence fixed? If not, the full-length sequence instead of 20bp protospacer should be provided in the Supplementary Data 1.

Response:
Thanks for the suggestion. These sequences are not fixed. The full-length sequences are provided in Supplementary Data 1.

Reviewer 2 comment 12:
8. The training and test set data (6 partitions) from SURRO-seq, Song and Arbab in supplementary Table 7 should be provided as supplementary data.

Response:
Thanks for the suggestion. These datasets are made available along with the software.