

Artificial Intelligence coupled to pharmacometric modelling to tailor malaria and tuberculosis treatment in Africa

Corresponding Author: Dr Miquel Duran-Frigola

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)

Thank you for the opportunity to review this manuscript, which investigates pharmacogenetics in an understudied population. The authors apply machine learning (ML) and artificial intelligence (AI) to identify drug-gene pairs relevant to key therapeutic areas, specifically malaria and tuberculosis, in African populations. These identified genes are then incorporated into pharmacometric modelling to assess their effects, including implications for dosing. I appreciate the significant effort and time invested in this research and in preparing the manuscript for publication.

The authors trained a supervised learning model for drug-pharmacogene prioritization using drug and gene descriptors, applied an LLM-based pharmacogene re-ranking approach, and performed PBPK and NLME modelling. They state that “*we propose a pipeline that couples a novel ML/AI component...*”; however, it is unclear which aspects are truly novel, given that LLMs have already been used for feature selection (e.g., <https://arxiv.org/abs/2407.02694>). Could the authors clarify what is specifically innovative in their approach? I outline additional concerns below.

1. Use of PharmGKB as ‘ground truth’

Less than 4% of PharmGKB data comes from African populations [PMID: 37450899], as the authors also noted in a separate study [PMID: 38594395]. This underrepresentation increases the likelihood that African-specific variants are missing, leading to potential misclassification in downstream classification; particularly, some instances labelled as “no drug-gene interaction” may be incorrect. Moreover, analyses focusing on African-specific variants could amplify this bias, as they rely on a subset of the data where misclassification is more likely. The authors later identify significant predictions or drug-gene pairs that are absent from PharmGKB, further suggesting that PharmGKB may be missing important drug-gene interactions. If these novel interactions are indeed valid, this heightens the concern that the training data may not be optimal. The authors should acknowledge this limitation and, if possible, estimate its impact on the downstream results.

It is commendable that the authors excluded level 4 evidence (*‘evidence does not support an association between the variant and the drug phenotype’*) from downstream analysis. However, level 3 evidence (which is based on a single study, preliminary findings, or inconsistent replication across studies) is also unreliable and should ideally be excluded. That said, removing true drug-gene associations may introduce misclassification, just as including false associations would. Can the authors perform a sensitivity analysis to estimate the impact of including low-quality (level 3 and probably level 6) data on the results?

2. NLME and PBPK modelling

In Figure 4C-D, the authors highlight the interactions between artemether and *CYP2B6**6, as well as rifampicin and *SLCO1B1* rs4149032, both of which are well-established drug-gene interactions (e.g., PMIDs 31549442 and 21709081). To enhance the novelty of their findings, the authors should prioritize new interactions, as demonstrating that rs4149032 influences rifampicin dosing does not contribute new knowledge. Furthermore, since some of these interactions are already present in the training data (PharmGKB), there is a risk that the model is merely recalling its training data rather than generating new insights.

3. LLM

The authors use GPT-4 (primarily gpt-4-turbo) for selecting 10 out of 50 genes, effectively performing variable or feature selection through prompt engineering, guided by prior knowledge of drug-gene relationships. Since, despite previous use,

LLMs are not yet an established method for variable selection and the context provided is most applicable to non-African populations, I wonder how accurately the LLM performed. Did the authors evaluate its accuracy, for example, by comparing its rankings against those provided by human experts?

The authors also state that *"the query was performed three times to minimize the chance of confabulation."* While LLM outputs can vary even when the prompt and context remain the same, setting temperature to 0 typically reduces variability. However, in the provided script (https://github.com/ersilia-os/pharmacogx-embeddings/blob/main/src/llm_reranking.py), the temperature setting is absent, meaning it defaulted to 1 (<https://platform.openai.com/docs/api-reference/chat/create>), which increases randomness. Given this, it is crucial that the authors report the consistency of model outputs across runs.

Additionally, the authors mention that *"consensus ranks with most trustworthy explanations"* (llm_reranking.py). How was trustworthiness determined? Did human reviewers assess and select the most reliable responses from the three outputs, or was an automated method used? Clarifying this process would strengthen the validity of the feature selection approach.

4. PGX signature

During the signature analysis, the authors state that *"the occurrence of words was proportional to the number of associated variants (3: 50+ variants, 2: 10-49 variants, 1: 1-9 variants)."* I wonder whether alternative categorizations (e.g., more granular categories) or a continuous scale (without categorization) were considered, and if not, why. The current grouping appears somewhat arbitrary – for example, 9 variants (1 word) are much closer to 10 variants (2 words) than 10 variants (2 words) are to 49 variants (also 2 words). A more refined classification or a continuous approach might better reflect the underlying distribution of variant associations.

5. Code availability

I was unable to demo at <https://pharmacogxembeddings.streamlit.app> due to the error "You do not have access to this app or it does not exist".

(Remarks on code availability)

The README file does not include instructions for installing and running the application, but I would have been able to run the code regardless. I reviewed most of the scripts, and they appeared well-structured – I have experience running similar scripts. Relevant details, such as the temperature setting for OpenAI's GPT and access limitations to the Streamlit app, are already addressed in the author comments above.

Reviewer #2

(Remarks to the Author)

The authors present an interesting and timely study that demonstrates the potential of integrating artificial intelligence/machine learning approaches with pharmacometrics and pharmacogenomics (PGx) to identify, through simulations, cases where treatment personalization for diseases such as tuberculosis (TB) and malaria can be improved. The results are promising and may contribute to the development of a methodology that incorporates relevant pharmacogenetic information in African populations affected by these diseases, thereby enabling more accurate treatment personalization and potentially reducing therapeutic failures. However, clinical PK implications interpretation should be understood with caution as they have not been validated with raw clinical data.

After reviewing the manuscript, I have the following specific comments:

First, I suggest softening the language in certain parts of the manuscript when discussing the implications of the findings. The results presented are based on simulations that have not yet been validated with clinical data. For instance, in the Discussion section, the authors state:

"In this study, we demonstrated how the systematic exploitation of publicly available PGx data across all genes and disease indications via a biomedical knowledge base, coupled with the use of an LLM for reasoned candidate selection, provides accurate suggestions for the improvement of pharmacometric modelling in African populations."

This statement may overstate the strength of the evidence, given the lack of clinical validation.

In line with this, and acknowledging the limitation the authors themselves mention regarding the absence of validation, it would be important to show how the simulations compare with available clinical data to demonstrate real-world impact—at least for the selected drugs i.e rifampicin and/or artemether.

I agree that this work, along with the simulations it presents, constitutes an important contribution to the field. However, validating these simulations would be of greater support.

Once the issue of validation has been addressed, I have the following minor comments:

In the presentation of the results, there is a frequent intermixing of methodological explanations that would be more appropriately placed in the Methods section. I suggest revisiting and reorganizing this part of the manuscript for improved clarity.

In the abstract, the methodology employed should be described with greater specificity to better inform the reader of the

approach taken.

Figure 1 should be reordered properly. Additionally, Figure 1A does not appear to add substantial information beyond what is already described in the main text and could be removed without loss of content.

The current layout of Figure 3 is confusing and would benefit from clearer presentation.

It should be indicated which PK models were used for each drug in the NLME modeling analysis to identify covariates and conduct the final Monte Carlo simulations. This information could be included in the Supplementary Materials. Were the models developed by the authors, or were previously published PK models from the literature used?

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

This is a well-written manuscript that integrates artificial intelligence (AI) with pharmacometric modeling to inform malaria and tuberculosis therapy in African populations. In this study, the authors developed a novel AI/machine learning approach to predict drug-gene pairs of pharmacogenetic interest, based on various large datasets, for a set of antimalarial and antituberculosis drugs relevant to African populations. Using the identified pharmacogenes, the authors further applied physiologically based pharmacokinetic (PBPK) modeling to predict the pharmacokinetics (PK) of artemether and rifampicin. The use of AI in individualized therapy has significantly increased in recent years and continues to attract substantial interest from researchers. This study presents an excellent integration of AI and pharmacometric modeling to address existing gaps in pharmacogenomics and drug response within the Africa population. The findings are clinically relevant and may help tailor treatment strategies for two major infectious diseases.

The AI and PBPK/NLME modeling analyses were well-executed, the figures are informative and of high quality, and the manuscript overall was a pleasure to read.

Below are some comments and suggestions for the authors:

- The number of pharmacogenes predicted by the PGx classifier for antimalarial and antituberculosis drugs appears to be greater than those reported in PharmGKB, as shown in Figure 2. To what extent can these predictions be supported or verified by existing literature? Please discuss.
- The AI-predicted pharmacogenes are primarily qualitative, with a subset further evaluated using PBPK modeling to identify sensitive genes. A key concern is the confidence level in these predictions. Is there any validation or external evidence supporting the sensitivity of these genes? The reviewer recommends discussing this in more detail.
- For PBPK modeling, certain gene-related kinetic parameters (e.g., SLCO1B1 rs4149032) were optimized in the model, which is appropriate. However, for other genes like AADAC and ABCB1—which are already implemented in the built-in PK-Sim model—optimization was also applied. Please clarify the rationale for optimizing parameters that already have default values in the software.
- The authors state they found evidence of variable PK and drug response in African cohorts, as shown in Table S5. However, the reviewer questions whether inter-individual variability (IIV)—a parameter used in NLME models—is the most appropriate metric to support this claim, as it can be influenced by numerous unknown factors. Typical values of CL and V may better reflect population-level differences. Please consider and discuss this point.
- Regarding PBPK predictions, it would be helpful to clarify that the simulations were conducted under single-dose drug administration conditions. Both artemether and rifampicin exhibit auto-induction of metabolism with repeated dosing, which appears to be beyond the scope of this study. Please confirm and clarify.
- Page 20, Figure 4B: Please clarify the meaning of the different colored lines. The reviewer assumes these represent PK profiles for different genotypes compared to the wild type. Explicit labeling or clarification in the figure legend would improve clarity.
- The description of the NLME methodology and results is limited. Please provide additional details in the Methods section (e.g., model structure, inter-individual variability model, model validation process) and include the PK parameter estimates in the supplementary material to enhance transparency and reproducibility.
- The reviewer encourages the authors to expand the discussion of future research directions, particularly in the context of drug development. The current AI-coupled modeling approach focuses on approved drugs for infectious diseases, which benefit from a wealth of existing data. However, a forward-looking perspective on how this approach could be applied in early-stage drug development would be valuable.

(Remarks on code availability)

The online code for data curation and training of the PGx classifier is adequately described and accompanied by a comprehensive README file, allowing the audience to follow the workflow easily.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you to the authors for addressing my concerns. There appears to be a minor typo in a figure description; specifically, "Figure R1. A, C and E. ...when 10% (A), 20% (B) and 50% (C)..." likely should read "Figure R1. A, C and E. ...when 10% (A), 20% (C) and 50% (E)..."

(Remarks on code availability)

The authors have improved the README file to provide more detailed instructions.

Reviewer #2

(Remarks to the Author)

They have addressed all the reviewer's comments. I do not have additional comments.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors have adequately addressed the reviewers' comments. This manuscript would be of interest to readers involved in drug development.

(Remarks on code availability)

No comments.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Reviewers' comments

We would like to thank the Reviewers for their helpful and useful feedback. We have responded and revised the manuscript accordingly. In brief, we have:

- Performed additional analyses on our pharmacogene ranking approach, with a special focus on the robustness and the accuracy of the methodology.
- Included a comparison between PK-simulated data and observed clinical data both for Rifampicin and Artemether.
- Modified figures to improve their clarity.
- Extended the Discussion section to include caveats of the study (especially related to clinical validation) and future perspectives.
- Improved our code repository with better documentation and updated code related to the current revision.

Below are detailed point-by-point responses to the Reviewer's comments point-by-point. Relevant changes to the manuscript text have been highlighted in yellow.

Reviewer #1

Thank you for the opportunity to review this manuscript, which investigates pharmacogenetics in an understudied population. The authors apply machine learning (ML) and artificial intelligence (AI) to identify drug-gene pairs relevant to key therapeutic areas, specifically malaria and tuberculosis, in African populations. These identified genes are then incorporated into pharmacometric modelling to assess their effects, including implications for dosing. I appreciate the significant effort and time invested in this research and in preparing the manuscript for publication.

We thank the Reviewer for the positive remarks.

The authors trained a supervised learning model for drug-pharmacogene prioritization using drug and gene descriptors, applied an LLM-based pharmacogene re-ranking approach, and performed PBPK and NLME modelling. They state that “we propose a pipeline that couples a novel ML/AI component...”; however, it is unclear which aspects are truly novel, given that LLMs have already been used for feature selection (e.g., <https://arxiv.org/abs/2407.02694>). Could the authors clarify what is specifically innovative in their approach? I outline additional concerns below.

The reference mentioned by the Reviewer is related to feature selection prior to model training, an excellent example of how LLMs are starting to be used in conjunction with classical (tabular, in this case) ML approaches. In the referenced article, the authors demonstrate that, given a task description and a collection of feature names, LLMs are able to prioritize predictive features for downstream ML modeling (e.g., classification). We formulated our problem slightly differently; more specifically, we divided our problem into two ranking tasks, namely ranking-1 and ranking-2, only one of which (ranking-2) used an LLM. Below, we clarify this two-step procedure and show how it differs to the referenced article.

In ranking-1, we aimed to predict the pharmacogenetic interaction of drug-gene pairs at scale using drug and gene featurization techniques. Many of those drug and gene representations are derived from highly specialized methodologies, such as the Chemical Checker or the Bioteque, that we previously developed and that are aimed at embedding prior knowledge about the genes and the compounds as found in bioinformatic and chemoinformatic databases, along with sequence and structure information. In ranking-1, the problem was formulated as a binary classification task, approached with an ensemble of predictors that took as features the concatenation of a given chemical and gene representation. This approach had several advantages, most notably:

- It allowed us to perform the study at a large scale, considering all possible drug-gene pairs, in a cost-effective manner.
- It enabled understanding (quantification) of the most predictive feature types, as shown in Figure 2C (here, features are not gene names or identifiers, but drugs or genes descriptors/embeddings).
- It allowed for learning patterns across genes and compounds, which is not immediately feasible if the approach is formulated as a feature selection task for each drug.

In ranking-1, for each drug, we selected 50 genes. In ranking-2, we aimed to re-rank these 50 genes to select the top 10 in a zero-shot manner based on LLMs. In the case of ranking-2, our approach is quite analogous to the one suggested in the referenced article. Indeed, for each drug, we elaborated a relatively complex prompt, containing drug and gene information (context) as well as a formulation of the ranking task (i.e., select 10 genes out of the 50). We decided to add the drug and gene information (context) explicitly because we wanted to maximize the transparency of the ML/AI pipeline; that said, in the prompt, we asked explicitly that the model use “its own knowledge” to make assessments, and not rely solely on the context provided in the prompt. In our case, a major reason for adding context was the fact that we wanted to provide summaries extracted from the DrugBank database, a well-curated resource of drug information.

In sum, the novelty of our approach lies in the use of embedding-based search techniques prior to LLM usage for drug-pharmacogene identification, followed by a “feature selection” procedure based on a prompt-engineered LLM querying, which is conceptually similar to the Jeong et al. (2025) preprint suggested by the Reviewer, with the exception that we rely on drug and gene summaries (and a list of the ranking-1 scores) in the prompt design, as opposed to the relatively more simplistic approach of relying on feature names alone. Jeong et al. is now referenced in

the main text, and a mention of “LLM-driven feature selection” is now included. Note that this work was published very recently and was not available at the time of the study. Finally, we believe that the main importance of our work, as previously noted by the Reviewer, is the coupling between drug-pharmacogene predictions and pharmacometric modeling. To the best of our knowledge, such a comprehensive approach combining different computational techniques has not been described thus far.

1. Use of PharmGKB as ‘ground truth’

Less than 4% of PharmGKB data comes from African populations [PMID: 37450899], as the authors also noted in a separate study [PMID: 38594395]. This underrepresentation increases the likelihood that African-specific variants are missing, leading to potential misclassification in downstream classification; particularly, some instances labelled as “no drug-gene interaction” may be incorrect. Moreover, analyses focusing on African-specific variants could amplify this bias, as they rely on a subset of the data where misclassification is more likely. The authors later identify significant predictions or drug-gene pairs that are absent from PharmGKB, further suggesting that PharmGKB may be missing important drug-gene interactions. If these novel interactions are indeed valid, this heightens the concern that the training data may not be optimal. The authors should acknowledge this limitation and, if possible, estimate its impact on the downstream results.

We thank the reviewer for raising this valid and relevant point. We have invested significant computational effort into addressing it. Indeed, in many cases, we are dealing with a “positive-unlabeled” scenario, where “negative” drug-gene pairs being fed into the model are “unknown” instances. We are aware of this situation, which, as the Reviewer rightly points out, might be amplified in the case of African-related variants. Importantly, our dataset is not only unlabeled but also highly imbalanced; that is, of all possible drug-gene pairs, only a very small portion represents “positive” pharmacogenetic interactions. We dealt with this imbalance by applying an ensemble-based approach that undersamples negative/unlabeled drug-gene pairs at a 1:10 ratio to the positives. This undersampling procedure is done multiple times, for multiple types of drug and gene embeddings/descriptors, which effectively means that the resulting ensemble-based classifier is more attentive to the positive pairs than to the negative ones.

It is noteworthy that during the undersampling of unlabeled samples as negatives, we exclude any drug-gene pairs with any reported PGx interaction at low evidence levels, whether PK or PD, reducing the chance of including false (unreliable) negatives in our training set. In sum, we use sampling strategies commonly applied in highly imbalanced, positive-unlabeled scenarios to achieve, to the best of our capacity, models that primarily exploit “positive” interactions and are robust to the possibility that some of the “negatives” might be undetected interactions.

To understand the effect of data contamination on our modelling exercise, we have now misclassified 10%, 20%, and 50% of the positive data as negative and included it in the randomly sampled negative data. We have retrained all models to assess performance degradation (Figure R1). We observe that our models remain robust when relatively small subsets (10% and 20% of the data) are misclassified, demonstrating that model predictions

remain relevant even in the presence of wrongly annotated training data. The subset most affected by the data misclassification exercise is the model trained with only ADME genes and only PK-related associations, as expected due to the small size of this filtered dataset. Remarkably, though, even in this scenario, the AUROC remains over 0.75 for both the 10% and 20% misclassification experiments. When half of the data is wrongly labeled, model performance decreases as expected and none of the subsets retains an AUROC above 0.75. Note, in addition, that—as expected—the number of predictive drug/gene representations decreases as more misclassifications are included (bottom row of Figure R1). This model evaluation has been included in the Methods section and as Supplementary Figure S4.

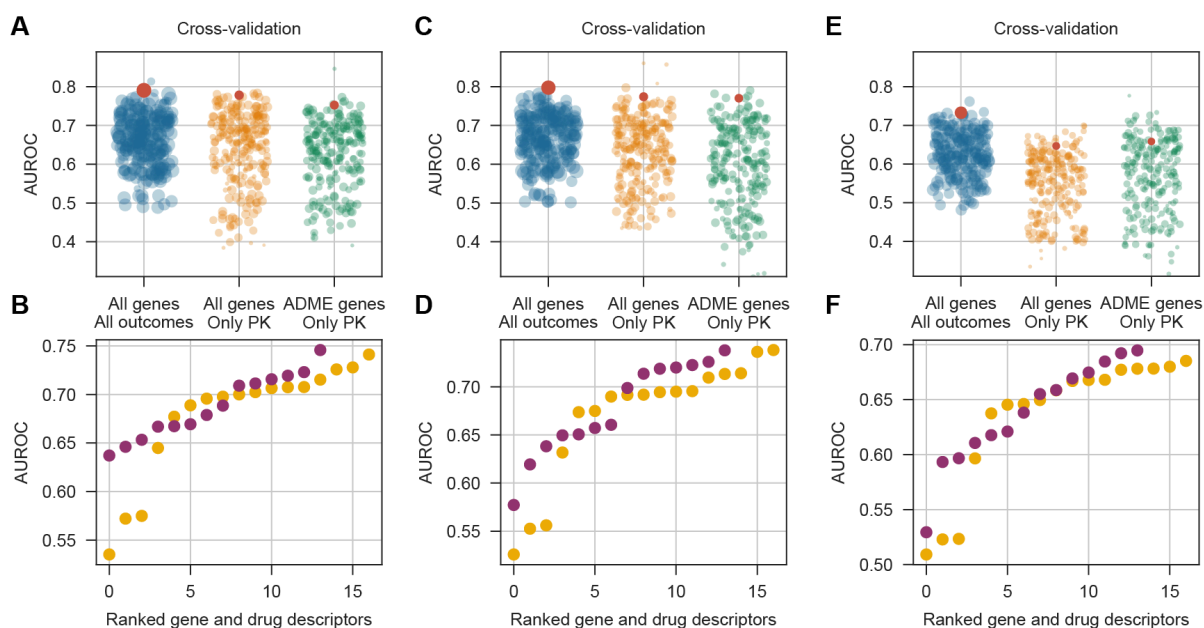


Figure R1. Effect of data contamination on model performance. **A, C and E.** Cross-validation performance measured by the Area Under the Curve (AUROC) of the different PGx classifiers and the performance of each of the three ensembles (red dot) when 10% (A), 20% (B) and 50% (C) of the positives are misclassified as negatives. **B, D and F.** Top quartile performance (AUROC) of classifiers using different individual drug and gene descriptors when 10% (B), 20% (D) and 50% (F) of the positives are misclassified as negatives.

It is commendable that the authors excluded level 4 evidence ('evidence does not support an association between the variant and the drug phenotype') from downstream analysis. However, level 3 evidence (which is based on a single study, preliminary findings, or inconsistent replication across studies) is also unreliable and should ideally be excluded. That said, removing true drug-gene associations may introduce misclassification, just as including false associations would. Can the authors perform a sensitivity analysis to estimate the impact of including low-quality (level 3 and probably level 6) data on the results?

We thank the Reviewer for this suggestion, which is tightly related to the comment above. Indeed, there is a trade-off between including and excluding lower-level evidence interactions.

As indicated by the Reviewer, the original data has been filtered to keep levels 1, 2, 3, 5, and 6, discarding level 4 associations (demonstrated non-associated variant-phenotype). We have performed further testing by restricting the data to very strict evidence (levels 1 and 2) and subsequently increasing the lower-quality data as follows: evidence levels 1, 2, and 3; evidence levels 1, 2, 5, and 6; and evidence levels 1, 2, and 6 (level 4 is always excluded from the positive training set, as it refers to validated non-associations). The performance of the models (measured by AUROC) remains excellent (above 0.75 for the ensemble models) in all cases, demonstrating that the models are robust to the presence or absence of higher- or lower-quality data (Figure R2). In Figure R2, the size of the dots is proportional to sample sizes; note how, as we become more restrictive with the evidence levels, the training data size shrinks significantly. This figure has now been referenced in the Methods section as Figure S5.

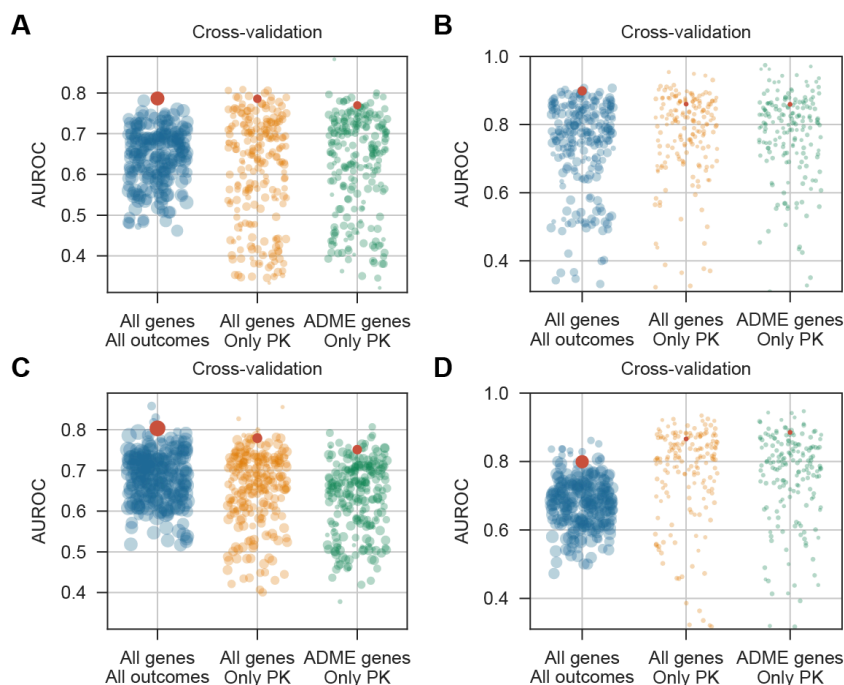


Figure R2. Cross-validation performance measured by the AUROC of the different classifiers and the performance of each of the ensembles (red dot) for **A.** Evidence levels 1, 2 and 3. **B.** Evidence levels 1 and 2. **C.** Evidence levels 1, 2, 5 and 6. **D.** Evidence levels 1, 2 and 6.

2. NLME and PBPK modelling

In Figure 4C-D, the authors highlight the interactions between artemether and CYP2B6*6, as well as rifampicin and SLCO1B1 rs4149032, both of which are well-established drug-gene interactions (e.g., PMIDs 31549442 and 21709081). To enhance the novelty of their findings, the authors should prioritize new interactions, as demonstrating that rs4149032 influences rifampicin dosing does not contribute new knowledge. Furthermore, since some of these interactions are already present in the training data (PharmGKB), there is a risk that the model is merely recalling its training data rather than generating new insights.

We agree that the interactions between *artemether*–CYP2B6*6 and *rifampicin*–SLCO1B1 rs4149032 are well-established in the literature. Our work goes further by incorporating

population specific allele frequencies in the modelling. For example, *SLCO1B1 rs4149032* has shown variable allele frequencies across African subpopulations. Our PBPK model accounts for allele frequency (e.g. *SLCO1B1 rs4149032* (as suggested by the Reviewer) or *CYP2B6*6*) to refine the predicted population pharmacokinetics (and corresponding optimal dose) in African populations. This context-specific application has rarely been addressed in previous studies. Secondly, in Figure 4A, our PBPK model included the *ABCB1* pharmacogene, which was predicted by the ML/AI pipeline and has not been reported in PharmGKB. In the PBPK analysis, we found this gene to be sensitive. However, its overall effect on exposure (AUC or C_{max}) was minimal. Importantly, its presence or absence in covariate modelling was not significant, suggesting that *ABCB1* may not warrant dose adjustment under the modelled conditions for either rifampicin or artemether. This demonstrates how coupling AI/ML to pharmacometric modelling allows to unravel new associations and ensures that only the clinically meaningful will proceed for further study. To address the Reviewer's comment, we have added more text in the Methods and Results section to bring out the fact that new ML/AI predicted pharmacogenes were integrated into the PBPK model and assessed their significance in NLME modelling: "The new ML/AI predicted pharmacogene interactions between artemether and rifampicin were introduced in the NLME model as covariates and assessed for their significance on the pharmacokinetic parameters. Other covariates introduced in the NLME model were the known enzyme/transporter variants prevalent in African populations". We have also clarified that the proposed optimal doses are not for all African populations as the optimal dose will depend on the frequency of the alleles that are significant in the African subpopulation of interest.

Finally, we would like to highlight that, as shown in Figure 3, in our PBPK sensitivity analysis (which we were able to conduct for 10 of the drugs), 25 of the (10x10=100) prioritized drug-gene pairs showed non-zero (>0.1) AUC or C_{max} sensitivity, 10 of which had a sensitivity >1.0. Of the 25 drug-gene pairs, 12 were not previously reported in PharmGKB, demonstrating the ability of the ML/AI pipeline to identify pharmacogenes of potential relevance to the downstream pharmacometric modeling.

3. LLM

The authors use GPT-4 (primarily gpt-4-turbo) for selecting 10 out of 50 genes, effectively performing variable or feature selection through prompt engineering, guided by prior knowledge of drug-gene relationships. Since, despite previous use, LLMs are not yet an established method for variable selection and the context provided is most applicable to non-African populations, I wonder how accurately the LLM performed. Did the authors evaluate its accuracy, for example, by comparing its rankings against those provided by human experts?

Yes, we did perform manual review of the results by DMPK/pharmacometrics experts. In particular, we prepared a web app for internal use (which we also make available via the following URL: <https://pharmacogx-embeddings.streamlit.app/>) to allow the experts to explore the embedding-based ranking (ranking-1) and the LLM-based re-ranking (ranking-2). Importantly, in ranking-2 we include "explanations" from the LLM, resulting from prompts containing explicit context from the drug and the gene, as well as the ranking-1 results with their associated scores. As discussed above, this can be seen as a variable/feature selection exercise, enriched with prior knowledge and natural language explanations. These explanations were important for the experts to gain trust in the predictions. For example, while the

metabolism of bedaquiline by CYP3A4 is reported in literature (PMID: 32331352), it is not annotated on PharmGKB. The model accurately ranks CYP3A4 as the most important enzyme but also accurately includes CYP3A5 (rank 3) with an explanation that infers its involvement in bedaquiline metabolism based on its contribution to the CYP3A family. This CYP3A4 and CYP3A5 co-ranking is also seen with other drugs not annotated on PharmGKB such as delamanid and piperazine. In addition, as mentioned in the text, the involvement of ABCB1 in rifampicin disposition is also not on PharmGKB and the pharmacogenetic link has only recently been published (PMID: 38315168). Other similar examples include the link between NAT2 and sulfadoxine, and CYP3A4 and delamanid.

Importantly, in Figure 2D, we show that while we are recapitulating PharmGKB-known interactions (orange), we are still including a substantial portion of newly predicted ones (green), especially for genes such as ABCB1, CYP2C19, CYP2D6, CYP3A4, and CYP2D7, which were associated with at least 10 drugs in ranking-1. Figure 2F demonstrates that the LLM reranking (ranking-2) did not have a drastic reordering of most genes. As explained in an earlier response, and precisely because LLMs have not yet been deeply investigated in the context of PGx, we preferred to apply a relatively conservative re-ranking (ranking-2) procedure using the LLM, with well-controlled and context-rich prompt engineering and mild filtering. In future work, we would like to be more expeditive in the prediction scope of the LLMs, potentially fine-tuning an LLM with PGx-specific literature to be run locally.

The authors also state that "the query was performed three times to minimize the chance of confabulation." While LLM outputs can vary even when the prompt and context remain the same, setting temperature to 0 typically reduces variability. However, in the provided script (https://github.com/ersilia-os/pharmacogx-embeddings/blob/main/src/llm_reranking.py), the temperature setting is absent, meaning it defaulted to 1 (<https://platform.openai.com/docs/api-reference/chat/create>), which increases randomness. Given this, it is crucial that the authors report the consistency of model outputs across runs.

We thank the Reviewer for raising this valid point. Indeed the temperature was defaulted to 1. To ensure this does not affect the results, we have evaluated the consistency of the model across the three rounds of prediction, which is reported in the following Figure R3. The top 10 gene selection performed by the LLM shows remarkable consistency, with over 50% of the drugs having more than 90% overlap in the top 10 genes. The lowest overlap is shown for Pretomanid, yet still with a remarkable 75% overlap (note that the random expected overlap between two re-ranking attempts is 20%). The overall consistency (mean overlap) is 88% across drugs. This figure has also been added as Figure S6A.

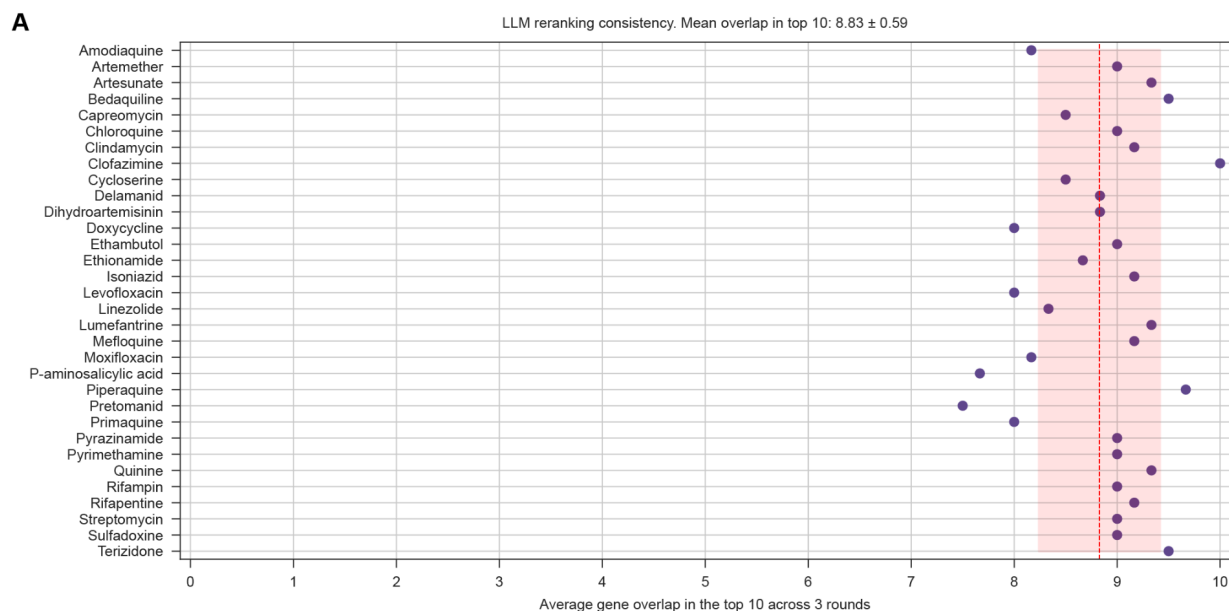


Figure R3. A. Overlap in the top 10 genes upranked by the LLM across three rounds for each drug.

Additionally, the authors mention that “consensus ranks with most trustworthy explanations” (llm_reranking.py). How was trustworthiness determined? Did human reviewers assess and select the most reliable responses from the three outputs, or was an automated method used? Clarifying this process would strengthen the validity of the feature selection approach.

In retrospect we realize that “trustworthy” may be a strong word with unclear statistical implications. We have removed it from the text. While we manually reviewed the explanations, we also used significant prompting to contextualize our queries, as shown in Table S4. In brief, prompts began as follows: “You are a pharmacogenetics expert. Your goal is to select the most trustworthy explanation of the pharmacogenetic relationship between a drug and a gene...”. Very importantly, note that the selection of the most plausible explanation did not alter the ranking of the drug-gene pairs. These explanations were simply intended to aid the DMPK/pharmacometrics experts when selecting examples for subsequent modeling.

Linked with the above comment, we have re-analyzed the responses by the LLM, evaluating the similarity between explanations for the same (red) and different (grey) drug-gene interactions, demonstrating that the LLM indeed produces very similar explanations for the same drug-gene pairs (Figure R4). Furthermore, GPT-based categorization of the explanations into predefined subgroups shows that most of the same drug-gene explanations are classified as “very similar,” whereas most of the different drug-gene interaction explanations are classified as “partly similar” (as expected, since they all belong to the same domain, pharmacogenetics). This analysis has been added together with Figure R3 as Figure S6.

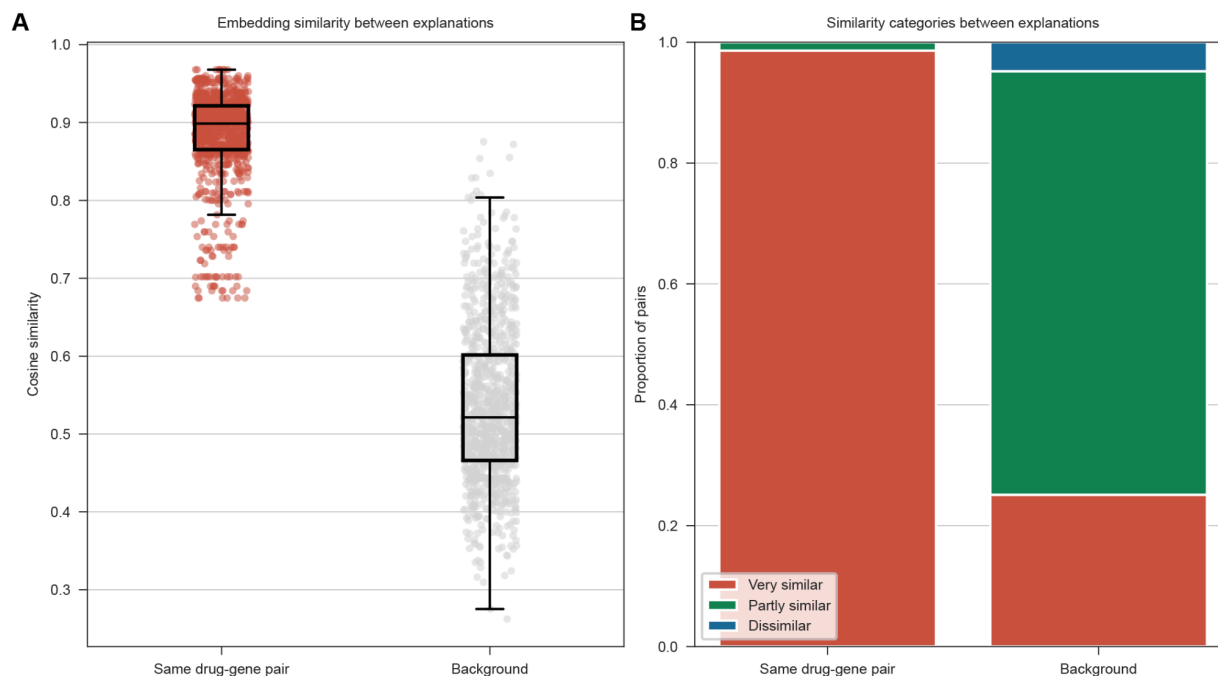


Figure R4. A. Cosine similarity between explanations related to the same drug-gene pair (red) vs unrelated pairs (grey) **B.** Categorization of similarity of the explanations in very similar, partly similar or dissimilar for same drug-gene pairs vs unrelated pairs.

In addition, in this table now included in the GitHub repository (https://github.com/ersilia-os/pharmacogx-embeddings/blob/main/paper_revision/results/explanations_for_reranking.csv), we provide instances where the three “explanations” can be found for each drug.

4. PGX signature

During the signature analysis, the authors state that “the occurrence of words was proportional to the number of associated variants (3: 50+ variants, 2: 10-49 variants, 1: 1-9 variants).” I wonder whether alternative categorizations (e.g., more granular categories) or a continuous scale (without categorization) were considered, and if not, why. The current grouping appears somewhat arbitrary – for example, 9 variants (1 word) are much closer to 10 variants (2 words) than 10 variants (2 words) are to 49 variants (also 2 words). A more refined classification or a continuous approach might better reflect the underlying distribution of variant associations.

We thank the Reviewer for raising this valid point that we did not sufficiently explore. The “topic modeling” strategy relies on word counts, so we could not use a continuous approach. To support our choice of categories (Figure R5A), we now demonstrate how alternative binning does not affect the signature analysis. We have chosen two alternative categorization strategies, one that expands the low range (quasi-logarithmic) (Figure R5B) and one that adds more granularity to the categorization using a bin of ten (Figure R5C). The different categorizations do not affect the signature analysis, as demonstrated by the stability of the top ten genes in each signature (Figure R5). We have added Figure R5 as Figure S3 in the manuscript.

A Variants	Words	Sign. 0	Sign. 1	Sign. 2	Sign. 3	Sign. 4	Sign. 5	Sign. 6	Sign. 7	Sign. 8	Sign. 9
1-9	1	CYP3A4	CYP2D6	CYP2C9	UGT1A9	CYP2B6	TPMT	DRD2	HLA-A	ESR1	GNB3
10-49	2	CYP3A5	SLC6A4	CYP2C8	UGT1A7	NAT2	TYMS	COMT	HLA-DRB1	ESR2	ACE
50+	3	NR1I2	SEPTIN3	G6PD	UGT1A10	CYP2C19	C18orf56	OPRM1	HLA-B	PGR	USP5
		SLCO1B1	WBP2NL	SLCO1B1	UGT1A8	CYP2E1	GATA3	ANKK1	HLA-DQB1	ERBB2	CDCA3
		ZSCAN25	ABCB1	NOS1AP	UGT1A4	GSTM1	MTHFR	SLC6A3	HLA-C	CYP2A6	ADD1
		ABCB1	APOE	VKORC1	UGT1A3	CYP2A6	PNPLA3	OPRD1	HLA-DQA1	CYP19A1	SCNN1G
		POR	FKBP5	CYP4F2	UGT1A5	TNF	BMP7	DBH	HLA-DRA	SULT1A1	SLC12A3
		ABCG2	PRKCE	STX4	UGT1A6	ABCB1	DOK5	SLC22A1	HLA-DRB5	VDR	ADRB1
		ABCC2	CHAT	ABCC8	UGT1A1	GSTT1	CPT1A	DRD3	PTPN22	MKI67	NOS3
		CYP2C19	CHRNA7	KCNJ11	ABCB1	NR1I2	IGHMBP2	BDNF	HSPA1B	ZNF613	ADRB2

B Variants	Words	Sign. 0	Sign. 1	Sign. 2	Sign. 3	Sign. 4	Sign. 5	Sign. 6	Sign. 7	Sign. 8	Sign. 9
1-4	1	CYP3A4	CYP2D6	CYP2C9	UGT1A1	CYP2B6	TPMT	DRD2	HLA-A	ESR1	GNB3
5-14	2	CYP3A5	SLC6A4	CYP2C8	UGT1A9	NAT2	TYMS	COMT	HLA-DRB1	CYP2A6	ACE
15-39	3	NR1I2	CYP2C19	G6PD	UGT1A7	CYP2C19	C18orf56	OPRM1	HLA-B	ESR2	ADD1
40-99	4	CYP2C19	WBP2NL	VKORC1	UGT1A8	CYP2E1	GATA3	OPRD1	HLA-DQB1	PGR	SLC12A3
100+	5	ABCB1	SEPTIN3	SLCO1B1	UGT1A10	CYP2A6	MTHFR	SLC22A1	HLA-C	ERBB2	SCNN1G
		ZSCAN25	BDNF	NOS1AP	UGT1A3	GSTM1	PNPLA3	SLC6A3	HLA-DQA1	CYP19A1	USP5
		SLCO1B1	ABCB1	CYP2C19	UGT1A4	ABCB1	BMP7	DBH	HLA-DRA	SULT1A1	CDCA3
		POR	FKBP5	CYP4F2	UGT1A6	NR1I2	DOK5	ANKK1	CYP2C19	MKI67	ADRB1
		ABCC2	APOE	STX4	UGT1A5	TNF	IGHMBP2	OPRK1	HLA-DRB5	ADRB2	ADRB2
		ABCG2	CYP3A5	ABCC8	ABCB1	GSTT1	CPT1A	DRD3	PTPN22	ZNF613	CYP11B2

C Variants	Words	Sign. 0	Sign. 1	Sign. 2	Sign. 3	Sign. 4	Sign. 5	Sign. 6	Sign. 7	Sign. 8	Sign. 9
1-10	1	CYP3A4	CYP2D6	CYP2C9	UGT1A9	CYP2B6	TPMT	DRD2	HLA-A	CYP2A6	ACE
11-20	2	CYP3A5	SLC6A4	CYP2C8	UGT1A1	NAT2	TYMS	COMT	HLA-DRB1	ESR1	GNB3
21-30	3	NR1I2	BDNF	G6PD	UGT1A7	CYP2C19	C18orf56	OPRM1	HLA-B	ESR2	ADD1
31-40	4	SLCO1B1	CYP2C19	SLCO1B1	UGT1A8	CYP2E1	GATA3	SLC22A1	HLA-DQB1	PGR	SLC12A3
41-50	5	ABCB1	WBP2NL	VKORC1	UGT1A10	GSTM1	PNPLA3	OPRD1	HLA-C	DPYD	SCNN1G
51-60	6	ZSCAN25	SEPTIN3	CYP2C19	UGT1A4	ABCB1	BMP7	DBH	HLA-DQA1	CYP19A1	CDCA3
61-70	7	CYP2C19	FKBP5	NOS1AP	UGT1A3	TNF	DOK5	SLC6A3	HLA-DRA	ERBB2	USP5
71-80	8	POR	APOE	STX4	UGT1A5	NR1I2	MTHFR	ANKK1	HLA-DRB5	CHRNA5	CYP11B2
81-90	9	ABCG2	ABCB1	CYP4F2	UGT1A6	GSTT1	IGHMBP2	OPRK1	HLA-DRB1	CDC148	ADRB1
91-100	10	ABCC2	CHAT	ABCC8	ABCB1	HLA-B	CPT1A	DRD3	CYP2C19	TNFRSF11B	ADRB2
100+	11										

Figure R5. Granular categorization of variants. **A.** Top ten genes in each PGx signature with the original variant categorisation into three discrete bins **B.** Top ten genes in each PGx signature when the variant categorisation is done with a quasi-logarithmic binning **C.** Top ten genes in each PGx signature with a granular variant categorisation into bins of 10.

5. Code availability

I was unable to demo at <https://pharmacogxembeddings.streamlit.app> due to the error “You do not have access to this app or it does not exist”.

There is a typo in the URL provided by the Reviewer (a hyphen is missing). The correct URL written in the manuscript was: <https://pharmacogx-embeddings.streamlit.app/>. This URL should work. Note that, to spare computational resources, if the app is not being used by other users, StreamLit (the cloud provider) puts the resource on sleep — so, it is possible that a message asking to launch the app appears when the Reviewer first requests access to it. Launching the app should take less than one minute.

Reviewer #1 (Remarks on code availability):

The README file does not include instructions for installing and running the application, but I would have been able to run the code regardless. I reviewed most of the scripts, and they appeared well-structured – I have experience running similar scripts. Relevant details, such as

the temperature setting for OpenAI's GPT and access limitations to the Streamlit app, are already addressed in the author comments above.

We thank the Reviewer for this positive feedback. We have improved the README file to provide instructions on how to install and run the computational pipeline.

Reviewer #2 (Remarks to the Author):

The authors present an interesting and timely study that demonstrates the potential of integrating artificial intelligence/machine learning approaches with pharmacometrics and pharmacogenomics (PGx) to identify, through simulations, cases where treatment personalization for diseases such as tuberculosis (TB) and malaria can be improved. The results are promising and may contribute to the development of a methodology that incorporates relevant pharmacogenetic information in African populations affected by these diseases, thereby enabling more accurate treatment personalization and potentially reducing therapeutic failures. However, clinical PK implications interpretation should be understood with caution as they have not been validated with raw clinical data.

We thank the Reviewer for the positive remarks about our work, and for highlighting that the results can be promising for the PGx field in Africa. We agree that the PK implications of our study should be understood with caution, since they lack prospective clinical validation on raw data. To the best of our capacity, in the updated version of the manuscript, we have included additional tables strengthening the pharmacometrics modeling aspects, as suggested by the Reviewer. Moreover, we have now included explicit text in the manuscript highlighting the caveats of the study related to the lack of raw clinical data validation.

After reviewing the manuscript, I have the following specific comments:

First, I suggest softening the language in certain parts of the manuscript when discussing the implications of the findings. The results presented are based on simulations that have not yet been validated with clinical data. For instance, in the Discussion section, the authors state: "In this study, we demonstrated how the systematic exploitation of publicly available PGx data across all genes and disease indications via a biomedical knowledge base, coupled with the use of an LLM for reasoned candidate selection, provides accurate suggestions for the improvement of pharmacometric modelling in African populations." This statement may overstate the strength of the evidence, given the lack of clinical validation.

We agree with the Reviewer. We have now softened the highlighted statement and others in the manuscript Discussion, and we have extended the caveats paragraph to more explicitly say that our study is a computational one and therefore results should be interpreted with caution. For example, the statement highlighted by the Reviewer now reads as follows: "In this study, we demonstrated how the systematic exploitation of publicly available PGx data across all genes and disease indications via a biomedical knowledge base, coupled with the use of an LLM for reasoned candidate selection, is able to suggest putative pharmacogenes that can be incorporated to pharmacometrics modelling and guide the design of clinical studies focused on understanding African pharmacogenomics." In the revised manuscript, we emphasize the fact

that prospective clinical validation is not pursued within this study, and we outline future directions for our work. The updated text is highlighted in yellow.

In line with this, and acknowledging the limitation the authors themselves mention regarding the absence of validation, it would be important to show how the simulations compare with available clinical data to demonstrate real-world impact—at least for the selected drugs i.e rifampicin and/or artemether. I agree that this work, along with the simulations it presents, constitutes an important contribution to the field. However, validating these simulations would be of greater support.

We agree with the Reviewer. We previously validated the PBPK predictions using clinical data cited in Table S6. In the current version of the manuscript, we have added a second table (Table R1) of observed (clinical PK) and simulated PK, including percentage bias for both Artemether and Rifampicin. As can be seen, for key PK parameters such as AUC, C_{max} and T_{max}, we achieved biases between observed and predicted values below acceptable bias of $\pm 15\%$, suggesting that our PBPK models were well adjusted to clinical PK data from African cohorts (Abdullahi et al. 2016 <https://doi.org/10.1002/jcph.1527>, Hodel et al. 2013 <https://doi.org/10.1128/aac.01700-12>, Naidoo et al. 2019 <https://doi.org/10.2217/pgs-2018-0166>, Chigutsa et al. 2011 <https://doi.org/10.1128/aac.01833-10>). Table R1 is now provided as a supplementary Table S7 and text has been updated in the Methods and Results sections correspondingly.

Drug	Parameter	Observed	Predicted	% Bias
Artemether	AUC _{0-8h} (ng*h/mL)	113.6	121.2	-6.7
	C _{max} (ng/mL)	32.5	35.3	-8.6
	T _{max} (h)	2	1.81	9.5
Rifampicin	AUC _{0-24h} (µg*h/mL)	70.1	64.3	8.3
	C _{max} (µg/mL)	6.6	5.9	10.6
	T _{max} (h)	6	5.4	10

Table R1. Comparison of pharmacokinetic parameters from observed and predicted concentration-time profiles of artemether and rifampicin.

Once the issue of validation has been addressed, I have the following minor comments:

In the presentation of the results, there is a frequent intermixing of methodological explanations that would be more appropriately placed in the Methods section. I suggest revisiting and reorganizing this part of the manuscript for improved clarity. In the abstract, the methodology employed should be described with greater specificity to better inform the reader of the approach taken.

We have now placed the methodological explanations in the Methods section for improved clarity. We kindly refer the Reviewer to this section (see, for example, “Africa-abundant and Africa-specific variants”). In addition, we have modified the Abstract with extra text informing the reader of the approach taken: “Our pipeline couples machine learning (ML) and artificial intelligence (AI) with physiologically-based pharmacokinetics (PBPK) and non-linear mixed effects (NLME) modeling to hypothesize which pharmacogenes could be of potential clinical interest, and which dose adjustments could be made to provide better treatment outcomes for the African population. Drug-gene pairs are first ranked with the latest knowledge embedding techniques, based on public structural and bioactivity data for drugs and genes, followed by a large language model (LLM) based refinement. Selected genes are then evaluated for their sensitivity in PBPK analysis, and relevant variants subsequently inspected with NLME for dose optimization.”

Figure 1 should be reordered properly. Additionally, Figure 1A does not appear to add substantial information beyond what is already described in the main text and could be removed without loss of content. The current layout of Figure 3 is confusing and would benefit from clearer presentation.

We have corrected the labelling of Figure 1 as suggested by the Reviewer. We have decided to leave Figure 1A for the reader to have a quick overview of the infectious vs non-communicable disease-related annotations, representing what is explained in the text. This is a core motivation of the work and, while the plot is simple, we feel it helps the reader understand the context of an otherwise quite populated and dense figure. In addition, we have improved the layout of Figure 3 to make it clearer.

It should be indicated which PK models were used for each drug in the NLME modeling analysis to identify covariates and conduct the final Monte Carlo simulations. This information could be included in the Supplementary Materials. Were the models developed by the authors, or were previously published PK models from the literature used?

We developed both Artemether and Rifampicin models from the concentration-time profiles that were predicted by the PBPK model. The final model for Artemether was best described by a 2-compartment PK model with linear elimination which consisted of CYP2B6 polymorphism as significant covariates. For Rifampicin, the final model was best described by a 1-compartment model and linear elimination with SLCO1B1 polymorphism as significant covariate. The final models with covariate effects were used to perform Monte Carlo simulations to determine the optimum dose for African populations.

To address the Reviewer's concern, text on this information has been added to Methods and Results sections of the manuscript: “The final model for artemether was best described by a 2-compartment PK model with linear elimination which had consisted of CYP2B6 polymorphism as covariate. For rifampicin, the final model was best described by a 1-compartment model and linear elimination with SLCO1B1 polymorphism as covariate. The final models were used to perform Monte Carlo simulations”. In addition, as suggested by the Reviewer, a supplementary Table S8 of the final PK model parameters for artemether and rifampicin has been added.

Reviewer #3 (Remarks to the Author):

This is a well-written manuscript that integrates artificial intelligence (AI) with pharmacometric modeling to inform malaria and tuberculosis therapy in African populations. In this study, the authors developed a novel AI/machine learning approach to predict drug-gene pairs of pharmacogenetic interest, based on various large datasets, for a set of antimalarial and antituberculosis drugs relevant to African populations. Using the identified pharmacogenes, the authors further applied physiologically based pharmacokinetic (PBPK) modeling to predict the pharmacokinetics (PK) of artemether and rifampicin.

The use of AI in individualized therapy has significantly increased in recent years and continues to attract substantial interest from researchers. This study presents an excellent integration of AI and pharmacometric modeling to address existing gaps in pharmacogenomics and drug response within the Africa population. The findings are clinically relevant and may help tailor treatment strategies for two major infectious diseases.

The AI and PBPK/NLME modeling analyses were well-executed, the figures are informative and of high quality, and the manuscript overall was a pleasure to read.

We thank the Reviewer for the positive feedback and encouraging comments about the overall manuscript narrative and the potential impact of the study for African populations.

Below are some comments and suggestions for the authors:

The number of pharmacogenes predicted by the PGx classifier for antimalarial and antituberculosis drugs appears to be greater than those reported in PharmGKB, as shown in Figure 2. To what extent can these predictions be supported or verified by existing literature? Please discuss.

This question is related to a similar comment made by Reviewer 1. We tried to address this point in several ways. First, we performed a cross-validation (CV) study whereby 20% of the PharmGKB data (drug-gene pairs) was held out for testing. On average, even in our initial (large-scale) ranking based on drug and gene embeddings, the ROC AUC score was above 0.8, indicating good classification capacity of drug-gene pairs. Importantly, we went beyond this systematic performance analysis by developing a web app intended for manual exploration of the results (<https://pharmacogx-embeddings.streamlit.app>). Predicted drug-gene pairs were presented in the app for our team to browse — in this process, we identified, for example, the case of ABCB1 involvement in rifampicin disposition, which was not reported in PharmGKB but was predicted by our method and has been described in the literature (PMID: 38315168). Other similar examples include CYP3A4 and CYP3A5 as relevant pharmacogenes for bedaquiline, NAT2 for sulfadoxine, and CYP3A4 for delamanid.

In addition, we argue that an orthogonal layer of validation is provided by the PBPK sensitivity analysis, in which we demonstrate that a significant number of the predicted drug-gene pairs affected PBPK modeling outcomes (Figure 3). Overall, we scoped the ML/AI part of the work as a ranking exercise, where we aimed to uprank, for each drug, genes of potential PGx relevance. We were able to demonstrate that our method significantly upranked relevant genes reported in the literature. These genes were then used as input for PBPK modeling to enable further

filtering and computational validation. We found that PBPK models tended to be sensitive to these genes. We have now included clarifying text in the Discussion section to highlight these points.

The AI-predicted pharmacogenes are primarily qualitative, with a subset further evaluated using PBPK modeling to identify sensitive genes. A key concern is the confidence level in these predictions. Is there any validation or external evidence supporting the sensitivity of these genes? The reviewer recommends discussing this in more detail.

This is an important point. Sensitivity analysis in PBPK modelling can only be performed if there is an external pharmacokinetic (PK) dataset. We were able to collect such data for Bedaquiline, Pretomanid, Amodiaquine, Artemether, Isoniazid, Lumefantrine, Mefloquine, Moxifloxacin, Quinine, and Rifampicin (Rifampin). For each predicted gene, a value of zero is assigned (as the starting value) to all kinetic parameters associated with it. This allows the algorithm to perform sensitivity analysis on the impact of that particular gene on the PK parameters of the drug of interest. Thus, the pharmacogene predicted qualitatively in the ML/AI step is assigned to a value that can be used to compute a sensitivity coefficient—a statistic that provides insight into whether the gene significantly affects PK parameters. To address the Reviewer's comment, we have expanded the Methods section to include additional details on this process. The sensitivity analysis results for the drugs listed above are now available in our GitHub repository (https://github.com/ersilia-os/pharmacogx-embeddings/blob/main/paper_revision/results/llm_top_10_with_pbpk_sensitivity.csv), highlighting how the ML/AI-based pharmacogene prioritization was able to identify genes likely to be sensitive in PBPK modelling.

Finally, to address a related comment by Reviewer 2, we have also included a supplementary Table S7, which contains observed (clinical PK) and simulated PK data, including percentage bias, for both Artemether and Rifampicin. We believe these validations further strengthen the relevance and robustness of the computational pipeline.

For PBPK modeling, certain gene-related kinetic parameters (e.g., SLCO1B1 rs4149032) were optimized in the model, which is appropriate. However, for other genes like AADAC and ABCB1—which are already implemented in the built-in PK-Sim model—optimization was also applied. Please clarify the rationale for optimizing parameters that already have default values in the software.

We thank the Reviewer for raising this important concern. The kinetic parameters of genes vary and depend on the genetics of the populations where the cell lines or microsomes are derived from. While the kinetic parameters for AADAC and ABCB1 are implemented in the built-in PK-Sim, it is noteworthy that kinetic parameters of these genes were also optimized (parameter identification). This is because the in-vitro tools used to determine kinetic parameters of these genes are not derived from the same population as the clinical PK data that they used to validate the PBPK prediction. In our case, we are using clinical data derived from African populations to validate PBPK predictions. Hence, AADAC and ABCB1 kinetic parameters in our study cannot be the same as those in PK-Sim since we are not using the same clinical dataset to validate predictions. For example the optimized k_{cat} for AADAC in PK-Sim is 9.87 min^{-1} while

in our study it's 6.5 min⁻¹. Similarly, the optimized kcat for SLCO1B1 in PK-Sim is 5.21 min⁻¹ while in our study kcat for SLCO1B1 rs4149032 is 2.87 min⁻¹. To address this comment, we have added a column in Table S6 showing the original PK-Sim kcat parameters and text in the Methods section covering this aspect: "When built-in PK-Sim parameters were available for a given gene, those were re-adjusted to fit clinical pharmacokinetic studies performed in African populations"

The authors state they found evidence of variable PK and drug response in African cohorts, as shown in Table S5. However, the reviewer questions whether inter-individual variability (IIV)—a parameter used in NLME models—is the most appropriate metric to support this claim, as it can be influenced by numerous unknown factors. Typical values of CL and V may better reflect population-level differences. Please consider and discuss this point.

We agree with the Reviewer on the fact that inter-individual variability (IIV) is influenced by many factors. And yes, the typical values of PK parameters in Table S5 reflect population-level, not individual-level, differences—in fact, these were extracted from population pharmacokinetics studies conducted in African populations. A few of these studies included pharmacogenetic factors as covariates in the population pharmacokinetic analyses to account for variation in PK parameters. However, most of these studies only tested demographic characteristics (e.g., age, weight, sex), concomitant medications, and comorbidities as covariates, which still could not fully explain the high variation observed in PK (i.e., high IIV). Although IIV is affected by many factors, not all of them are clinically significant or warrant dose adjustment. Our study is focused on genetics as a source of PK variation and assesses its impact on dosing in different African populations. Therefore, all typical PK values and the proposed doses are intended for the population level, not for individual patients. We have now clarified this point in the Methods section.

Regarding PBPK predictions, it would be helpful to clarify that the simulations were conducted under single-dose drug administration conditions. Both artemether and rifampicin exhibit auto-induction of metabolism with repeated dosing, which appears to be beyond the scope of this study. Please confirm and clarify.

Our modelling and simulations did not consider the auto-induction with repeated dosing as it is beyond the scope of the current manuscript. We have clarified this point in the Discussion: "known clinical phenomena such as auto-induction of metabolism with repeated dosing in artemether and rifampicin are outside the scope of the current work described in this manuscript."

Page 20, Figure 4B: Please clarify the meaning of the different colored lines. The reviewer assumes these represent PK profiles for different genotypes compared to the wild type. Explicit labeling or clarification in the figure legend would improve clarity.

Indeed, colors represent different genotypes compared to the WT. We have clarified this in the figure legend.

The description of the NLME methodology and results is limited. Please provide additional details in the Methods section (e.g., model structure, inter-individual variability model, model

validation process) and include the PK parameter estimates in the supplementary material to enhance transparency and reproducibility.

We agree with the reviewer. We have provided details of NLME methodology including the final population pharmacokinetics parameters for Artemether and Rifampicin (see the new Table S8).

The reviewer encourages the authors to expand the discussion of future research directions, particularly in the context of drug development. The current AI-coupled modeling approach focuses on approved drugs for infectious diseases, which benefit from a wealth of existing data. However, a forward-looking perspective on how this approach could be applied in early-stage drug development would be valuable.

We thank the Reviewer for this suggestion. Indeed, in a short article published in *Nature* last year (Turon et al., 2024, DOI: 10.1038/d41586-024-01001-y), we discussed some of the potential future directions, including the use of foundation models and large-scale datasets to better exploit AI in the context of PGx in Africa. We have now embedded some of our thoughts on this in the updated version of the Discussion section of the revised manuscript. To more specifically address the question about applying this approach in early-stage drug development, where drug data is not yet abundant, we argue that the key lies in the use of embeddings for drugs and genes. In the present manuscript, we have demonstrated that embedding drug and gene information (even at the pre-clinical level) can provide predictive power for PGx interactions. Importantly, we have a track record in developing knowledge embedding techniques for drugs and genes, including inference and imputation in data-sparse contexts. These techniques are embodied in the Chemical Checker and Bioteque frameworks (e.g., Duran Frigola et al., *Nat Biotechnol*, 2020; Bertoni et al., *Nat Commun*, 2021; Fernandez-Torras et al., *Nat Commun*, 2023; Comajuncosa-Creus et al., *Nat Protocol*, 2025). The central idea behind the Chemical Checker, for example, is to infer “bioactivity signatures” from chemical structure alone, or from any other available compound data—this aligns closely with the goal of predicting late-stage drug development outcomes from early-stage data points. The approach presented here naturally incorporates descriptors and embeddings from both the Chemical Checker and Bioteque as feature data for predictive modeling. Therefore, as suggested by the Reviewer, future work should focus (ideally prospectively) on the potential of the approach when only early-stage data is available. In fact, this is an avenue we are already beginning to explore. We believe that combining the current approach with (a) in vitro and in vivo ADMET data and (b) whole-genome data from biobanks holds great promise for anticipating PGx effects earlier in the drug development pipeline—potentially even at the variant-level rather than just at the gene-level. Please note that we have included additional text along these lines in the Discussion section.

Reviewer #3 (Remarks on code availability):

The online code for data curation and training of the PGx classifier is adequately described and accompanied by a comprehensive README file, allowing the audience to follow the workflow easily.

Please note that we have further improved the README file to make it more readable and include better installation instructions.

Reviewer #1 (Remarks to the Author):

Thank you to the authors for addressing my concerns. There appears to be a minor typo in a figure description; specifically, "Figure R1. A, C and E. ...when 10% (A), 20% (B) and 50% (C)..." likely should read "Figure R1. A, C and E. ...when 10% (A), 20% (C) and 50% (E)..."

Thanks for pointing this out, we have corrected the typo in the figure S1 legend.

Reviewer #1 (Remarks on code availability):

The authors have improved the README file to provide more detailed instructions.

Thanks, we appreciate the importance of clear README files for reproducibility

Reviewer #2 (Remarks to the Author):

They have addressed all the reviewer's comments. I do not have additional comments.

Thanks

Reviewer #3 (Remarks to the Author):

The authors have adequately addressed the reviewers' comments. This manuscript would be of interest to readers involved in drug development.

Thanks

Reviewer #3 (Remarks on code availability):

No comments.