

Dear Reviewer,

Thank you very much for your insightful and constructive comments on our manuscript (metabolites-4193302). We have carefully considered all your suggestions and have revised the manuscript accordingly. All changes in the revised version are highlighted in **blue** to distinguish them from first-round revisions (which were in red). Below we provide our point-by-point responses to your comments.

Comments1. The first issue concerns how performance superiority is framed. I noticed that the revised manuscript now reports repeated cross-validation results with mean values and standard deviations, which is certainly an improvement. Even so, some of the wording still suggests a stronger and more universal advantage than the tables really show. In places where the margins are small, or where the model is not uniformly best across all settings, I would encourage the authors to be more measured. A brief statistical comparison against the strongest competing methods would help. If the authors prefer not to add formal tests, then the language of the claims should be adjusted accordingly. At this stage, what I would like to see is not necessarily more experimentation, but a tighter match between the evidence and the wording.

**Response 1:** We thank the reviewer for this careful observation. We fully agree that the language describing performance should be precisely aligned with the experimental evidence. To address this, we have adjusted several statements in the manuscript where the wording was overly absolute or did not accurately reflect the results shown in Table 1 (e.g., M-GNN is the highest on STRINGdb and IRefIndex but second-highest on CPDB). Specifically:

- In the **Abstract**, we changed “consistently achieves the highest AUPRC” to “consistently achieves the highest or second-highest AUPRC”.
- In **Section 3.1**, we replaced “superior classification ability” with “strong classification ability” and “optimal performance” with “strong performance”.

These revisions are highlighted in **blue** to distinguish them from first-round changes (which were in red). No additional experiments were performed; the wording now matches the actual data more closely. We believe this adjustment addresses the reviewer’s concern.

--- Page 1, Abstract, line 14

--- Page 11, Section 3.1 Performance Evaluation of the M-GNN, lines 406-409

Comments2. I also remain unconvinced that the current discussion of ranking-oriented evaluation goes far enough. I understand the authors' point that AUC and AUPRC are standard and informative, and I do not regard the absence of precision@k or NDCG as fatal. Still, for a task that is ultimately about prioritizing candidate driver genes, those measures are not peripheral. They speak directly to how useful the model would be in practice. Since these metrics have not been added in the revision, I think the manuscript should say more explicitly why they are absent and what that means for the interpretation of the reported results. Right now, this point is acknowledged, but it still feels slightly deferred rather than fully confronted.

**Response 2:** We thank the reviewer for this important follow-up. In response, we have added a dedicated paragraph in the **Discussion** section (highlighted in **blue**) that now explicitly addresses:

- **Why ranking metrics are not included:** (i) AUC and AUPRC are the established standards in the field, as exemplified by recent works such as CancerGATE [ 51] and MLGCN-Driver [ 52], which use them exclusively; (ii) all five baseline models were originally evaluated only with these metrics, and their raw per-gene predictions are unavailable, making a fair ranking-based comparison infeasible without re-implementing and retraining each baseline—a process that would introduce new variability and risk of unfairness.
- **What this means for interpreting our results:** The repeated cross-validation experiments (Supplementary Figure S2) confirm the robustness of our model, and the consistent superiority in AUC and AUPRC across most networks (Table 1), together with the ablation study (Table 2), provides strong evidence of its effectiveness. While ranking metrics would add complementary practical insight, their absence does not invalidate our core conclusions.

We believe this addition makes the discussion more balanced and transparent, directly confronting the limitation rather than merely deferring it. All changes are highlighted in blue.

--- Page 16-17, Discussion, lines 591-607

Comments3. The revision on negative sampling is helpful, but I would encourage the authors to present it with a bit more caution. I appreciate that the manuscript now openly recognizes the risk that some unlabeled genes may in

fact be true drivers, and the sensitivity analyses across random seeds and class ratios are worthwhile additions. Even so, these checks do not really eliminate the underlying labeling problem; they only suggest that the reported performance is not overly fragile under the chosen setup. I would therefore recommend softening any implication that the issue has been settled. A more explicit statement in the limitations or discussion section would make the paper more balanced and, in my view, more credible.

**Response 3:** We thank the reviewer for this important clarification. We fully agree that the sensitivity analyses (Supplementary Figure S2 and Supplementary Table S5) demonstrate robustness under the chosen experimental settings but do not eliminate the inherent labeling uncertainty—i.e., that some undiscovered driver genes may still reside in the negative set. To avoid any implication that the issue has been fully resolved, we have added an explicit statement in the **Discussion** section (highlighted in **blue**) that acknowledges this limitation, clarifies that our checks only indicate non-fragility rather than a solution to the labeling problem, and suggests future directions to further mitigate this concern. We believe this addition makes the paper more balanced and credible.

--- Page 17, Discussion, lines 613-618

Comments4. I had a similar reaction to the comparison framework. The revised manuscript does a better job of explaining why the selected baselines were chosen, and I appreciate that the authors focused on conceptually relevant graph-based competitors rather than assembling an arbitrary benchmark. Even so, the comparison space remains relatively narrow. I do not think the paper needs a wholesale expansion of experiments at this stage, but I do think readers should be told more plainly that the study is benchmarking the proposed method primarily against GNN-based alternatives, not against the full landscape of possible driver-gene prediction strategies. That is a reasonable design choice, but it should be framed as such.

**Response 4:** We thank the reviewer for this helpful suggestion. To make the scope of our comparison clearer, we have added a brief statement at the beginning of the baseline description paragraph in Section 3.1 (highlighted in **blue**), explicitly stating that this study benchmarks M-GNN primarily against GNN-based competitors rather than attempting a comprehensive comparison across all driver gene prediction strategies. This addition complements the existing explanation for excluding non-GNN methods and frames our comparison choice transparently. We

believe this makes the evaluation scope more explicit and appropriately qualified.

--- Page 11, Section 3.1 Performance Evaluation of the M-GNN, lines 411-413

We sincerely appreciate your time and expertise in reviewing our work. We hope that the revised manuscript now meets your expectations. Please do not hesitate to contact us if any further clarification is needed.

Sincerely,

Wen Zhu (on behalf of all authors)

School of Mathematics and Statistics, Hainan Normal University

Email: syzhuwen@163.com