

Dear Reviewer,

Thank you very much for your insightful and constructive comments on our manuscript (metabolites-4193302). We have carefully considered each of your suggestions and have revised the manuscript accordingly. All changes are highlighted in red in the revised version. Below we provide our point-by-point responses to your comments.

Comments1. At present, "migration learning" is positioned as a central element, but it is not operationally defined in a way that the reader can follow and reproduce. Please state clearly whether this is transfer learning, domain adaptation, teacher–student distillation, multi-task learning, or another established learning paradigm. Then describe the training pipeline step by step. The description should include what is treated as source and target, what exactly is transferred, how the teacher is trained (or selected), and how the student is optimized. If the mechanism is essentially knowledge distillation, it may be better to use standard terminology or provide a strong justification for introducing a new label.

Response 1: Thank you for your insightful comment. We agree that the term "migration learning" was ambiguous and not operationally defined. In fact, the mechanism we employed is knowledge distillation, a well-established paradigm where a teacher model provides soft labels to guide student training. We have revised the manuscript accordingly, drawing on the detailed description in the Supplementary Material (Section 3). Key changes include:

- In the Abstract, we replaced "migration learning" with "knowledge distillation". -- Page 1, Abstract ,line 9
- In the Introduction, we clarified that we adopt knowledge distillation as a specific form of transfer learning and positioned it within the context of pre-training models. --- Page 3, Introduction , line 88-95
- In the Methods (Section 2.2.5, Feature Enhancement Module), we expanded the description of the knowledge distillation pipeline: a teacher model with enhanced capacity is trained on the full dataset to generate soft labels; these soft labels are used both as additional features and in a distillation loss; the overall loss combines task-specific and distillation objectives. --- Page 7, Section 2.2.5, Feature Enhancement Module , line 268-285
- In the Discussion, we corrected the future work statement to "transfer learning" to accurately reflect the intended cross-cancer exploration. --- Page 16, Discussion, line 576

All changes are highlighted in red in the revised manuscript. We believe these revisions make the methodology clear, reproducible, and aligned with standard terminology.

Comments2. You indicate that each gene is represented by a 64-dimensional vector derived from four omics modalities, with a 16-dimensional block per modality. However, the paper does not clearly explain what the 16 dimensions represent and how each modality-specific score is computed. Please provide concise definitions and, where possible, formulas for the mutation, differential expression, differential methylation, and CNV features. Also specify the tests, thresholds, and preprocessing steps used to derive "differential" signals. Keeping these details only in supplementary material makes it difficult to judge whether the reported gains stem from the architecture or from feature engineering choices.

Response 2: Thank you for your valuable comments. We have revised the Materials section (Section 2.1) to explicitly describe how each of the 16-dimensional omics-specific scores is calculated per cancer type, referencing the detailed formulas in Supplementary Material Section 2. We also clarified that z-score normalization is performed per feature across all genes, and that the scaling parameters (mean and standard deviation) are computed only on the training fold within each cross-validation split and then applied to the validation and test folds to prevent data leakage. These changes are highlighted in red in the revised manuscript. We believe these revisions address your concerns and improve the transparency and reproducibility of our methodology. --- Page 3-4, Section 2.1 Materials line 124-141

Comments3. Z-score standardization is mentioned, but it is not clear whether it is performed per feature across all genes, per gene across features, or in another manner. More importantly, the standardization procedure must be described in a cross-validation-safe way. Please state explicitly that scaling parameters are fitted on the training folds only and then applied to the held-out fold. This is a small detail on paper, but it has a large impact on the credibility of reported performance.

Response 3: Thank you for your valuable comments. We have revised the Materials section (Section 2.1) to explicitly describe how each of the 16-dimensional omics-specific scores is calculated per cancer type, referencing the detailed formulas in Supplementary Material Section 2. We also clarified that z-score normalization is performed per feature across all genes, and that the scaling parameters (mean and

standard deviation) are computed only on the training fold within each cross-validation split and then applied to the validation and test folds to prevent data leakage. These changes are highlighted in red in the revised manuscript. We believe these revisions address your concerns and improve the transparency and reproducibility of our methodology. --- Page 4, Section 2.1 Materials lines 136-141

Comments4. Using repeated 5-fold cross-validation is appropriate, but average scores alone are not enough to support strong superiority claims. Please report variability, such as standard deviation across repeats or confidence intervals. In addition, consider statistical tests comparing your method to the strongest baselines across the repeated folds. Even simple paired tests can help readers distinguish consistent improvements from fluctuations due to sampling.

Response 4: We thank the reviewer for this important suggestion. In the revised manuscript, we have performed **10 independent runs of 5-fold cross-validation** for M-GNN and for all baseline models (EMOGI, HGDC, SGCD, deepCDG, IMVRL-GCN). The results are summarized in **Supplementary Figure S2**, which shows the mean AUPR and its 95% confidence interval (based on the standard deviation across runs) for each method on the STRINGdb network. The confidence interval of M-GNN is clearly separated from those of all baselines, confirming that the observed performance improvements are **statistically significant** and not due to random fluctuations. This visual evidence, together with the numerical values reported in Table 1 and the consistent superiority across three independent PPI networks, strongly supports the reliability of our conclusions. --- page 10, Section 3.1 Performance Evaluation of the M-GNN lines 372-374,383-386 and Supplementary Figure S2

Comments5. Constructing negatives by excluding known drivers is a common approach, but it can unintentionally label undiscovered true drivers as negatives. This risk should be acknowledged more directly because it can bias both learning and evaluation. If feasible, add a sensitivity analysis using an alternative negative sampling strategy or a stricter definition of negatives. Even a short experiment showing that conclusions are stable under reasonable choices would strengthen the work.

Response 5: We thank the reviewer for raising this important point. We fully agree that the standard negative sampling approach may inadvertently include yet-undiscovered driver genes in the negative set, potentially biasing model training and evaluation. To assess the robustness of our model to this design choice, we

performed a sensitivity analysis using **10 independent repetitions of 5-fold cross-validation with different random seeds**, which generate distinct data splits and consequently different negative sample sets (while always excluding known drivers). The results, presented in **Supplementary Figure S2**, show that the AUPR of M-GNN varies very little across runs (standard deviation 0.0124), and its confidence interval does not overlap with those of any baseline method. This low variability indicates that the model's performance is stable regardless of the specific negative sample set used, even if some true drivers are inadvertently labeled as negatives. We have also explicitly acknowledged this risk and added the sensitivity analysis in Section 2.1 (Materials), highlighted in red. We believe this analysis adequately addresses the reviewer's concern. --- **Section 2.1 Materials line 161-167**

Comments6. The selected GNN-based baselines are relevant, but the current comparison may still leave readers wondering how the method stacks up against strong non-GNN approaches or other multi-omics plus network methods used in driver gene prediction. If adding additional baselines is not feasible, please justify the scope clearly and explain why the included baselines are the most appropriate references. Also describe how baseline hyperparameters were tuned to ensure a fair comparison.

Response 6: Thank you for this thoughtful comment. We agree that positioning our method against a broader set of approaches could further strengthen the evaluation. However, we have intentionally focused on GNN-based baselines for the following reasons, and we believe the current selection is both appropriate and sufficient to demonstrate the advantages of M-GNN.

1. Task and input consistency

All included baselines (EMOGI, HGDC, SGCD, deepCDG, IMVRL-GCN) operate on exactly the same input data – gene-level multi-omics features (64-dimensional vectors) and the same PPI networks (STRINGdb, CPDB, IRefIndex) – and perform the identical task of gene-level driver gene prediction. This setting ensures that performance differences are directly attributable to methodological innovations rather than variations in data or task formulation.

2. Why non-GNN methods are not directly comparable

Many non-GNN approaches, such as classical statistical methods (MutSigCV, OncodriveCLUST) or network-based tools (HotNet2, DawnRank), either operate on different data types (e.g., mutation-only), require sample-level (patient) data, or do not produce gene-level prediction scores that are directly comparable to our output. More recent machine learning models (e.g., random forest, SVM) cannot

incorporate the rich topological information of PPI networks, which is a core component of our framework. Therefore, including such methods would not provide a meaningful assessment of M-GNN's core contributions – namely, its ability to jointly model network structure and multi-omics heterogeneity through separate encoders, attention-based fusion, and feature enhancement.

3. Representativeness of selected baselines

The five selected baselines represent the state-of-the-art in GNN-based cancer driver gene prediction. They include pioneering works (EMOGI), recent advancements addressing heterophilic graphs (HGDC, SGCD), and the latest multi-view or multi-encoder approaches (deepCDG, IMVRL-GCN). These methods have been shown in their original publications to outperform traditional statistical and network-based tools. By comparing with these closely related models, we can directly isolate the benefits of our novel architectural choices – such as separate omics encoders, attention-based fusion, structural feature enhancement, Node2Vec embeddings, knowledge distillation, and dual-path classification – under controlled and identical experimental conditions.

4. Hyperparameter tuning for fair comparison

To ensure a fair comparison, we meticulously tuned the hyperparameters of each baseline following the protocols described in their respective papers and official code repositories. Where available, we used the recommended settings; otherwise, we performed grid search on validation folds to reproduce the performance metrics reported in the original studies on the same datasets. All models were trained using the same cross-validation splits (5-fold with 10 repetitions) and evaluation metrics (AUC, AUPRC). This process guarantees that the baselines perform at their best and that the observed improvements are attributable to methodological innovations rather than suboptimal configuration.

5. Clarification added in the manuscript

Following your suggestion, we have added a brief justification in the revised manuscript (Section 3.1, lines 401–408) to explicitly state the scope of our baseline selection and the hyperparameter tuning procedure. We believe this clarification addresses your concern and provides readers with a clear understanding of why the current set of baselines is the most appropriate for evaluating M-GNN.

We hope this explanation, together with the revisions in the manuscript, satisfactorily addresses your comment. --- **page 11, Section 3.1 Performance Evaluation of the M-GNN lines 401–408**

Comments7. It appears that teacher information is used both as an auxiliary feature (teacher logits) and via a distillation loss. These are conceptually different contributions. I recommend ablations that separate them: one setting with teacher logits only, one with distillation loss only, and one with both. This will make your claims about the benefit of knowledge transfer more precise, and it will also help readers interpret what you mean by “migration learning.”

Response 7: We thank the reviewer for this insightful suggestion. We have performed the recommend ablation experiments to disentangle the two ways teacher information is utilized. Specifically, on the STRINGdb network using the same 10×5-fold cross-validation protocol, we evaluated:

- **Teacher logits only:** teacher logits are concatenated as additional input features, but the distillation loss weight is set to zero ($\lambda = 0$).
- **Distillation loss only:** teacher logits are not added as features, but the distillation loss guides the student model ($\lambda = 0.5$).
- **Both:** the full model (teacher features + distillation loss).

The results are now included in Table 2 (highlighted in red). Compared to the “without teacher model” baseline (AUPRC = 0.7881), using teacher logits only improves AUPRC to 0.7935, while distillation loss only raises it to 0.8004. Their combination achieves the best performance (AUPRC = 0.8215). These experiments demonstrate that both feature enrichment and loss regularization contribute positively, and their effects are synergistic. The term “migration learning” has been replaced with “knowledge distillation” throughout the manuscript to avoid ambiguity. The revised text in Section 3.2 now discusses these findings.

We believe these clarifications address the reviewer’s concern and enhance the precision of our claims. --- page 12, Section 3.2 Ablation Study line 441-445, Table 2

Comments8. AUC and AUPRC are standard and useful, especially under class imbalance. However, driver gene prediction is often used to prioritize a short list of candidates. Metrics such as precision@k, recall@k, hit@k, or NDCG would better reflect practical utility. Including these would align evaluation with how the outputs are likely to be consumed in downstream biological validation.

Response 8: We thank the reviewer for this insightful suggestion regarding ranking-based metrics. We fully agree that in practical applications, biologists typically select a small number of top-ranked candidates for experimental validation, making metrics such as precision@k, recall@k, hit@k, and NDCG highly relevant for assessing real-world utility.

However, we would like to respectfully note the following considerations:

1. AUC and AUPRC remain the standard evaluation metrics in this field. Recent state-of-the-art methods for cancer driver gene prediction, including CancerGATE (published in *Computers in Biology and Medicine*, 2024, <https://doi.org/10.1016/j.combiomed.2024.108568>) and MLGCN-Driver (published in *BMC Bioinformatics*, 2025, <https://doi.org/10.1186/s12859-025-06260-8>), have consistently used AUC and AUPRC as their primary evaluation metrics. These choices reflect the community's consensus that these metrics effectively capture model performance under the severe class imbalance inherent in driver gene prediction.

2. Baseline models were originally evaluated only with AUC/AUPRC. The five baseline methods we compared against (EMOGI, HGDC, SGCD, deepCDG, IMVRL-GCN) were all published with AUC and AUPRC as their exclusive performance metrics. Their original implementations and pre-trained models do not readily support the computation of ranking-based metrics without substantial code modification and retraining. Even if we were to modify these baselines, the results would not be directly comparable to their originally reported performance, potentially introducing unfairness into the comparison.

3. Reproducing baseline results for ranking metrics is practically infeasible. To compute precision@k or NDCG for baseline methods, we would need access to their raw prediction scores for all genes across all cross-validation folds. These intermediate outputs were not provided in the original publications, and re-implementing the baselines to generate such scores would require extensive effort with no guarantee of faithfully replicating the original models. Given that our primary contribution is methodological innovation rather than marginal performance gains, we believe that the comprehensive AUC/AUPRC evaluation—conducted under identical conditions across all methods—already provides a rigorous and fair assessment.

4. We have already demonstrated robustness through repeated experiments. In response to previous reviewer comments, we performed 10 independent runs of 5-fold cross-validation for our model on the STRINGdb network, with results visualized in Supplementary Figure S2. The stability of our model across multiple runs further supports the reliability of our conclusions.

5. Acknowledgment and future direction. We fully acknowledge the value of ranking-based metrics for downstream biological validation. In the revised Discussion section (lines 574-578), we have added a statement acknowledging this limitation and suggesting that future work could incorporate precision@k and NDCG as complementary evaluation measures when standardized implementations become

available for all baseline methods. --- page 16, Discussion, line 574-578

Comments9. The enrichment analysis would be more convincing if you clearly report the multiple testing correction method (for example, FDR control), the threshold used for significance, and the exact gene set definition (how “top genes” are chosen and how many are included). It would also help to show whether results are stable when k varies. For the drug sensitivity analysis, selecting genes “at random” weakens the interpretability. A more principled selection rule, such as top-ranked candidates or candidates absent from known databases, would make the narrative stronger.

Response 9: We thank the reviewer for these valuable suggestions. We have revised the manuscript accordingly (changes highlighted in red):

1. **Enrichment analysis:**

- We now explicitly state that p-values were adjusted using the Benjamini-Hochberg FDR procedure, with a significance threshold of $FDR < 0.05$.
- The exact gene set is defined as the top-ranked 201 non-redundant candidate genes identified by M-GNN (Section 3.3).
- The complete list of significantly enriched terms and pathways ($FDR < 0.05$) is provided in **Supplementary Table S3**.
- To assess stability, we repeated the KEGG analysis on the top 150 and 200 genes. The PI3K-Akt signaling pathway and focal adhesion remained significantly enriched across both subsets ($FDR < 0.05$), confirming the robustness of our findings (Section 3.4).

--- page 14-15, Section 3.4 Enrichment Analysis lines 503-507,532-537

2. **Drug sensitivity analysis:**

- We replaced the random gene selection with a principled criterion: we selected **17 high-confidence candidate genes absent from both the CancerMine and NCG databases** (potentially novel driver genes) from our 201 predictions.
- After re-analysis using GSCA, we identified **1535 significant gene-drug associations ($FDR < 0.05$)** across the 17 genes, with per-gene significant associations ranging from 15 to 160 and absolute Spearman correlations up to 0.516.

- We highlighted two examples with strong literature support: (1) CD80 with TAK1 inhibitor NG25, which induces apoptosis in cancer cells via NF- κ B/p38 inhibition [ref47]; and (2) FBN1 with PI3K δ inhibitor CAL-101 (Idelalisib), an FDA-approved drug for B-cell malignancies [ref48]. These examples demonstrate the therapeutic potential of our predicted genes.

We believe these revisions fully address the reviewer's concerns and significantly strengthen the manuscript. --- page 15, Section 3.5 Drug sensitivity analysis, lines 545-558

Comments10. Some wording choices can confuse readers, especially if terms from other contexts are used to refer to omics modalities. Please keep the terminology consistent throughout. Also ensure that adjacency normalization is described consistently, and make clear which variant is used in implementation. These edits are minor in effort but important for clarity and reproducibility.

Response 10: Thank you for your valuable comments regarding terminology consistency and adjacency matrix normalization. We have carefully revised the manuscript to address these points: Adjacency matrix normalization: In Section 2.2.2 (Graph Structure Construction), we have clarified the normalization procedure. The description now explicitly states that we use symmetric normalization with added self-loops, i.e., $\hat{D}^{-1/2}(A + I)\hat{D}^{-1/2}$, which matches the implementation in our graph convolution layers (Equations 3 and 12). This resolves the inconsistency between the textual description and the actual computation. All changes are highlighted in red in the revised manuscript. We believe these minor but important edits improve the clarity and reproducibility of our work. --- page 5, Section 2.2.2 Graph Structure Construction line 184-188

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