The reviewer’s comments are presented in black

The authors’ responses are highlighted in blue.

Reviewer's Responses to Questions

Note: In order to effectively convey your recommendations for improvement to the author(s), and help editors make well-informed and efficient decisions, we ask you to answer the following specific questions about the manuscript and provide additional suggestions where appropriate.

1. Are the objectives and the rationale of the study clearly stated?

Please provide suggestions to the author(s) on how to improve the clarity of the objectives and rationale of the study. Please number each suggestion so that author(s) can more easily respond.

Reviewer #2: Yes

2. If applicable, is the application/theory/method/study reported in sufficient detail to allow for its replicability and/or reproducibility?

Please provide suggestions to the author(s) on how to improve the replicability/reproducibility of their study. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #2: Mark as appropriate with an X:

Yes [x] No [] N/A []

Provide further comments here:

3. If applicable, are statistical analyses, controls, sampling mechanism, and statistical reporting (e.g., P-values, CIs, effect sizes) appropriate and well described?

Please clearly indicate if the manuscript requires additional peer review by a statistician. Kindly provide suggestions to the author(s) on how to improve the statistical analyses, controls, sampling mechanism, or statistical reporting. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #2: Mark as appropriate with an X:

Yes [x] No [] N/A []

Provide further comments here:

4. Could the manuscript benefit from additional tables or figures, or from improving or removing (some of the) existing ones?

Please provide specific suggestions for improvements, removals, or additions of figures or tables. Please number each suggestion so that author(s) can more easily respond.

Reviewer #2: No further improvements are required.

5. If applicable, are the interpretation of results and study conclusions supported by the data?

Please provide suggestions (if needed) to the author(s) on how to improve, tone down, or expand the study interpretations/conclusions. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #2: Mark as appropriate with an X:

Yes [x] No [] N/A []

Provide further comments here:

6. Have the authors clearly emphasized the strengths of their study/theory/methods/argument?

Please provide suggestions to the author(s) on how to better emphasize the strengths of their study. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #2: Yes

7. Have the authors clearly stated the limitations of their study/theory/methods/argument?

Please list the limitations that the author(s) need to add or emphasize. Please number each limitation so that author(s) can more easily respond.

Reviewer #2: Yes

8. Does the manuscript structure, flow or writing need improving (e.g., the addition of subheadings, shortening of text, reorganization of sections, or moving details from one section to another)?

Please provide suggestions to the author(s) on how to improve the manuscript structure and flow. Please number each suggestion so that author(s) can more easily respond.

Reviewer #2: No further improvements are required.

9. Could the manuscript benefit from language editing?

Reviewer #2: No

Reviewer #2: The authors have made detailed responses to the reviewers. The evaluation metrics of predictive modeling is worthy of investigation. It is suggested to review more existing studies investigating the gap between prediction accuracy metrics and model performance, such as <https://doi.org/10.1016/j.est.2024.112126>.

Thank you for your suggestion. We have incorporated three additional studies to further address this important topic (L647-665). The first study investigates how discrepancies between evaluation metrics and real-world performance can arise from both data and modeling perspectives. We then include the paper you recommended as a concrete example of this phenomenon. Finally, we introduce the third study, which addresses these discrepancies by minimizing the domain differences that cause them.