nature portfolio

Peer Review File

Convolutional neural networks quantify antibiotic resistance in Mycobacterium tuberculosis with diagnostic grade accuracy and predict treatment response

Corresponding Author: Ms Sanjana Kulkarni

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)
Overview
The authors present a convolutional neural network model capable of predicting Minimum Inhibitory Concentration (MIC) values from Mycobacterium tuberculosis gene sequences, evolutionary information, biochemical properties of proteins and rare variants. They demonstrate the utility of their new model on rifampicin-susceptible samples where patients had unfavourable outcomes.
The authors do an excellent job of describing their work, especially in the methods section and their code is well documented. As constructive feedback, I would aim to highlight the novelty of this approach more and its relevance to clinical predictions in light of previous MIC prediction papers for TB. The manuscript feels slightly unfinished in places e.g. (ref) still within the manuscript so could do with an extra proofread. I attach my comments below, but overall it is a well-executed piece of research.

Comments
Abstract
1. I think the abstract is well-written. However, I would rephrase 'These results demonstrate the value of encoding multiple dimensions of biological data in machine learning of function or cellular phenotypes' to just 'drug-resistant phenotypes' as this manuscript mainly focuses on the prediction of MIC as the outcome (of course there are wider clinical implications and phenotypes of interest associated with this).
Introduction
1. Minor comment, perhaps double check for double spaces throughout the manuscript e.g. line 34: 'The prediction of antibiotic resistance'
2. Drug-resistance across pathogens could potentially be underpinned by things outside of the genome. I would perhaps tone down line 34 to reflect this. Pathogens can also refer to species beyond bacteria and so you could make it more specific to Mycobacterium where there is often one highly penetrating mutation resulting in resistance.
3. To none-machine learning audiences, it may be useful to provide more context to why there is a need to synthesize genetic data e.g. due to a lack of bedaquiline resistant samples being available to train models (hopefully I am correct in your line of thinking). Potentially I would add a sentence to say why ML models are useful in clinical settings here (line 41).
4. I believe there is another MIC prediction paper for TB available: Quantitative measurement of antibiotic resistance in Mycobacterium tuberculosis reveals genetic determinants of resistance and susceptibility in a target gene approach | Nature Communications. It might be worth mentioning this in the introduction to position your model in comparison and highlight its novelty.
Results
1. On line 73, it mentions that there are unpublished sources, have they since been published?
2. On line 79 it would be nice for you to summarise the lineage distribution beyond referencing figure 1. Whilst there are probably many, are there any dominant sublineages in the dataset?
3. In section B the title is 'CNN architecture with multimodal input improves prediction accuracy', but no predictive accuracy is actually referred to. I would sprinkle in some numbers into the paragraph and comment on the actual performance of the models, as it mostly contains methodological insights so far.
4. 'Lineage defines hundreds to thousands of linked variants across the genome that can encode MIC effects or interact with

canonical resistance mutations.' I would rephrase this sentence to make it clearer that there is an association between genomic background, often classified into lineages, and the DR phenotype.
5. On line 115-116 - 'The learned relationships between lineages and MICs are supported by the average MICs by lineage'- can you describe this relationship further numerically- do you see increased performance for certain lineages?
6. Online line 125 you mention 'EA ranged from 79-93% (average 89%)' – can you break this down per drug e.g. higher performance for RIF and INH perhaps because they are first-line drugs and there are more data? Do you see such trends?
7. On line 145, 'The essential agreement rates of the CNNs are comparable to those reported by the CRyPTIC using XGBoost models trained on whole-genome k-mers.'- Can you add a reference here?
8. Line 151, ' To validate the variants the CNNs learned from'- here I think you mean you want to identify the variants that have higher importance in the model, perhaps rephrase the topic sentence.
9. Line 195- Change to 'RpoB (Fig. 4a), and GyrA and GyrB (Fig. 4b).'
10. Line 197- 'hot spots'. Super minor, but I believe this is one word.
11. 'These include PncA and KatG, which activate the pro-drugs pyrazinamide and isoniazid, respectively, and Rv0678, a transcription factor that represses expression of bedaquiline efflux pumps mmpL5 and mmpS5 (ref).' - I think you mean to add a reference here.
12. I am not so familiar with 'cold spots' could you perhaps define this. It is a nice, novel approach to looking at DR and protein sequences and it would be more impactful if these terms are defined.
13. 'Mutations in the Rv0678 hot spot likely inhibit or reduce DNA binding upstream of the bedaquiline efflux pumps'- If using the term 'likely' can you add a reference to support this, alternatively tone down the language if this is more hypothetical.
14. I am a bit confused here: 'We overlaid CNN predictions on the AlphaFold confident substructure of EthA (residues 2-483 of 489 total residues with high structure confidence scores, pLDDT > 70).'. What did you overlay exactly? Was it the feature importance scores or the important mutations you observed?
15. Results sections J and K are well-written and I enjoyed reading these results.

Discussion

1. Reference WHO perhaps: 2023 WHO catalog classification method, despite being trained on a fraction of the data.
2. Line 327: low CD4+ T cell
3. In the GitHub page it mentions that pkl models etc are excluded due to size- this leads to a larger overall question: 'Can we actively implement ML models in clinical practice? Or do computational requirements etc limit this?'

Methods

1. To Supplementary Table 1 can you add the drugs to their associated drug-resistant candidate genes.
2. Line 417: ' We additionally removed isolates whose MIC range spans the drug's critical concentration from model performance comparisons due to the inability to know if these isolates are drug-resistant or drug-susceptible for binary classification analysis'- I think these are potentially some of the interesting results, can you comment in the results about their distribution. What do you think this means clinically (perhaps add to the Discussion)- is using a cut-off a good idea?
3. 'Isolates were split into training (80%)' – did you use a validation set during model training? What proportion of the dataset?
4. For part B, referring to the augmentation of the bedaquiline data, did you vary the MIC to assess its impact on predictive performance? Perhaps mention this as a limitation if not already discussed.
5. 'Additional variant filtering and annotation was done using bcftools (v1.9)'- perhaps add some detail on what was filtered e.g. did you filter for missingness/ heterozygous SNPs? Or reference it to where it is discussed later as per line 468 onwards.

Code Availability

1. I have not tested the code independently. However, it is well-documented.

Figures

1. Perhaps I am mistaken in my interpretation, but I find Figure 2 slightly confusing. Why do the green and orange points not correspond/ line up to the categories on the x axis?
2. I really like Figure 4 and appreciate the difficulty to visualise large amounts of data- could any of the panels be moved to the supplementary/ made larger in any way? I struggle to see the heatmaps, breaking them up may help this.
3. Same concept for Figure 5- lovely visualisations but I struggle to take the overall message due to the amount of data presented/ small font size in places.

General Comment

The manuscript is well-written, but I would check for minor grammatical errors throughout. I appreciate it has probably gone through a few rounds of edits, but it could do with a minor proofread e.g. adding a ',' after however, double spacing etc.

(Remarks on code availability)

I have not tested the code independently. However, it is well-documented. If possible share the pkl file via zenodo or some other means as it would be great for others to be able to use this model.

Reviewer #2

(Remarks to the Author)

Here, the authors present convolutional neural network (CNN) models that integrate genomics, amino acid biochemical properties, and lineage information to predict quantitative minimum inhibitory concentrations (MICs) for eight anti-tubercular drugs in Mycobacterium tuberculosis with high accuracy. The models outperform simple regression approaches, recapitulate known resistance mechanisms and drug-binding regions, and identify both borderline resistance effects and novel candidate resistance associated variants. The manuscript is very well written, and the results novel with direct clinical impact. I have only two minor comments that I feel would make the paper stronger.

Minor comments:

1. The paper compares the CNNs to simple linear regression, however comparing the CNNs directly to other, stronger methods, such as gradient boosting, k-mer-based models, or other nonlinear sequence models, using the same training and test splits would make it clearer if CNNs are the optimal model for this task. This limitation could be acknowledged in the discussion, noting that additional comparisons may clarify the advantages of the proposed model.

2. The link between higher predicted rifampicin MICs below the resistance threshold and worse treatment outcomes could be more interpreted carefully. The predicted MICs were split at the median likely because of the small numerical range of MIC values in this cohort. Again, stating this limitation in the discussion and clarifying whether data constraints did not allow for the MIC to be modelled as a continuous variable would strengthen the interpretation of this result.

(Remarks on code availability)
Code is clear and readable.

Reviewer 1 Comments:

Abstract:

1. I think the abstract is well-written. However, I would rephrase 'These results demonstrate the value of encoding multiple dimensions of biological data in machine learning of function or cellular phenotypes' to just 'drug-resistant phenotypes' as this manuscript mainly focuses on the prediction of MIC as the outcome (of course there are wider clinical implications and phenotypes of interest associated with this).

   Fixed this.

Introduction:

1. Minor comment, perhaps double check for double spaces throughout the manuscript e.g. line 34: 'The prediction of antibiotic resistance'

   Fixed this here and everywhere else.

2. Drug-resistance across pathogens could potentially be underpinned by things outside of the genome. I would perhaps tone down line 34 to reflect this. Pathogens can also refer to species beyond bacteria and so you could make it more specific to Mycobacterium where there is often one highly penetrating mutation resulting in resistance.

   Fixed this.

3. To none-machine learning audiences, it may be useful to provide more context to why there is a need to synthesize genetic data e.g. due to a lack of bedaquiline resistant samples being available to train models (hopefully I am correct in your line of thinking). Potentially I would add a sentence to say why ML models are useful in clinical settings here (line 41).

   We added some additional sentences here (lines 53-65 in the revised manuscript).

4. I believe there is another MIC prediction paper for TB available: Quantitative measurement of antibiotic resistance in Mycobacterium tuberculosis reveals genetic determinants of resistance and susceptibility in a target gene approach | Nature Communications. It might be worth mentioning this in the introduction to position your model in comparison and highlight its novelty.

   The prediction paper is actually this paper by the same group; the above paper is an association study. We've included a reference to their work in the introduction now at line 63.

Results:

1. On line 73, it mentions that there are unpublished sources, have they since been published?

   No, and we wish they were, but it's unfortunately not up to us. Some of the whole-genome sequencing data has been deposited in SRA and ENA (and we've included the

public IDs in Supplementary Data 1), but the MICs have not been made public.

2. On line 79 it would be nice for you to summarise the lineage distribution beyond referencing figure 1. Whilst there are probably many, are there any dominant sublineages in the dataset?

   We've added this now to lines 89-91 in the revised manuscript.

3. In section B the title is 'CNN architecture with multimodal input improves prediction accuracy', but no predictive accuracy is actually referred to. I would sprinkle in some numbers into the paragraph and comment on the actual performance of the models, as it mostly contains methodological insights so far.

   We've renamed this section to "CNN with multimodal input more accurately models genotype to phenotype than nucleotide-only input" because we wanted to focus on the methodological improvements.

   But we have added a line at the end about a comparison of the model performances for pyrazinamide with and without amino acid properties (lines 111-112). We did not do a comparison of models with and without amino acid properties for all drugs.

4. 'Lineage defines hundreds to thousands of linked variants across the genome that can encode MIC effects or interact with canonical resistance mutations.' I would rephrase this sentence to make it clearer that there is an association between genomic background, often classified into lineages, and the DR phenotype.

   We've rephrased this part of the text and added some additional information (lines 115-121).

5. On line 115-116 - 'The learned relationships between lineages and MICs are supported by the average MICs by lineage'- can you describe this relationship further numerically- do you see increased performance for certain lineages?

   Yes, we computed the mean absolute error (MAE) for each of the four major lineages in each drug dataset. We have added this:

   "For the two fluoroquinolones and pyrazinamide, the MAEs for lineage 2 isolates are significantly larger than the overall MAEs ($p < 0.03$, two-sided Mann-Whitney U test). For both rifampicin and ethambutol, MAEs for lineage 2 isolates are significantly lower than overall ($p < 0.003$), and MAEs for lineage 4 isolates are significantly larger ($p < 0.02$). Bedaquiline MAE for lineage 3 isolates is also significantly larger than overall ($p = 0.0003$). All other comparisons were not significant."

   to lines 137-142. This doesn't seem to suggest any broad trends in differences in performance by lineage, however, just drug-specific observations.

6. Online line 125 you mention 'EA ranged from 79-93% (average 89%)' – can you break this down per drug e.g. higher performance for RIF and INH perhaps because they are first-line drugs and there are more data? Do you see such trends?

   Not exactly, because essential agreements are inflated for drugs with small MIC ranges

because one doubling in either direction encompasses more of the range than for drugs with larger MIC ranges. Rifampicin and isoniazid have the largest MIC ranges in our dataset (15 and 13 doublings, respectively) because they are first-line drugs, but not the highest EAs. The essential agreement rates we computed, in decreasing order, are levofloxacin (93%), bedaquiline (92%), ethambutol (91%), rifampicin (90%), ethionamide (89.1%), isoniazid (88.8%), moxifloxacin (86%), and pyrazinamide (79%).

This is similarly observed in the CRyPTIC paper, where the EAs of the 8 drugs we built models for, in decreasing order, are ethambutol (95.0%), levofloxacin (94.9%), isoniazid (94.2%), bedaquiline (93.3%), rifampicin (92.9%), ethionamide (91.3%), and moxifloxacin (90.9%). Pyrazinamide is not in the CRyPTIC dataset.

7. On line 145, 'The essential agreement rates of the CNNs are comparable to those reported by the CRyPTIC using XGBoost models trained on whole-genome k-mers.'- Can you add a reference here?

   Added this.

8. Line 151, ' To validate the variants the CNNs learned from'- here I think you mean you want to identify the variants that have higher importance in the model, perhaps rephrase the topic sentence.

   Rephrased this (line 180).

9. Line 195- Change to 'RpoB (Fig. 4a), and GyrA and GyrB (Fig. 4b).'

   Fixed (line 228).

10. Line 197- 'hot spots'. Super minor, but I believe this is one word.

    Fixed this here and everywhere else in the manuscript.

11. 'These include PncA and KatG, which activate the pro-drugs pyrazinamide and isoniazid, respectively, and Rv0678, a transcription factor that represses expression of bedaquiline efflux pumps mmpL5 and mmpS5 (ref).' - I think you mean to add a reference here.

    Added it.

12. I am not so familiar with 'cold spots' could you perhaps define this. It is a nice, novel approach to looking at DR and protein sequences and it would be more impactful if these terms are defined.

    The explanation had been moved to the methods section. We moved some of it here to make it easier to understand coldspots when reading the results (lines 224-226).

13. 'Mutations in the Rv0678 hot spot likely inhibit or reduce DNA binding upstream of the bedaquiline efflux pumps'- If using the term 'likely' can you add a reference to support this, alternatively tone down the language if this is more hypothetical.

    This is more hypothetical, so we have replaced 'likely' to 'may.' We also fixed the language about 4 monomers; there are only 2, but there are 4 in the crystal structure

(PDB ID 4nb5), so we have updated the clustering results to reflect using a distance map with 2 monomers, rather than 4 (lines 243-248).

14. I am a bit confused here: 'We overlaid CNN predictions on the AlphaFold confident substructure of EthA (residues 2- 483 of 489 total residues with high structure confidence scores, pLDDT > 70).'. What did you overlay exactly? Was it the feature importance scores or the important mutations you observed?

    We apologize, this was incorrect and arose during editing rounds. The figure is the same as all the other structures. We colored the five hotspot residues red. We think it would be better for consistency with the other structures to NOT color the three residues with the lowest MIC predictions blue. These aren't true coldspots, and we had thought it might make it easier to highlight them, but it probably causes more confusion than it's worth. The residues with low predicted MICs can be seen in Figure 5e.

Discussion:

1. Reference WHO perhaps: 2023 WHO catalog classification method, despite being trained on a fraction of the data.

   Added this reference.

2. Line 327: low CD4+ T cell

   Fixed this here and everywhere else, except for when it says HIV / CD4 status.

3. In the GitHub page it mentions that pkl models etc are excluded due to size- this leads to a larger overall question: 'Can we actively implement ML models in clinical practice? Or do computational requirements etc limit this?'

   Not necessarily. The files being referred to contain the genetic data for all regions of interest (Tier 1 + Tier 2 loci according to the 2023 WHO mutation catalog) for all isolates per drug, which is on the order of 10,000 isolates.

   The same genetic data for a single isolate is approximately 1 megabyte, which is not prohibitively large. Predicting MICs for new isolates using a pre-trained model is very fast (less than 1 second), so one could individually predict an MIC for every new isolate, or batch them (not strictly needed since the time is so low) into small enough batches to avoid exceeding standard RAM.

   We have now added an additional directory to the GitHub repository called "test_example" that includes all the input data files for 10 samples and a script to get MIC predictions for these 10 samples. This data is not too large to deposit here.

Methods:

1. To Supplementary Table 1 can you add the drugs to their associated drug-resistant candidate genes.

   Added them.

2. Line 417: ‘ We additionally removed isolates whose MIC range spans the drug’s critical concentration from model performance comparisons due to the inability to know if these isolates are drug-resistant or drug-susceptible for binary classification analysis’- I think these are potentially some of the interesting results, can you comment in the results about their distribution. What do you think this means clinically (perhaps add to the Discussion)- is using a cut-off a good idea?

   203 isolates that span the CC were excluded for 4 drugs: bedaquiline, ethambutol, ethionamide, and isoniazid. 60% of the excluded cases are for ethambutol. Many *embB* mutations (the main source of resistance mutations) confer MICs close to the CC, according to the 2023 WHO *Mtb* resistance mutation catalogue, which may explain the large number for ethambutol. Overall, 74% of these excluded isolates have CNN-predicted MICs within 1 doubling of the critical concentration. We have added this information to **Supplementary Results Section G** and **Supplementary Figure 11.**

   We’re cautious about interpreting this biologically. These isolates are essentially caused by experimental conditions; different labs use different testing concentrations, and the breakpoints for some drugs have been revised since some of these data were collected.

   While it’s true that isolates with intermediate resistance are phenotypically different from those that are susceptible or have high levels of resistance, we don’t think we can comment much on them given that most of what we see is technical artifacts caused by needing to measure MICs semi-continuously. Using a cut-off is not ideal but is necessary for clinical tractability and decision-making.

3. ‘Isolates were split into training (80%)’ – did you use a validation set during model training? What proportion of the dataset?

   Yes, we have rephrased that text now to make that clear. Lines 462-465: “Isolates were split into training (80%), validation (10%), and testing (10%) sets, stratified by binary resistance phenotype and primary lineage. The validation set was used to determine when to stop training the CNNs and the regularization parameter for regression models. The final model performance metrics were computed on the test set.”

4. For part B, referring to the augmentation of the bedaquiline data, did you vary the MIC to assess its impact on predictive performance? Perhaps mention this as a limitation if not already discussed.

   No, we did not test this. We added this limitation in **Supplementary Results Section B.**

5. ‘Additional variant filtering and annotation was done using bcftools (v1.9)’- perhaps add some detail on what was filtered e.g. did you filter for missingness/ heterozygous SNPs? Or reference it to where it is discussed later as per line 468 onwards.

   It’s probably best to remove the bcftools text. We used bcftools to do position-level filtering to get variants within loci coordinates, but filtering sites for missingness was done in the custom python script described in the next section, “Featurizing nucleotide sequences.”

Figures:

1. Perhaps I am mistaken in my interpretation, but I find Figure 2 slightly confusing. Why do the green and orange points not correspond/ line up to the categories on the x axis?

   The green and orange points don't line up to the x-axis ticks perfectly because we dodged them so that the green and orange points wouldn't be on top of each other.

   We've treated the x-axis like a categorical variable (even though it's a continuous integer position) to avoid having a lot of empty space for all the other codons at which nonsense mutations are not observed in the dataset. We're open to suggestions to make the figure clearer, maybe by increasing the spacing between codon positions or adding dashed black vertical lines to separate the sites from each other so it's easier to see which position the points correspond with?

2. I really like Figure 4 and appreciate the difficulty to visualise large amounts of data- could any of the panels be moved to the supplementary/ made larger in any way? I struggle to see the heatmaps, breaking them up may help this.

   We broke them a bit by separating the first two panels (RpoB and GyrA/B) to Figure 4 and the rest to Figure 5, with the split being essential genes (Figure 4) and non-essential genes (Figure 5). Let us know if this helps or if the panels should be further split.

3. Same concept for Figure 5- lovely visualisations but I struggle to take the overall message due to the amount of data presented/ small font size in places.

   We apologize, this figure's legend was not properly updated. We have fixed the legend and condensed some of the panels to make it easier to read. Figure 6 (newly named because we split Figure 4 into two) is now only about the composite outcomes models, and Supplementary Figure 8 is now only about the TCC models.

Reviewer 2 Comments:

1. The paper compares the CNNs to simple linear regression, however comparing the CNNs directly to other, stronger methods, such as gradient boosting, k-mer-based models, or other nonlinear sequence models, using the same training and test splits would make it clearer if CNNs are the optimal model for this task. This limitation could be acknowledged in the discussion, noting that additional comparisons may clarify the advantages of the proposed model.

   This is a fair point, as linear regression is the simplest alternative model and is limited by its assumption of a linear function between the input and output. To address this, we built an XGBoost model to predict rifampicin MIC (with the lineage SNPs and amino acid properties) using the same inputs for the CNN and ridge regression models.

   The XGBoost model was trained with a maximum allowed tree depth of 10, a learning rate of 0.1, and the same custom loss function and stopping criterion as for the CNN. We also applied L2 regularization and used 5-fold cross-validation to select the best regularization strength, as we did for regression. These are the results (all tests are one-sided Mann-Whitney U tests):

   The CNN model has significantly better mean absolute error (MAE) and essential

agreement than XGBoost (p = 0.004 and 0.006, respectively). Regression has better MAE than XGBoost, (0.608 vs. 0.613, p = 0.048), and XGBoost has significantly better essential agreement than regression (81.7% vs. 81.0%, p = 0.014).

XGBoost models require tabular input, like regression models, so the advantage of preserving nucleotide order and sequence context is still with the CNNs.

We have now added this information to lines 381-387 in the Discussion: “Other non-linear models, such as XGBoost models, may provide better accuracy but remain limited by their reliance on tabular representations of independent features. For comparison, we trained an XGBoost model to predict rifampicin MIC. Although it achieved slightly higher essential agreement than regression, it had slightly worse error and was outperformed by the CNN on both metrics, highlighting the importance of capturing local sequence context and positional dependencies among variants.”

2. The link between higher predicted rifampicin MICs below the resistance threshold and worse treatment outcomes could be more interpreted carefully. The predicted MICs were split at the median likely because of the small numerical range of MIC values in this cohort. Again, stating this limitation in the discussion and clarifying whether data constraints did not allow for the MIC to be modelled as a continuous variable would strengthen the interpretation of this result.

   We agree on this point. We have added additional cautionary language to **Results Section K** (lines 334-335) and the Discussion (lines 351-358).