nature portfolio

Peer Review File

Deep-learning-enabled multi-omics analyses for prediction of future metastasis in cancer

Corresponding Author: Dr Qin Ma

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)
The authors present a heterogeneous graph-based model, EmitGCL, designed to predict future metastasis and identify corresponding biomarkers by integrating single-cell RNA-seq data from both primary and metastatic tumor samples. The model demonstrates promising performance in comparative analyses. However, I have several major concerns regarding the methodological clarity and biological validation of the findings.
Major:
1) The model relies on graph-based cell embeddings, which can be sensitive to random initialization and runtime settings (e.g., seed, batch order). How do the authors ensure the robustness and reproducibility of their predictions under different seeds or conditions? Has the model been run multiple times to assess variability?
2) The authors state that occult metastatic cells and metastatic precursor cells are predicted to belong to the same cluster based on the EmitGCL framework. However, in Figure 3a, these two cell types appear to occupy different clusters, which contradicts the earlier claim. The authors should clarify it.
3) How are batch effects handled during the integration of multi-patient or multi-batch scRNA-seq data? Was any standard correction method applied prior to graph construction? The effectiveness of the model may be compromised if batch-specific biases dominate the clustering or graph topology. Could the authors clarify how the cell-gene graph is constructed?
4) Tumor cells are highly heterogeneous, and boundary or intermediate-state tumor cells tend to exhibit higher metastatic potential. Identifying cells with high CNV scores, stemness, and metastasis is not surprising. Hence, the bioinformatic evidence provided is not sufficient to confirm the accuracy. Given the increasing availability of spatial transcriptomics (ST) data, can EmitGCL be extended to integrate multi-slice ST datasets? If so, testing whether the model's predictions align with spatially resolved metastatic progression could significantly strengthen the study's impact and biological credibility.
5) While the framework seems to function as a cell embedding approach for integrating large-scale scRNA-seq datasets, the manuscript lacks a presentation of the overall clustering result, which is essential to evaluate the integration quality and biological interpretability.
6) The benchmark comparison appears to be biased, as only EmitGCL incorporates the tissue-of-origin information of the cells, whereas the other methods do not.
7) I am unable to access the GitHub code.

Minor:
The manuscript contains inconsistent tense usage, which may affect readability. Please revise the text to ensure grammatical consistency throughout.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
This study applies a heterogeneous graph method, previously developed by the authors, to represent cells and genes in a graph and integrate single-cell RNA-seq (scRNA-seq) data from multiple studies, including primary and metastatic samples. The primary aim is to create a robust framework capable of handling the sparsity and complexity inherent to scRNA-seq

data, with the goal of identifying precursor metastatic cells and their underlying biomarkers for metastasis prediction. The method and benchmarking are quite complex, the authors make a good effort with some helpful schematics, but several technical details still require clearer explanation. Most importantly, several claims seem overstated given the evidence presented, for example the statement in the abstract that “we present EmitGCL, a deep-learning framework that accurately predicts future metastasis and its corresponding biomarkers”. These conclusions appear limited to the datasets analyzed, and would be more accurate if accompanied by specific performance metrics.

Main comments:

1. The authors define the term “occult metastatic cells” in the introduction, but they do not clearly specify how scRNA-seq cells are labeled in practice. It appears that the labeling is derived from the metastatic annotations in the original study, but a precise description of the criteria and procedure used is essential.

2. The validation of the FPR using primary lung cancer and lymph node sites without metastasis is interesting. It would be valuable to further assess this using scRNA-seq datasets from healthy tissues, for example, leveraging the Human Cell Atlas to evaluate a range of cell types, from fast- to slow-growing cells, including fibroblasts.

3. In Figure 2h, it is somewhat puzzling that all the other methods exhibit a FPR of 1. This implies they label every negative sample as positive, suggesting that these models are fundamentally ill-suited for the task and therefore may not serve as appropriate benchmarks.

4. The authors should assess and comment on the potential contribution of fibroblast cells to the EMT signature. Several of the original dataset publications report fibroblast-associated signatures.

5. YY1 knockout is pan-lethal across +1,000 cancer cell lines (https://depmap.org/portal/gene/YY1?tab=overview). Have the authors considered that both the wound healing and transwell migration assays could be confounded by this? In other words, their observations might simply be explained by cell death.

6. There is a general lack of labels across both the main and supplementary figures, particularly of axis labels. Additionally, it appears that post-processing of the figures has introduced inconsistencies in some panels. For example, in Figure 4e, the second bar displays a higher value than the first despite being visibly smaller, which suggests an issue introduced during figure editing. This is evident when comparing with the original plot, which is one of those that can be found in the source code provided in the accompanying Jupyter notebooks. In my view, presenting the original versions would likely offer a better representation of the results.

7. More details on the assembly and integration of the datasets would be important. Was any normalization performed? While batch effect correction using ComBat is mentioned in the methods, wouldn't ComBat-seq be more appropriate for count data? It would also be helpful to clarify the batch and group parameters used in the correction.

8. Several parts of the Methods section is very similar to the original publication from the authors (PMID: 38184630), thus instead of duplicating text it would be more appropriate to reference the original formulation and mention the differences.

9. Multiple gene signatures of EMT exist, particularly those that separate genes that should be up/down regulated. Given the central role of EMT enrichment in this analysis, the authors should show that this can be recapitulated with other EMT signatures.

10. The GitHub link is currently inaccessible, but the code is provided to the reviewers as a zip file. Will the authors make the repository publicly available upon publication? The code would benefit from improved organization; for example, Biomarker_inference.ipynb contains execution errors, and code comments are sparse and often not in English.

11. Page 5: Improve sentence: “GiniClust17, was shown as the second-best tool for rare cell identification using clustering methods, while CellSIUS27 was the second-best tool for rare cell identification using classification methods.” for example “GiniClust17 and CellSIUS27 were the second-best tool for rare cell identification using clustering and classification methods, respectively.”

(Remarks on code availability)
A README file in the root of the code repository is missing. As mentioned in the main comments: The code would benefit from improved organization; for example, Biomarker_inference.ipynb contains execution errors, and code comments are sparse and often not in English.

Reviewer #3

(Remarks to the Author)
The manuscript presents a new graph contrastive learning approach, call EmitGCL, to detecting occult metastatic cells and metastatic precursor cells from scRNAseq data. Based on EmitGCL, the authors reported the identification and validation of two predictive biomarkers for future breast cancer metastasis. Furthermore, they demonstrated YY1 TF as a key driver of breast cancer metastasis, which was validated in-silico and by CRISPR-based migration assays. Overall, the manuscript is well-written, the deep-learning approach is reasonable and scalable, and the findings were carefully validated.

The above being said, I have the following major concerns.

(1) Insufficient and unclear benchmarking results.

According to the description of model inputs in the Methods section, EmitGCL is trained by a pair of scRNAseq datasets, one from the primary site, and the other from a metastatic site. Thus, to gather the baseline performance of the approach in terms of false positive rates and true positive rates, it would be natural to use unmatched/mismatched pairs, in addition to matched pairs, as inputs to EmitGCL, such as a pair of primary site datasets (with one mislabeled as metastatic), a pair of metastatic site datasets (with one mislabeled as primary), and a pair of mismatched primary-metastatic datasets, all from two different patients. Personally, I would also gather some additional scRNAseq datasets collected from healthy samples to assist the mismatched pair generation process.

Moreover, in the reported results, all cells in the matched dataset pair were used in training EmitGCL. This naturally raises the concern on whether there is overfitting. One approach to make sure this is not the case is to leave out 20% of cells (or even 30% as one aims to identify rare cell types here) in the training process. Given the subgraph sampling approach the authors took when training the model, this is very much doable. Such training-test sample splitting is a common practice for ML or DL approaches, I anticipate that they can also be implemented on other methods in comparison. Then, one can perform inference on the left-out test cells and calculate various performance metrics on them across all methods considered in the benchmarking study.

(2) Model training.

In equation (16), the authors proposed to train the weights by using a multi-task loss function that has four terms. While each of the four terms makes sense, I am concerned that they are just summed up without proper weights. In particular, it is not self-evident that the relative magnitudes of the four terms are balanced. For example, it is possible that two terms are much bigger than the other two and so the loss becomes effectively the sum of the two dominating terms. The authors should report the sizes of these terms across the datasets that they have fitted the model on. If they are not balanced, then a proper weighing scheme should be in place, ideally with recommended default weights.

In addition to the foregoing issue, there are ambiguities in the current description of the bipartite graph subsampling scheme. In the paragraph after equation (4), it is stated that “The number of reserved genes is denoted as N_g.” However, neither “reserved genes” nor “N_g” is defined. Moreover, it is stated that “In each site, the subgraph incorporates 30 randomly selected cells …”. I am assuming it means that each subsampled subgraph has 60 randomly selected cell nodes (30 from either site) plus their chosen neighbor gene nodes. If the numbers of cells are very different in primary vs metastatic sites, what are the ramifications of this balanced sampling scheme?

(3) Batch effect correction.

The current work used ComBat as the batch correction method. However, according to the following paper “A benchmark of batch-effect correction methods for single-cell RNA sequencing data”[https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1850-9], ComBat is not as good as Harmony, LIGER, and Seurat v3. Since the EmitGCL model also take pairs of datasets as input, it makes sense to adopt one of the best batch effect correction methods available.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
The authors evaluated the predictive value of EmitGCL for metastasis. The authors indicated they identified HSP90AA1 and HSP90AB1 are a reliable biomarker to predict breast cancer metastasis. Although it is a interesting study, it has fundamental flaws.

1. All the data presented in this manuscript showed correlation of genes between detected metastasis and primary tumors. No patient in this study had occult metastasis which should be defined as metastasis but not detected.

2. There is no description of the 420 patients the authors used as validation cohorts.

3. The authors failed to show expression difference of HSP90AA1 and HSP90AB1 in breast tumors that developed metastasis vs that did not develop metastasis.

4. In figure 4. The authors showed survival difference but not metastatic rate, which is very odd.

5. To validate the biomarkers, the authors should obtain primary tumor samples from patients who did not have metastasis at the time of surgery. The authors should follow these patients to validate their biomarkers. Although this step requires IRB approval and involvement of clinical teams, it is the best way to validate the biomarkers. I do not see any convincing validation in this manuscript.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed most of my concerns.
However, regarding comment 1), I do not think that evaluating robustness solely by ARI and NMI between clusterings obtained under different initialization seeds and the reference clustering is sufficient. Because a key claim of EmitGCL is its ability to detect rare cell types and to identify biologically meaningful marker genes via the attention mechanism, the robustness of these specific outputs should also be quantified. It would be helpful if the authors could design more targeted metrics to assess (i) whether the same rare cell populations are consistently recovered across seeds and (ii) whether the marker genes highlighted by the attention weights are stable and biologically coherent across different runs.

Regarding comment 4), I appreciate the authors' detailed response and agree that a full spatial extension of EmitGCL may ultimately require a dedicated algorithm. However, I am still not fully convinced by the argument that the lack of multi-slice ST data precludes any spatial validation at this stage. In recent years, a growing number of multi-slice or multi-region ST datasets across several cancer types have become publicly available, and even single-slice ST data can already provide spatially resolved patterns of metastatic propensity, niche composition, and rare cell populations.

Given that a central claim of EmitGCL is the ability to detect rare metastatic precursors and biologically meaningful markers, an analysis on at least one existing ST dataset (e.g., assessing whether EmitGCL-derived risk scores or marker signatures align with spatially localized high-risk niches or early metastatic regions) would substantially strengthen the biological credibility of the method. Even a proof-of-concept application on publicly available ST data, without developing a fully new spatial framework, would be much more convincing than limiting this point to a discussion of future work.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have addressed most of my previous comments. However, I still find that some of the general claims are overstated given the results presented, and several figures are unnecessarily over-processed which introduce errors (e.g. Figure 4e). While I agree with the authors that applying their method to healthy cells is not the primary aim of the study, I still believe there is value in leveraging the large number of available scRNA-seq datasets from healthy cells as controls.

(Remarks on code availability)
The code is now available in github with some level of documentation. The tutorial available needs improvement (e.g. contains absolute paths which are not portable to other users).

Reviewer #3

(Remarks to the Author)
Two of the concerns I had in my previous round report have not been carefully addressed by this revision. Please see below for details.

(1) My concern: Insufficient and unclear benchmarking results. In my previous round report, I suggested the inclusion of unmatched/mismatched pairs for understanding the false positive behavior of EmitGCL. The authors dismissed this suggestion in their response letter and replied "Accordingly, mismatched or healthy pairs are not an appropriate baseline for estimating false- or true-positive rates in the intended use case." While I am not exactly sure what "the intended use case" is in authors' mind, I imagine one unavoidable use case is to detect whether a biopsy samples from a patient's primary tumor contains metastatic precursors, and another is to detect whether a potential metastatic site with no visible tumor actually contains occult metastatic cells. Both primary-healthy and primary-primary test cases are ideal for understanding the method's false positive rate when calling occult metastatic cells. Moreover, these cases are not artificial. It is possible that a suspected metastatic site is actually the primary site of a separate tumor or simply a malign one. Thus, I do not understand whey the authors dismissed my initial comment.

(2) My concern about model training objective in my previous round report "In equation (16), the authors proposed to train the weights by using a multi-task loss function that has four terms. While each of the four terms makes sense, I am concerned that they are just summed up without proper weights." In their response letter, the authors declared that "Each loss component is computed as the mean over its contributing elements within each mini- batch, which places the four terms on comparable scales." and thus regarded weighing the four terms as unnecessary. However, this claim was not supported by

the results reported in Supplementary Fig. 3d. For example, the range of the contrastive loss term is almost ten times that of the KL loss term for HN272. In addition, I cannot understand the reasoning behind the weight search grid reported in Supplementary Fig. 3e. The current weights are all ones, and the search grid in the figure set possible weights for some terms to be all below 0.2 and for some others to be all above 5, essentially making the search grid uninformative for balancing the four terms. I encourage the authors to take this issue more seriously, as there four terms are intrinsically different and there is no obvious reason for assigning them equal weights.

(Remarks on code availability)
N/A.

Reviewer #5

(Remarks to the Author)
This manuscript introduces EmitGCL, a deep-learning framework that uses knowledge-aware graph contrastive learning on paired single-cell omics data to predict future cancer metastasis. The model demonstrates strong performance in identifying rare metastatic cell types, proposes new biomarkers, and validates a key transcription factor via computational and experimental approaches. While the study addresses an important clinical problem and shows technical sophistication, major limitations in clinical translatability, model generalizability, and mechanistic depth prevent it from meeting the high standards of this journal. Substantial additional validation is required before publication can be recommended.
Issues that need to be addressed:
1. The study cohort is underpowered, with only 28 patients spanning six highly heterogeneous cancer types. This limited sample size is inadequate for reliably training a complex deep learning model, raising concerns about overfitting and generalizability. Substantial expansion of patient numbers within each cancer type is required to validate model robustness.
2. The clinical validation relies predominantly on a prediction case from a single patient's lymph node data. This isolated success may be coincidental and does not demonstrate general applicability. It is essential to include predictions from multiple independent patients to systematically assess the model's reproducibility and clinical utility.
3. For the proposed biomarkers HSP90AA1 and HSP90AB1, the study lacks functional mechanistic investigation. Correlation data alone are insufficient to establish their causal roles in metastasis. Functional experiments directly testing their impact on migration, invasion, or other metastatic phenotypes are necessary.
4. The observed phenotype cannot be strictly attributed to YY1 loss without demonstrating that re-introducing wild-type YY1 into knockout cells restores migratory capacity, thus establishing definitive causality.
5. To study the function of YY1 in mouse models, additional experiments should be conducted, such as comparing the lung colonization of YY1 overexpressing cells with control cells, to provide stronger support for its role as a therapeutic target.
6. All analyses are based on retrospective data with known outcomes, creating risk of "answer leakage." The model may simply identify features of already-metastatic patients rather than genuinely predicting future metastasis. Prospective validation on samples from treatment-naive, non-metastatic patients with long-term follow-up is crucial.

(Remarks on code availability)
Coding is not my area of expertise. Please prioritize opinions from experts in code-related fields.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I have no further comments.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors present a substantial amount of work to address the reviewers' comments. My outstanding points have been addressed, including code availability and documentation and a normal control. However, I still would like to mention figure presentation: although they have addressed in part my previous point on Figure 4e (now Figure 4g), this panel was simply the most obvious example of a more systematic underlying problem that runs across all figures. The issues highlighted in that panel are still reflected in several other panels (e.g., Figure 2e, f, g, h; Figure 3b, c; Figure 4c, g, j). As we have seen in previous rounds of revision, the post-processing of these panels introduces mistakes and omits genuinely informative aspects of the plots, such as axis ranges (apart from specific cases like UMAP representations, where this is appropriate). I would recommend that the authors work on this aspect prior to publication.

(Remarks on code availability)
The authors addressed my previous comments.

Reviewer #3

(Remarks to the Author)
The authors have successfully addressed my remaining concerns with the second revision. I do not have any additional concerns.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
EmitGCL is a deep-learning framework that utilizes knowledge-aware graph contrastive learning to predict future metastasis by identifying occult metastatic and precursor cells. It demonstrates superior performance, outperforming existing tools by 12.44% in true positive rate across various cancer types. The study identified HSP90AA1 and HSP90AB1 as predictive biomarkers and the transcription factor YY1 as a driver of breast cancer metastasis, supported by in vitro and in vivo validation and clinical survival analysis.

(Remarks on code availability)

Reviewer #1

The authors present a heterogeneous graph-based model, EmitGCL, designed to predict future metastasis and identify corresponding biomarkers by integrating single-cell RNA-seq data from both primary and metastatic tumor samples. The model demonstrates promising performance in comparative analyses. However, I have several major concerns regarding the methodological clarity and biological validation of the findings.

Major:

1) The model relies on graph-based cell embeddings, which can be sensitive to random initialization and runtime settings (e.g., seed, batch order). How do the authors ensure the robustness and reproducibility of their predictions under different seeds or conditions? Has the model been run multiple times to assess variability?

Response: Thank you for raising the two key aspects: (i) the model's sensitivity to random initialization, and (ii) the reproducibility of the reported results.

(i) To evaluate robustness to random initialization, we tested five datasets (one randomly selected per cancer type), each run with 20 different random seeds. The seed used in the main text served as the reference. For each alternative seed, we computed the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) relative to the reference clustering. Higher ARI/NMI values indicate greater similarity and see the following figure for more details. The means and standard deviations of ARI and NMI are shown as boxplots in the revised manuscript (Supplementary Fig. 3b, also shown below for convenience). The small standard deviations demonstrate that EmitGCL produces stable results across different initializations. We also put this new section into the Methods section "Statistics and reproducibility".

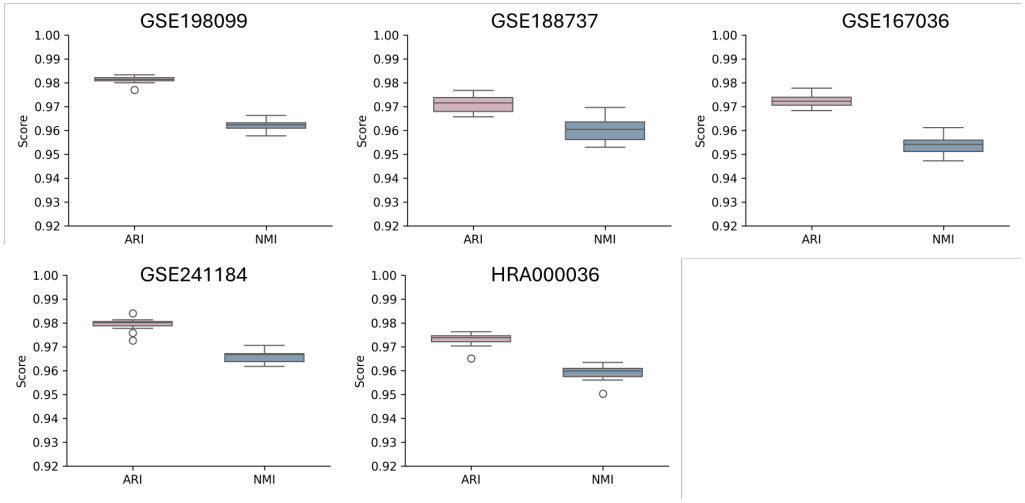


Figure legend: Robustness of EmitGCL with respect to parameter initialization evaluated across multiple cancer types.

(ii) We have also addressed the reproducibility of our results in the Methods section ("Statistics and Reproducibility"), and Supplementary Fig. 3 provides additional evidence supporting the reproducibility of our findings.

2) The authors state that occult metastatic cells and metastatic precursor cells are predicted to belong to the same cluster based on the EmitGCL framework. However, in Figure 3a, these two cell types appear to occupy different clusters, which contradicts the earlier claim. The authors should clarify it.

Response: We would like to clarify that Fig. 3a shows two separate datasets derived from distinct lymph nodes (LN-1 and LN-2) of the same patient, as indicated in the figure legend (Top: LN-1; Bottom: LN-2). Each lymph node was independently profiled and analyzed in parallel using the EmitGCL framework. Therefore, the occult metastatic cells and metastatic precursor cells are not shown as separate clusters within a single dataset. Rather, in the LN-1 dataset (top panel of Fig. 3a), Cluster 2 contains both occult metastatic cells and metastatic precursor cells, as further illustrated in Fig. 3b. Likewise, in the LN-2 dataset (bottom panel of Fig. 3a), Cluster 4 contains both cell types, which is again confirmed in Fig. 3b.

3) How are batch effects handled during the integration of multi-patient or multi-batch scRNA-seq data? Was any standard correction method applied prior to graph construction? The effectiveness of the model may be compromised if batch-specific biases dominate the clustering or graph topology. Could the authors clarify how the cell-gene graph is constructed?

Response: In our framework, conventional batch-effect correction across multi-patient scRNA-seq data is not directly relevant, because EmitGCL is trained on each patient individually using paired scRNA-seq datasets from the primary and metastatic sites. Thus, no standard integration or batch-effect removal step is applied before graph construction. To address patient heterogeneity and ensure marker robustness, we adopt a post-hoc consensus strategy. Specifically, candidate metastatic markers are prioritized based on their frequency across patients: genes consistently identified as metastatic markers in multiple patients are ranked higher and selected with greater priority than those identified in only a few patients. This procedure reduces the risk that patient-specific noise or batch-specific biases dominate the downstream analysis. Conceptually, this corresponds to a typical post-hoc strategy, as discussed in our recent Nature Biotechnology article (Qin Ma et al., Nature Biotechnology, 2025).

As described in the newly added Methods section (“Heterogeneous Bipartite Graph Construction”), we construct a cell–gene bipartite graph directly from raw count data, without prior normalization or correction. The graph contains two node types, genes and cells, and an edge is drawn between a gene and a cell only if that gene is expressed in that cell. This design preserves the original expression landscape within each sample pair, ensuring that biologically relevant variation is maintained.

4) Tumor cells are highly heterogeneous, and boundary or intermediate-state tumor cells tend to exhibit higher metastatic potential. Identifying cells with high CNV scores, stemness, and metastasis is not surprising. Hence, the bioinformatic evidence provided is not sufficient to confirm the accuracy. Given the increasing availability of spatial transcriptomics (ST) data, can EmitGCL be extended to integrate multi-slice ST datasets? If so, testing whether the model’s predictions align with spatially resolved metastatic progression could significantly strengthen the study’s impact and biological credibility.

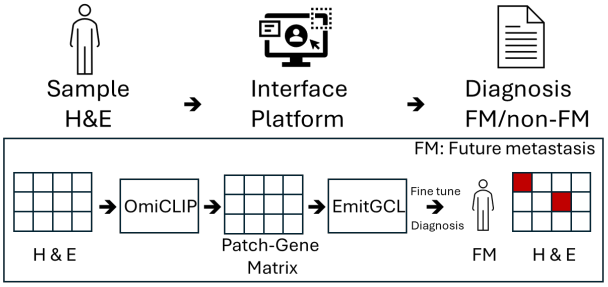
Response:

i) We fully agree that elevated CNV, stemness, and metastasis scores are necessary but not sufficient indicators of metastatic potential. Our analyses in Fig. 2e and Supplementary Fig. 2 were not intended as conclusive validation of metastatic cell identification, but rather to show that EmitGCL outperforms baseline methods in capturing tumor features commonly associated with metastasis. To provide more rigorous validation, we designed multi-level benchmarks:

- Patient-level validation: In Fig. 2g, EmitGCL shows superior predictive power. Baseline methods fail to identify occult metastases in several clinically confirmed metastatic cases, whereas EmitGCL succeeds.
- Non-metastatic cases: In Fig. 2h, we present two lung cancer patients who did not develop metastasis. EmitGCL correctly predicts no metastatic potential (FPR = 0), while all baseline methods falsely predict metastasis (FPR = 1).
- Longitudinal validation: In Fig. 3, we highlight a breast cancer patient whose two lymph nodes were initially diagnosed as non-metastatic but were later confirmed to harbor metastases. EmitGCL correctly predicted metastatic risk in both lymph nodes and the matched primary tumor sample, whereas baseline methods did not.

Together, these results demonstrate that EmitGCL achieves higher true positive rates (TPR) and lower false positive rates (FPR), and that our validation strategy extends beyond surface-level associations (e.g., CNV scores) to show consistency with clinical outcomes and biological plausibility.

ii) We agree that extending EmitGCL to spatially resolved datasets is an exciting future direction that could strengthen both impact and biological credibility. At present, however, there are practical challenges. Our current model focuses on detecting rare metastatic precursors at the early dissemination stage, which may be difficult to capture in ST data due to limited resolution and the absence of true 3D spatial context. Multi-slice ST profiling from the same patient is rarely available in clinical practice, especially before metastasis occurs, limiting both dataset availability and standardization. For these reasons, we believe that a dedicated algorithm will be needed to adapt EmitGCL to spatial data, other than a simple extension of the current algorithm. Nonetheless, we agree that this is a critical next step and have added this point to the Discussion section of the revised manuscript. In fact, inspired by this suggestion, our group has already initiated a follow-up project to tackle this problem. Specifically, we are developing a new framework that leverages foundation models trained on large-scale ST and H&E datasets: H&E-stained slides serve as input, ensuring clinical feasibility. Bulk RNA-seq datasets are integrated as molecular references to enhance interpretability. The framework builds on OmiCLIP, a spatial foundation model from Dr. Guangyu Wang's lab (a co-author of EmitGCL), enabling robust gene expression inference from histopathology. A conceptual overview of this ongoing project is presented in the following figure, which illustrates how the proposed spatial version of EmitGCL would integrate histopathology with molecular data to predict future metastasis and highlight high-risk regions within H&E slides.



5) While the framework seems to function as a cell embedding approach for integrating large-scale scRNA-seq datasets, the manuscript lacks a presentation of the overall clustering result, which is essential to evaluate the integration quality and biological interpretability.

Response: As noted in our response to Comment 4 (Reviewer #1), our framework does not integrate data across different samples or patients. Instead, the model is trained independently on each sample or on matched primary–metastatic pairs. Therefore, standard evaluations of integration quality, such as overall clustering across multiple datasets or assessment of batch effect removal, are not applicable in our setting.

6) The benchmark comparison appears to be biased, as only EmitGCL incorporates the tissue-of-origin information of the cells, whereas the other methods do not.

Response: We agree that EmitGCL is the only model in our comparison that incorporates tissue-of-origin information. This is intentional and represents a key innovation of our framework. By integrating biologically relevant prior knowledge and applying contrastive learning, EmitGCL is specifically designed to model the trajectory of cellular dissemination and to predict future metastatic progression—a task that current general-purpose methods are not designed to address. To ensure a fair and informative evaluation, we benchmarked EmitGCL against several widely used approaches for clustering and rare cell type identification. Although these methods do not leverage tissue-of-origin information and are not tailored to metastasis prediction, the comparison highlights differences in biological interpretability and underscores the limitations of existing tools in this context. In turn, their lack of specialized design emphasizes the necessity of problem-specific frameworks such as EmitGCL to achieve accurate and clinically meaningful predictions.

In addition, we emphasize that our in-house tool, MarsGT, which does not incorporate any organ or tissue-of-origin information and is designed solely for rare cell type identification, was also included as a comparator. As shown in Fig. 2f–h, MarsGT performs better than other general-purpose methods but remains inferior to EmitGCL. This performance gradient further underscores the importance of integrating organ-level prior knowledge, EmitGCL's key innovation, for achieving accurate metastasis prediction.

7) I am unable to access the GitHub code.

Response: Thank you for pointing this out. As noted in our initial submission, the code was not made publicly available on GitHub at that time due to an ongoing patent application. However, a code.zip file was provided through the submission system to ensure accessibility for reviewers.

In the current revision, we have updated the GitHub repository and released the code under a non-commercial license, which is now accessible. Please feel free to visit the updated repository (https://github.com/mtduan/emitgcl).

Minor:

The manuscript contains inconsistent tense usage, which may affect readability. Please revise the text to ensure grammatical consistency throughout.

Response: We have made the necessary revisions to ensure consistent tense usage throughout the manuscript.

Reviewer #2

This study applies a heterogeneous graph method, previously developed by the authors, to represent cells and genes in a graph and integrate single-cell RNA-seq (scRNA-seq) data from multiple studies, including primary and metastatic samples. The primary aim is to create a robust framework capable of handling the sparsity and complexity inherent to scRNA-seq data, with the goal of identifying precursor metastatic cells and their underlying biomarkers for metastasis prediction. The method and benchmarking are quite complex; the authors make a good effort with some helpful schematics, but several technical details still require clearer explanation. Most importantly, several claims seem overstated given the evidence presented, for example, the statement in the abstract that “we present EmitGCL, a deep-learning framework that accurately predicts future metastasis and its corresponding biomarkers”. These conclusions appear limited to the datasets analyzed and would be more accurate if accompanied by specific performance metrics.

Main comments:

1. The authors define the term “occult metastatic cells” in the introduction, but they do not clearly specify how scRNA-seq cells are labeled in practice. It appears that the labeling is derived from the metastatic annotations in the original study, but a precise description of the criteria and procedure used is essential.

Response: Thank you for the helpful comments and suggestions. In our study, the term “occult metastatic cells” refers to tumor cells that have recently disseminated to a new organ and appear at the metastatic site as a rare population (Harper et al., *Nature*, 2016, Celià-Terrassa & Kang, *Genes Dev*, 2016). Importantly, this is not a label annotated in the original studies but rather a population identified by our algorithm based on specific criteria. In practice, we define occult metastatic cells as those that (1) are classified as tumor cells at the metastatic site, and (2) constitute less than 3% of the total cell population in that site, reflecting their rarity. This definition is intended to capture early or rare metastatic events that may not yet have formed overt lesions. A detailed description of the criteria and identification procedure is provided in the Methods section, “EmitGCL identified occult metastatic and precursor cells.”

2. The validation of the FPR using primary lung cancer and lymph node sites without metastasis is interesting. It would be valuable to further assess this using scRNA-seq datasets from healthy

tissues, for example, leveraging the Human Cell Atlas to evaluate a range of cell types, from fast- to slow-growing cells, including fibroblasts.

Response: Our framework is specifically designed to model and predict tumor cell dissemination. For this reason, we focus exclusively on malignant cells and do not apply predictions or trajectory analysis to non-malignant cell types such as fibroblasts. Nevertheless, we agree that assessing model specificity in the context of healthy tissue references (e.g., Human Cell Atlas) could be an interesting direction for future validation, and we have noted this point in the revised Discussion.

3. In Figure 2h, it is somewhat puzzling that all the other methods exhibit an FPR of 1. This implies they label every negative sample as positive, suggesting that these models are fundamentally ill-suited for the task and therefore may not serve as appropriate benchmarks.

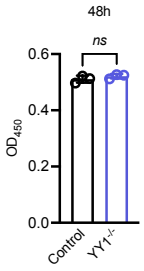
Response: We would like to clarify that the false positive rate (FPR) in our study is defined at the patient level (Methods, “False Positive Rate”). An FPR of 1 indicates that a patient who did not develop metastasis was incorrectly predicted to have future metastatic potential (i.e., predicted to contain occult metastatic cells). In Fig. 2h, we evaluated two lung cancer patients who did not develop metastasis. EmitGCL correctly predicted both patients as non-metastatic, resulting in an FPR of 0. In contrast, all baseline methods predicted the presence of occult metastatic cells in both patients, yielding an FPR of 1. This highlights that existing methods lack specificity in this setting, whereas EmitGCL avoids such false positives.

4. The authors should assess and comment on the potential contribution of fibroblast cells to the EMT signature. Several of the original dataset publications report fibroblast-associated signatures.

Response: We would like to clarify that all EMT-related analyses in our study were conducted exclusively on tumor cells. Fibroblast cells were not included in the calculation of EMT scores or in the benchmarking of EMT-associated signatures. Thus, the fibroblast-associated EMT signals reported in the original datasets do not influence our results or interpretation. Our benchmarking focuses solely on tumor-intrinsic EMT programs, and we have added this clarification to the revised manuscript.

5. YY1 knockout is pan-lethal across +1,000 cancer cell lines (https://depmap.org/portal/gene/YY1?tab=overview). Have the authors considered that both the wound healing and transwell migration assays could be confounded by this? In other words, their observations might simply be explained by cell death.

Response: We thank the reviewer for this important point. To exclude a viability confounder, we measured cell viability in the exact time window used for our wound-healing and Transwell assays (48 h) using a CCK-8 assay. YY1 knockout did not reduce viability compared with control in EO771 cells (n =3; mean ± SEM A450:0.521 ± 0.005 vs. 0.511 ± 0.007; two-sided unpaired t-test, p =0.2871). These data indicate that the genetic deletion of YY1 has no impact on the survival of breast cancer cell lines (Supplementary Fig. 9a); thus, the YY1-mediated migration is not attributable to cell death in our experiments. The experiment workflow had been added in the Methods section “Cell Viability Assay”.



Supplementary Fig. 9a. Cell viability was measured using a CCK-8 assay at 48 h in control and YY1 knockout EO771 cells

6. There is a general lack of labels across both the main and supplementary figures, particularly of axis labels. Additionally, it appears that post-processing of the figures has introduced inconsistencies in some panels. For example, in Figure 4e, the second bar displays a higher value than the first despite being visibly smaller, which suggests an issue introduced during figure editing. This is evident when comparing with the original plot, which is one of those that can be found in the source code provided in the accompanying Jupyter notebooks. In my view, presenting the original versions would likely offer a better representation of the results.

Response: We have addressed these issues by adding appropriate axis labels to all main and supplementary figures (Fig. 4c, e, and Supplementary Fig. 2b) and correcting inconsistencies introduced during post-processing. For example, in Figure 4e, the bar heights have been adjusted to accurately reflect the underlying data. Wherever possible, we now use plots generated directly from the original source code to ensure visual consistency and data integrity.

7. More details on the assembly and integration of the datasets would be important. Was any normalization performed? While batch effect correction using ComBat is mentioned in the methods, wouldn't ComBat-seq be more appropriate for count data? It would also be helpful to clarify the batch and group parameters used in the correction.

Response: We integrated the UCSC Xena breast cancer cohorts (Chin, Miller, Desmedt, TCGA, Caldas). These datasets were preprocessed with R/affy using the RMA procedure (background correction, quantile normalization, and probe summarization) to obtain log2-scale continuous expression matrices. Our analyses, therefore, used continuous values rather than integer counts. Given the continuous inputs and the need to integrate across platforms (microarray and RNA-seq), batch correction was performed with ComBat, rather than ComBat-seq.

Parameter settings:

- key = 'batch': ['Chin','Miller','Desmedt','TCGA','Caldas'], capturing study/platform-level effects;
- group (covariates): not included, to avoid regressing out biologically relevant differences related to outcome/subtype and to retain signals useful for prediction.

8. Several parts of the Methods section is very similar to the original publication from the authors (PMID: 38184630), thus instead of duplicating text it would be more appropriate to reference the original formulation and mention the differences.

Response: We have removed redundant and overlapping text from the Methods section and now reference our previous publication (PMID: 38184630) where appropriate. The revised text highlights only the key differences and updates that are specific to the current study.

9. Multiple gene signatures of EMT exist, particularly those that separate genes that should be up/down regulated. Given the central role of EMT enrichment in this analysis, the authors should show that this can be recapitulated with other EMT signatures.

Response: We agree and have performed a signature-robustness analysis across our benchmark cohorts. For each cancer type, we randomly selected ≥1 dataset and recomputed EMT scores using three independent up/down sources:

HALLMARK_EMT (MSigDB Hallmark). We retrieved *HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION* (e.g., ABI3BP, ACTA2, ADAM12, ANPEP, AREG, BGN, BMP1) via msigdbr and used the per-cell UCell score as an EMT activity measure (higher = stronger EMT program) :

$$EMT_{polarity}^{HALLMARK} = UCell(HALLMARK)$$

FOROUTAN (MSigDB; TGFB EMT UP/DOWN). We retrieved *FOROUTAN_TGFB_EMT_UP* (e.g., ABCA1, ACKR3, ACTN1, ADAM12, ADAM19, ANGPTL4) and *FOROUTAN_TGFB_EMT_DN* (e.g., ABLIM1, AGR2, ALDH1A3, ALDH3A2, AQP3, ARHGAP29) via msigdbr and likewise computed per-cell UCell scores:

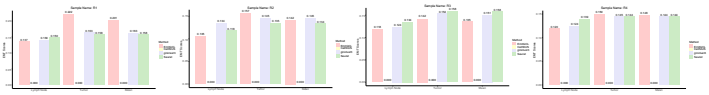
$$EMT_{polarity}^{FPROUTAN} = UCell(UP) - UCell(DOWN)$$

KOHN (MSigDB curated sets). We obtained the paired epithelial and mesenchymal EMT gene sets from MSigDB via msigdbr—*KOHN_EMT_EPITHELIAL* (e.g., CLDN3, CLDN4, CLDN7, ESRP1, ESRP2, GRHL1, EPN3, EHF, ADAP1, ATP2C2) and *KOHN_EMT_MESENCHYMAL* (e.g., VIM, ZEB1, ZEB2, SNAI1, SNAI2, MSN, QKI, EMP3, CCDC88A). Per cell, we computed signature scores using the UCell R package (Seurat function AddModuleScore_UCell):

$$EMT_{polarity}^{KOHN} = UCell(Mes) - UCell(Epi)$$

Across datasets, the three scores were highly concordant, yielding consistent sample rankings and group differences. Together, these results demonstrate that our conclusions are robust to the choice of EMT gene signatures.

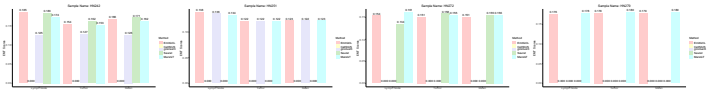
GSE162631



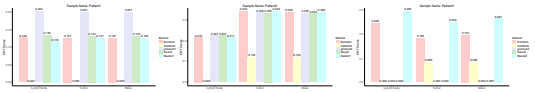
GSE180286

GSE241184

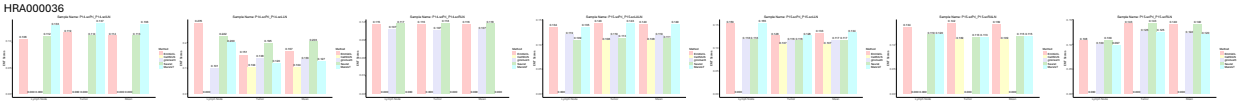
GSE188737



GSE197177

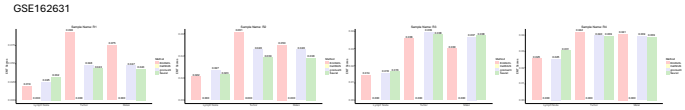


HRA000036

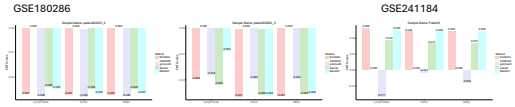


(1) Calculate EMT score based on *HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION*

GSE162631

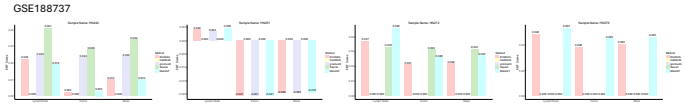


GSE180286

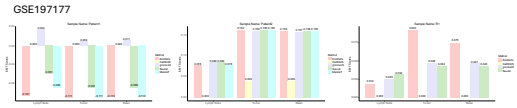


GSE241184

GSE188737

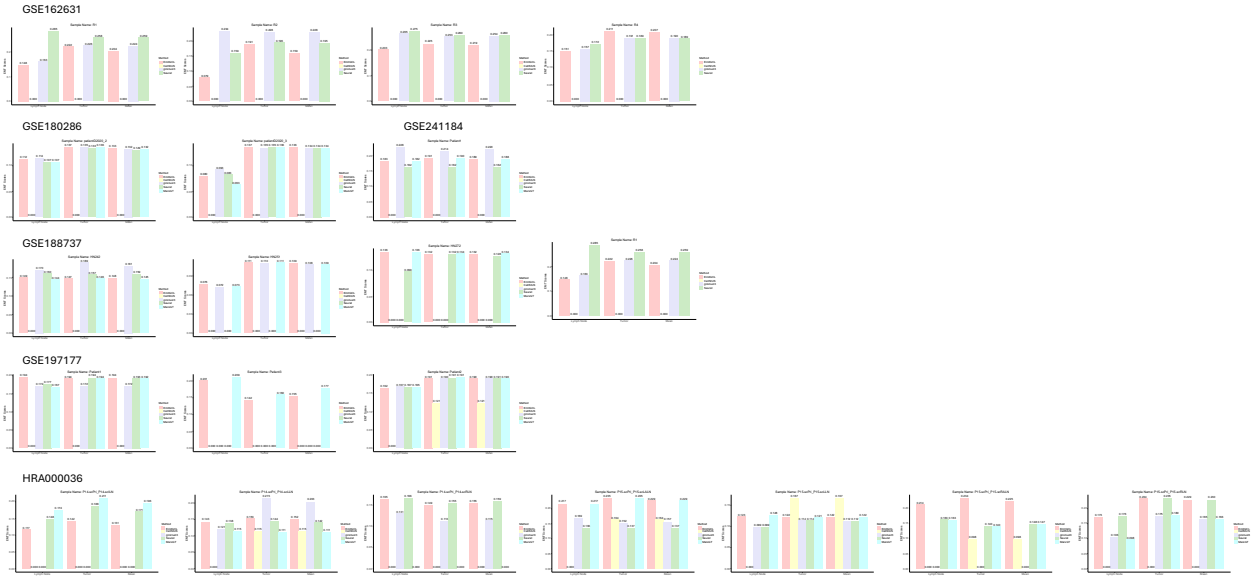


GSE197177



HRA000036

(2) Calculate EMT score based on FOROUTAN_EMT



(3) Calculate EMT score based on KOHN_EMT

10. (Remarks on code availability) The GitHub link is currently inaccessible, but the code is provided to the reviewers as a zip file. Will the authors make the repository publicly available upon publication? The code would benefit from improved organization; for example, Biomarker_inference.ipynb contains execution errors, and code comments are sparse and often not in English.

Response: Thank you for your feedback. In the revised version, we have uploaded the code to GitHub under a non-commercial license, and it is now accessible. The errors in the previous version were caused by a key/configuration issue; all steps now run end-to-end. We have also reorganized and updated the codebase to improve usability, including fixing execution errors in notebooks (e.g., Biomarker_inference.ipynb) and adding clearer English-language comments to enhance clarity and reproducibility.

11. Page 5: Improve sentence: “GiniClust17, was shown as the second-best tool for rare cell identification using clustering methods, while CellSIUS27 was the second-best tool for rare cell identification using classification methods.” for example “GiniClust17 and CellSIUS27 were the second-best tool for rare cell identification using clustering and classification methods, respectively.”

Response: The suggested changes have been addressed. In addition, we carefully reviewed the full manuscript and applied similar improvements where applicable to ensure consistent and concise expression throughout the text.

(Remarks on code availability):

A README file in the root of the code repository is missing. As mentioned in the main comments: The code would benefit from improved organization; for example, Biomarker_inference.ipynb contains execution errors, and code comments are sparse and often not in English. Response: Thank you for your feedback. In the revised manuscript, we have uploaded the code

to GitHub under a non-commercial license, where it is now publicly available. The response to this point is the same as that to comment 10.

Reviewer #3

The manuscript presents a new graph contrastive learning approach, call EmitGCL, to detecting occult metastatic cells and metastatic precursor cells from scRNAseq data. Based on EmitGCL, the authors reported the identification and validation of two predictive biomarkers for future breast cancer metastasis. Furthermore, they demonstrated YY1 TF as a key driver of breast cancer metastasis, which was validated in-silico and by CRISPR-based migration assays. Overall, the manuscript is well-written, the deep-learning approach is reasonable and scalable, and the findings were carefully validated.

The above being said, I have the following major concerns.

1) Insufficient and unclear benchmarking results.

According to the description of model inputs in the Methods section, EmitGCL is trained by a pair of scRNAseq datasets, one from the primary site, and the other from a metastatic site. Thus, to gather the baseline performance of the approach in terms of false positive rates and true positive rates, it would be natural to use unmatched/mismatched pairs, in addition to matched pairs, as inputs to EmitGCL, such as a pair of primary site datasets (with one mislabeled as metastatic), a pair of metastatic site datasets (with one mislabeled as primary), and a pair of mismatched primary-metastatic datasets, all from two different patients. Personally, I would also gather some additional scRNAseq datasets collected from healthy samples to assist the mismatched pair generation process. Moreover, in the reported results, all cells in the matched dataset pair were used in training EmitGCL. This naturally raises the concern on whether there is overfitting. One approach to make sure this is not the case is to leave out 20% of cells (or even 30% as one aims to identify rare cell types here) in the training process. Given the subgraph sampling approach the authors took when training the model, this is very much doable. Such training-test sample splitting is a common practice for ML or DL approaches, I anticipate that they can also be implemented on other methods in comparison. Then, one can perform inference on the left-out test cells and calculate various performance metrics on them across all methods considered in the benchmarking study.

Response:

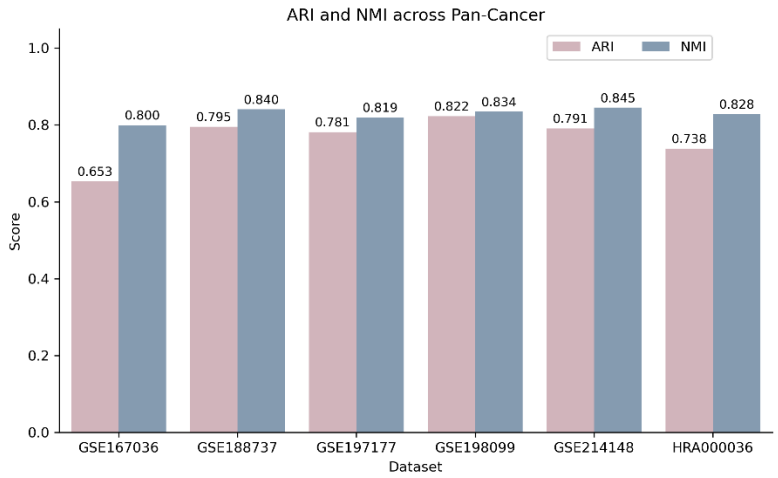
i) We appreciate the suggestion to evaluate mismatched pairs. EmitGCL is designed to learn patient-specific correspondences between paired primary and metastatic scRNA-seq samples using contrastive learning (rather than transfer learning).

- Healthy samples. Occult or metastatic precursor cells are malignant by definition. If a healthy sample is used, the model would have no malignant signal to align to, and thus cannot identify occult or metastatic precursors.
- Mislabeled or unmatched pairs.

- ✧ Primary–primary (one mislabeled as metastatic): the model may highlight divergent cancer subpopulations, but these would reflect within-primary evolution rather than a primary-to-metastasis transition.
- ✧ Metastasis–metastasis (one mislabeled as primary): similarly, any “precursor-like” cells would reflect within-metastasis variation, not a primary-to-metastasis trajectory.
- ✧ Cross-patient pairs: pairing samples across patients breaks the intended patient-specific alignment. While the model might still assign “precursor-like” scores, these cannot be interpreted as metastatic risk for either patient.

Accordingly, mismatched or healthy pairs are not an appropriate baseline for estimating false- or true-positive rates in the intended use case.

ii) Thank you for the suggestion regarding potential overfitting. To address this, we randomly split each dataset into training (80%) and testing (20%) subsets and retrained EmitGCL. We then evaluated the clustering results on the held-out test cells using ARI and NMI (see figure below). Both ARI and NMI remained high, demonstrating that EmitGCL generalizes well and does not overfit to the training cells.



Supplementary Fig. 3c: ARI and NMI with 8:2 train-test split.

2) Model training.

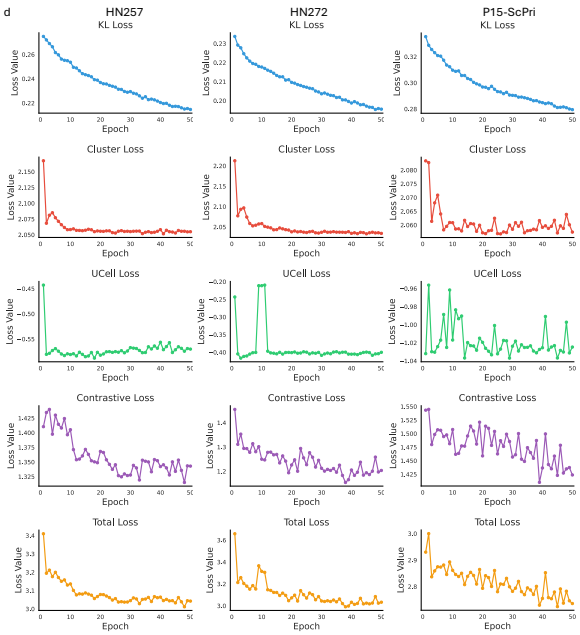
In equation (16), the authors proposed to train the weights by using a multi-task loss function that has four terms. While each of the four terms makes sense, I am concerned that they are just summed up without proper weights. In particular, it is not self-evident that the relative magnitudes of the four terms are balanced. For example, it is possible that two terms are much bigger than the other two and so the loss becomes effectively the sum of the two dominating terms. The authors should report the sizes of these terms across the datasets that they have fitted the model on. If they are not balanced, then a proper weighing scheme should be in place, ideally with recommended default weights. In addition to the foregoing issue, there are ambiguities in the current description of the bipartite graph subsampling scheme. In the paragraph after equation (4), it is stated that “The number of reserved genes is denoted as N_g.” However, neither “reserved genes” nor “N_g” is defined. Moreover, it is stated that “In each site, the subgraph

incorporates 30 randomly selected cells …". I am assuming it means that each subsampled subgraph has 60 randomly selected cell nodes (30 from either site) plus their chosen neighbor gene nodes. If the numbers of cells are very different in primary vs metastatic sites, what are the ramifications of this balanced sampling scheme?

Response:

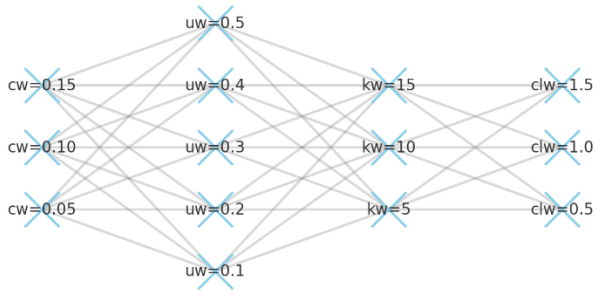
1) Each loss component is computed as the mean over its contributing elements within each mini-batch, which places the four terms on comparable scales. We therefore fixed the weights equally ($\lambda_1=\lambda_2=\lambda_3=\lambda_4=1$) across all experiments and did not expose them as user-tunable parameters. To evaluate balance, we examined three representative datasets (HN257, HN272, and P15-ScPri). In all cases, the four loss components decreased smoothly and converged, with magnitudes remaining within the same order of magnitude throughout training. No single term exceeded another by more than ~10×. We have added the per-epoch trajectories in the figure below, which demonstrate that all four loss terms evolve in a balanced manner.

.



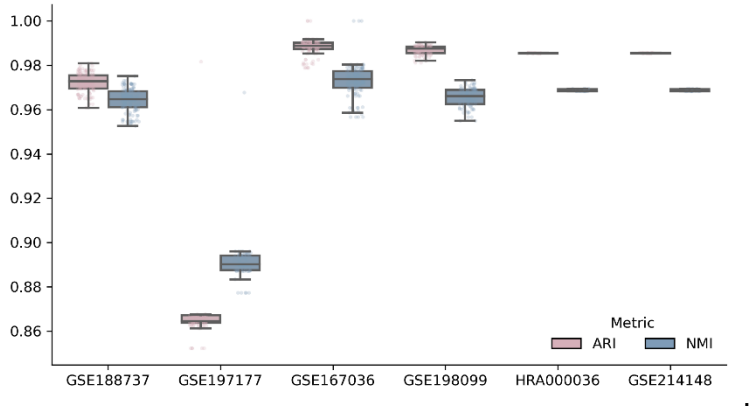
Supplementary Fig. 3d. Loss trajectory for per epoch.

To test whether our results depend on the choice of loss weights, we conducted a grid search of 135 combinations (see figure below) using one dataset per cancer type.



Supplementary Fig. 3e. Grid search for parameter combinations.

The results show that our model was insensitive to the weight settings across this range (see figure below).



Supplementary Fig. 3f. NMI and ARI for 135 parameter combinations in each cancer type.

2)

i) For the first comment: Definition of reserved genes and N_g.

We would like to clarify that the term “reserved genes” refers to the genes associated with a given cell $v_{j_0}^C$ that exhibit high rare-signal probabilities, as computed using equation (4). These genes are retained for downstream processing, and their total number is denoted as N_g . This is not a newly introduced concept but rather a shorthand notation indicating the subset of genes selected based on their high correspondence probability with cell $v_{j_0}^C$. We have revised the relevant paragraph to make this definition more explicit and avoid ambiguity.

ii) For the second comment: unbalanced cell numbers in different sites.

We would like to clarify that the subgraphs for the primary and metastatic sites are sampled independently. Each subgraph contains 30 randomly selected cells from a single site only, along

with their corresponding gene neighbors. Therefore, each subgraph is site-specific and includes only the cells from either the primary or the metastatic location—not both.

In our benchmarking experiments, we specifically included cases where there is a substantial difference in the number of cells between primary and metastatic sites (see figure below). Since the sampling is conducted independently, this imbalance does not affect the structure of individual subgraphs, and the comparison across sites remains valid.

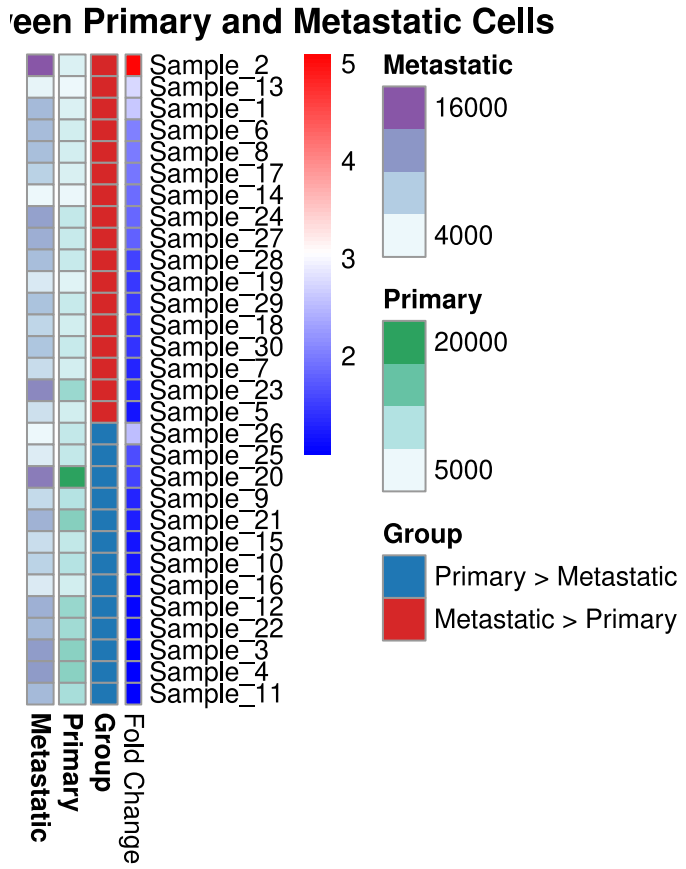


Figure legend. Primary and metastatic cell number of each sample/data

3) Batch effect correction.

The current work used ComBat as the batch correction method. However, according to the following paper “A benchmark of batch-effect correction methods for single-cell RNA sequencing data”[https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1850-9], ComBat is not as good as Harmony, LIGER, and Seurat v3. Since the EmitGCL model also take pairs of datasets as input, it makes sense to adopt one of the best batch effect correction methods available.

Response: We would like to clarify that we did not apply any batch correction methods to the single-cell RNA-seq datasets, as each sample was processed and analyzed independently. Since there is no technical batch effect across samples in our framework, batch correction at the scRNA-seq level is not necessary.

The ComBat method was used only for bulk RNA-seq data, specifically to validate the robustness of the biomarkers identified in our breast cancer analysis. ComBat operates on normalized gene expression matrices and is well-suited for correcting batch effects in bulk data, where our primary concern was to preserve gene-level expression values (Johnson et al., *Biostatistics*, 2007). This is important because our method requires information at the gene dimension rather than low-dimensional embeddings.

The benchmarking study you referenced focuses on batch correction methods for single-cell RNA-seq data, such as Harmony, LIGER, and Seurat v3. These methods are indeed powerful for integrating scRNA-seq datasets, but they do not apply to bulk RNA-seq analysis. Moreover, most of these methods only generate corrected cell embeddings rather than adjusted expression matrices, which would not meet the needs of our bulk RNA-seq-based validation.

Reviewer #4

The authors evaluated the predictive value of EmitGCL for metastasis. The authors indicated they identified HSP90AA1 and HSP90AB1 are a reliable biomarker to predict breast cancer metastasis. Although it is a interesting study, it has fundamental flaws.

1. All the data presented in this manuscript showed correlation of genes between detected metastasis and primary tumors. No patient in this study had occult metastasis which should be defined as metastasis but not detected.

Response: We thank the reviewer for the helpful input. Our approach aligns with how most metastasis risk scores are constructed, using already metastatic cases. Clinically evident metastases are typically driven by dominant subclones that have successfully colonized distant organs. However, emerging evidence suggests that minor subclones may also disseminate concurrently or sequentially and remain underrepresented at early stages. We define these newly seeded, low-abundance disseminated clones as part of the occult metastatic pool, as they reflect early, subclinical metastatic events. Our model is specifically designed to capture such rare metastatic-competent populations within the primary tumor—regardless of whether overt metastasis has been clinically detected. Moreover, we included a real clinical validation case in Fig. 3, where we analyzed a representative breast cancer patient from our dataset who had a primary tumor and two lymph nodes (LN-1 and LN-2). Their initial diagnosis was that the nodes were “initially deemed negative” by conventional methods with the patient declared to have “no evidence of disease.” Our model predicted occult metastatic cells were in both LN-1 and LN-2, despite being rare (3% of the total cell population) in these locations. Follow-up confirmed the development of a metastatic disease, validating the initial prediction by our model.

Detection of occult metastasis is a general problem across cancer types, and many studies have attempted to identify occult metastasis or micrometastases with little success. Current clinical methods rely on i) imaging techniques that still lack the sensitivity to detect minimal residual disease or small metastases, ii) pathological examination, where small metastases can be missed, and iii) molecular markers that have been limited due to false-positive results and sensitivity issues. This is the reason why EmitGCL was developed, and its exact purpose is to fill this unmet clinical need. EmitGCL can do this due to: a) knowledge-aware graph contrastive learning that does not rely on simple differential gene expression analysis that often fails to find common

biomarkers across patient cohorts, b) a high sensitivity to detect occult metastatic cells and has a low false positive rate in non-metastatic cases, and c) using attention scores that reflect a gene's contribution to distinguishing a particular cell cluster, which allowed identification of HSP90AA1 and HSP90AB1 as biomarkers.

2. There is no description of the 420 patients the authors used as validation cohorts.

Response: These samples are clearly described in the Data Description section and fully listed in Supplementary Data 1. Our Fig. 4 is entirely based on this validation cohort, which includes five public breast cancer datasets.

3. The authors failed to show expression difference of HSP90AA1 and HSP90AB1 in breast tumors that developed metastasis vs that did not develop metastasis.

Response: We appreciate this point of solving the metastasis problem using the differentially expressed gene (DEG) strategy. Here are some discussions to support our EmitGCL design, rather than directly checking DEGs.

1) Why are DEGs not enough to distinguish whether the patient will develop metastasis?

- As we claimed in the Introduction of our manuscript: "Computational studies identify predictive biomarkers through differentially expressed genes based on bulk RNA-seq data. Still, no common biomarkers across different cohort groups of patients have been identified, limiting their generalizability and effectiveness. These underscore the need for more sensitive and specific biomarkers to determine whether a cancer patient with no detectable metastasis is likely to develop it over time (i.e., future metastasis)."
- As highlighted in a recent Science paper (Forrest et al., *Science*, 2025): "Case-versus-control classifications oversimplify diseases that exist on a spectrum, further reducing the accuracy of risk estimates. Machine learning (ML) offers a scalable solution by integrating large-scale electronic health record (EHR) and genetic data to assess penetrance in a data-driven, quantitative, and precise manner." This supports our view that case-versus-control DEG analysis alone is insufficient for accurate risk score prediction. Moreover, we think that a simple difference in expression is not the only, or even the best, measure for predictive behavior.
- The biological rationale is that HSP90AA1 and HSP90AB1 are essential molecular chaperones that maintain the integrity and function of many oncogenes, which is critical for cancer cells to drive the processes and proliferation required for cells to achieve metastasis. Beyond protecting the integrity of oncogenes, these chaperone complexes participate in cell stress responses and general protein homeostasis, both of which are also required for cells going through the stressful process of metastasis. The expression levels that facilitate the ability of HSP90AA1/HSP90AB1 to perform these functions may not be significantly different from those in non-metastasis tumor cells when looking at bulk data, but rather the attention scores that are most predictive. High attention scores do not necessarily require the obvious differential expression that would be detected in routine DEG analysis.

- Future metastasis involves cancer cells that initially appear as a rare population at the metastatic site. However, conventional DEG analysis first selects highly variable genes, which tend to overlook the critical signals carried by such rare cells. In our framework, we address this limitation by leveraging precursor cells associated with early dissemination to predict future metastasis.

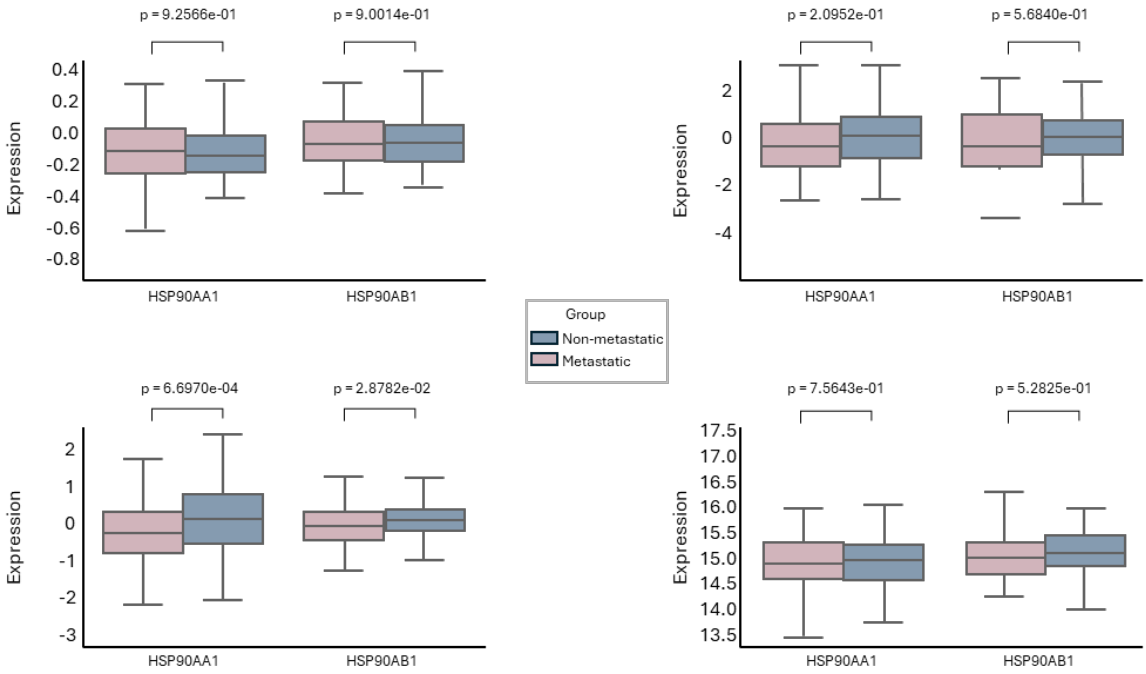
The above consideration is the entire rationale of having this new deep learning model beyond simple DEG analysis. More importantly, our approach focuses on model attention-derived genes, which are more consistent across patients, other than DEGs. Figure 4 shows that attention to HSP90AA1/AB1 can be used to predict future metastasis, not their raw expression levels.

2) What do we do in EmitGCL to overcome this limitation?

- Our approach focuses on model attention-derived genes. Simply speaking, each cell has many genes expressed at different levels. The attention mechanism is like a spotlight: It shines brighter on the more informative genes (e.g., future metastatic). It dims on the genes that are common, noisy, or not helpful. This way, the model learns to focus on the right subset of genes for each cell, depending on whether it is a primary cell, a metastatic precursor, or an occult metastatic cell.

3) Experiment results support the main idea above.

- Using 420 samples, we performed a DEG analysis and found no consistently significant DEGs (Supplementary Fig. 6a), suggesting that a DEG-based approach cannot resolve this problem.
- Attention-based genes are more consistent across patients, other than DEGs. Figure 4 shows that attention to HSP90AA1/AB1 can be used to predict future metastasis, not their raw expression levels.
- We examined HSP90AA1 and HSP90AB1, genes highlighted by our model's attention mechanism, to illustrate expression differences between the Non-Metastatic and Metastatic groups. These differences were not detected by standard DEG analyses (Supplementary Fig. 6b).



Supplementary Fig. 6b. Differential expressions of HSP90AA1 and HSP90AB1 between Non-Metastatic and Metastatic.

4. In figure 4. The authors showed survival difference but not metastatic rate, which is very odd.

Response: The x-axis in Fig. 4d and g represents time to metastasis, not general survival. We will update the legend to clarify this. This is an important distinction as metastasis-free survival is a measure of metastatic rate, which is critical for this model as a predictor of metastasis. The updated figure is shown below:

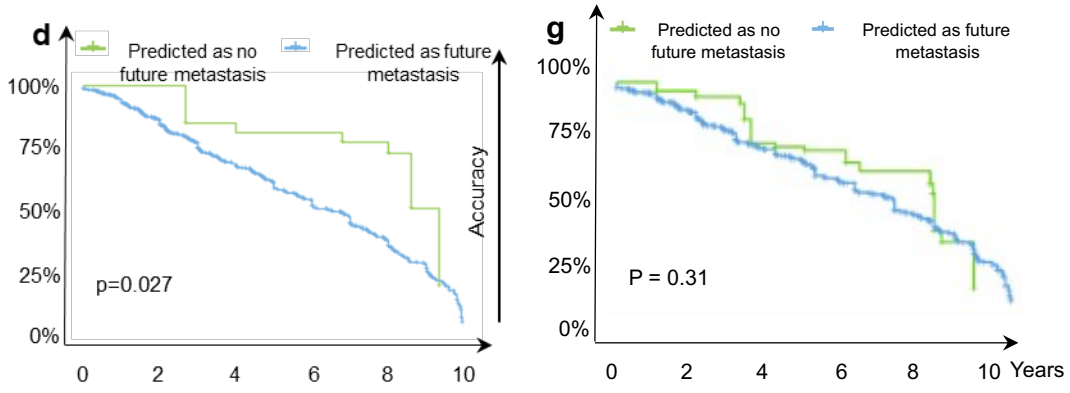


Fig. 4d, Survival curve stratified by predicted future metastasis results using attention core of HSP90AA1 and HSP90AB1. **g,** Metastasis-free survival curve stratified by predicted future metastasis results using the attention core of ACTB.

5. To validate the biomarkers, the authors should obtain primary tumor samples from patients who did not have metastasis at the time of surgery. The authors should follow these patients to validate their biomarkers. I do not see any convincing validation in this manuscript.

Response: We thank the reviewer for this point, as we agree that a prospective clinical validation is the ultimate goal. We wanted to highlight that our study does include a real-world validation case, shown in Figure 3. In this case, we analyzed a representative patient from our dataset who had a primary tumor and two lymph nodes (LN-1 and LN-2). At their initial diagnosis and surgery, these lymph nodes were “initially deemed negative” by conventional pathological evaluation with “no evidence of disease.” Our model predicted the presence of occult metastatic cells in both LN-1 and LN-2. Follow-up confirmed the development of metastasis, thereby providing a real-world validation scenario that aligns with your suggestion.

Reviewer #1 (Remarks to the Author):

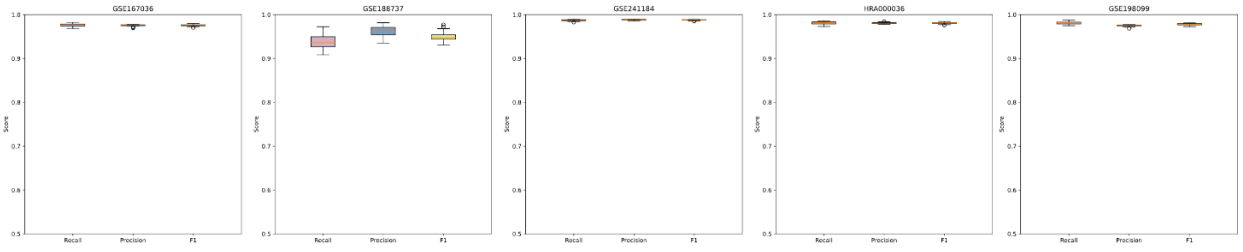
The authors have addressed most of my concerns. However, regarding comment 1), I do not think that evaluating robustness solely by ARI and NMI between clustering obtained under different initialization seeds and the reference clustering is sufficient. Because a key claim of EmitGCL is its ability to detect rare cell types and to identify biologically meaningful marker genes via the attention mechanism, the robustness of these specific outputs should also be quantified. It would be helpful if the authors could design more targeted metrics to assess (i) whether the same rare cell populations are consistently recovered across seeds and (ii) whether the marker genes highlighted by the attention weights are stable and biologically coherent across different runs.

Response: Thanks. We agree that ARI and NMI are primarily driven by major cell populations and therefore are insufficient to evaluate the robustness of EmitGCL with respect to its core claims. In this revision, we performed the following two robustness analyses targeting the key biological outputs highlighted by the reviewer.

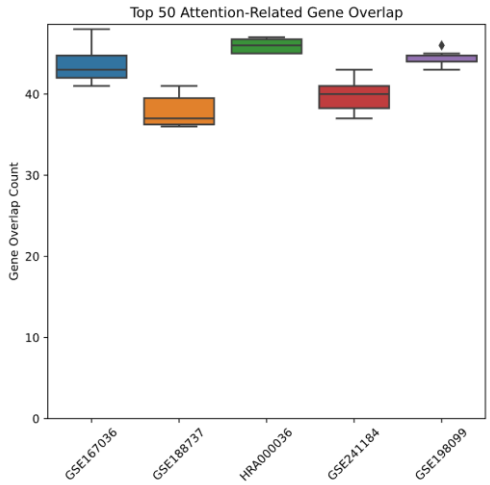
1. For the comments: whether the same rare cell populations are consistently recovered across seeds. We evaluated EmitGCL under 20 independent random initializations and assessed whether the same rare cell populations were consistently recovered across runs. Robustness was quantified using F1 score at the rare-cell level, explicitly measuring the reproducibility of rare cell identification rather than global clustering agreement. As shown below, this analysis was conducted across five independent datasets, each corresponding to a different cancer type (GSE167036, GSE188737, GSE241184, HRA000036, and GSE198099). Across all datasets, EmitGCL achieves consistently high F1 scores with minimal variance across random seeds (see Table below). These results demonstrate that the same rare cell populations are reproducibly identified across different initializations, indicating that EmitGCL's ability to identify rare cell states is stable and not driven by stochastic effects.

Dataset	Recall Mean	Recall Median	Recall Variance	Precision Mean	Precision Median	Precision Variance	F1 Mean	F1 Median	F1 Variance
GSE167036	0.975841	0.975895	0.000011	0.975156	0.975492	0.000005	0.975495	0.975543	3.704039e-06
GSE188737	0.939160	0.935780	0.000383	0.961689	0.962264	0.000217	0.950123	0.948837	1.431699e-04
GSE241184	0.986979	0.987006	0.000003	0.988261	0.988545	0.000001	0.987619	0.987675	9.520830e-07
HRA000036	0.980924	0.981857	0.000010	0.980838	0.980680	0.000003	0.980878	0.981029	4.079372e-06
GSE198099	0.980955	0.980996	0.000009	0.974947	0.975285	0.000004	0.977940	0.978515	4.721753e-06

The detailed boxplot of F1 score for robustness of rare cell detection across random initializations are shown in the Figure below.



2. For the comments: whether the marker genes highlighted by the attention weights are stable and biologically coherent across different runs. To assess the robustness of the attention mechanism, we examined the overlap of the top 50 genes ranked by attention weights across the same 20 random seeds. We quantified the number of shared genes across runs to evaluate whether the same marker genes were consistently highlighted. As shown below, 72–94% of attention-ranked genes are reproducibly recovered across random seeds, demonstrating strong stability of attention-based gene prioritization.



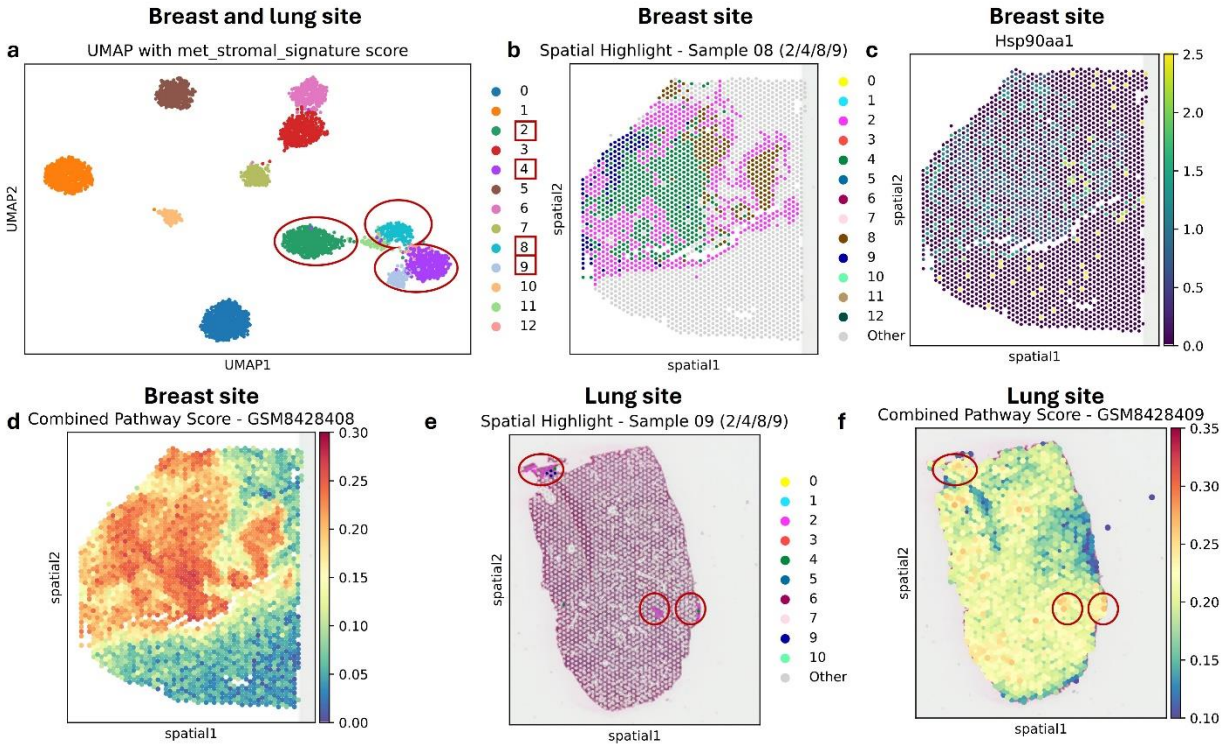
Together, these analyses complement ARI and NMI by evaluating the robustness of EmitGCL's rare cell discovery and attention-based marker identification, providing a more targeted and biologically relevant assessment of model stability. These results have been included in **Supplementary Fig. 3**.

Regarding comment 4), I appreciate the authors' detailed response and agree that a full spatial extension of EmitGCL may ultimately require a dedicated algorithm. However, I am still not fully convinced by the argument that the lack of multi-slice ST data precludes any spatial validation at this stage. In recent years, a growing number of multi-slice or multi-region ST datasets across several cancer types have become publicly available, and even single-slice ST data can already provide spatially resolved patterns of metastatic propensity, niche composition, and rare cell populations. Given that a central claim of EmitGCL is the ability to detect rare metastatic precursors and biologically meaningful markers, an analysis on at least one existing ST dataset (e.g., assessing whether EmitGCL-derived risk scores or marker signatures align with spatially localized high-risk niches or early metastatic regions) would substantially strengthen the biological credibility of the method. Even a proof-of-concept application on publicly available ST data, without developing a fully new spatial framework, would be much more convincing than limiting this point to a discussion of future work.

Response: We thank the reviewer for this constructive suggestion. In this revision, we performed additional spatial analyses using publicly available datasets (GSE273439), which is a paired mouse breast cancer–to–lung metastasis datasets with spatial transcriptomic measurements. We applied EmitGCL directly to the spatial transcriptomic data and projected EmitGCL-derived risk scores and marker signatures onto the spatial coordinates.

As shown in the Figure below, EmitGCL identified groups 2, 4, 8, and 9 from the scRNA-seq data (**a**). When mapped to spatial transcriptomics data of the primary breast tumor, these groups localized to distinct high-risk regions (**b**). To validate these spatial predictions, we first examined the expression of the predicted biomarker Hsp90aa1 (note that this dataset does not include Hsp90ab1). Elevated Hsp90aa1 expression was specifically observed within EmitGCL-defined high-risk regions (**c**). We further computed spatial scores using established metastasis-associated hallmark gene signatures, including angiogenesis, invasion–metastasis cascade, PI3K/Akt/mTOR signaling, MAPK/ERK signaling, Hippo signaling, and extracellular matrix (ECM) degradation. These hallmark scores showed strong spatial concordance with EmitGCL-derived risk maps in the primary breast tumor (**d**). Meanwhile, applying the same mapping strategy to lung tissue revealed that EmitGCL-defined high-risk groups localized to spatially restricted regions suggestive of early metastatic niches (**e**). To independently validate these predictions, we applied the same metastasis hallmark scoring framework to the lung dataset. Regions with elevated hallmark scores overlapped with EmitGCL-predicted high-risk niches, supporting the presence of spatially localized areas with high metastatic potential consistent with early metastatic colonization (**f**).

Together, these analyses provide a direct spatial validation of EmitGCL on existing ST data, demonstrating that the model can recover biologically meaningful, spatially localized metastatic risk regions even without a specialized spatial modeling framework. We have incorporated these results as a proof-of-concept spatial validation in the revised manuscript, addressing the reviewer's concern and strengthening the empirical support for EmitGCL's ability to detect rare metastatic precursors and relevant molecular programs.



Reviewer #2 (Remarks to the Author):

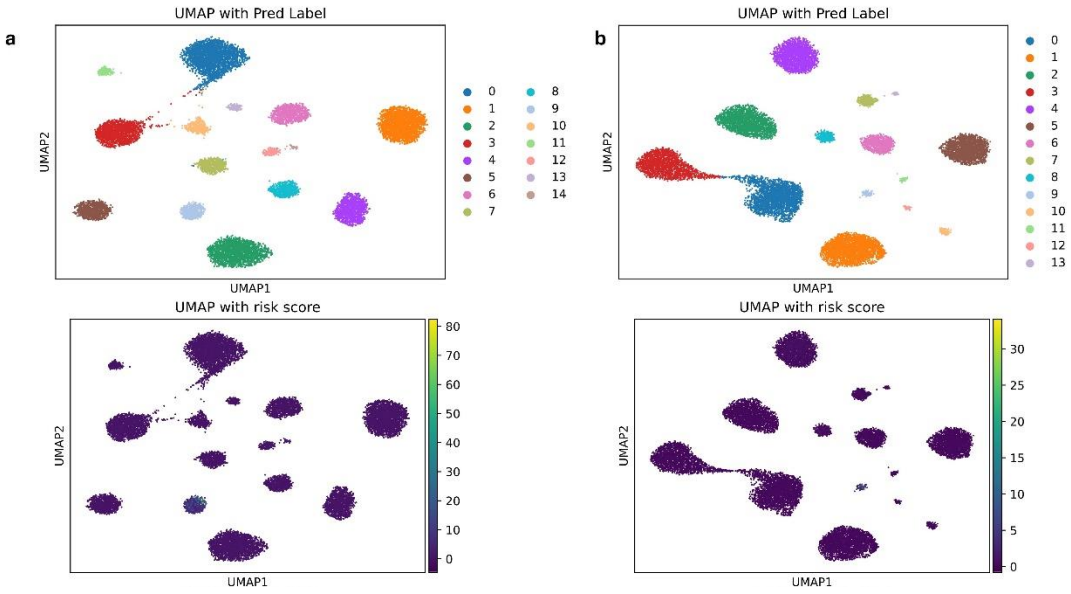
The authors have addressed most of my previous comments. However, I still find that some of the general claims are overstated given the results presented, and several figures are unnecessarily over-processed which introduce errors (e.g. Figure 4e). While I agree with the authors that applying their method to healthy cells is not the primary aim of the study, I still believe there is value in leveraging the large number of available scRNA-seq datasets from healthy cells as controls.

Response: We thank the reviewer for their careful evaluation. We have addressed the remaining concerns as follows.

1. Regarding overstated claims, we carefully reviewed the manuscript and revised several statements that could be interpreted as overstated. These modifications were made primarily in the Abstract and Results sections and are highlighted in red in the revised manuscript to ensure clarity and transparency.

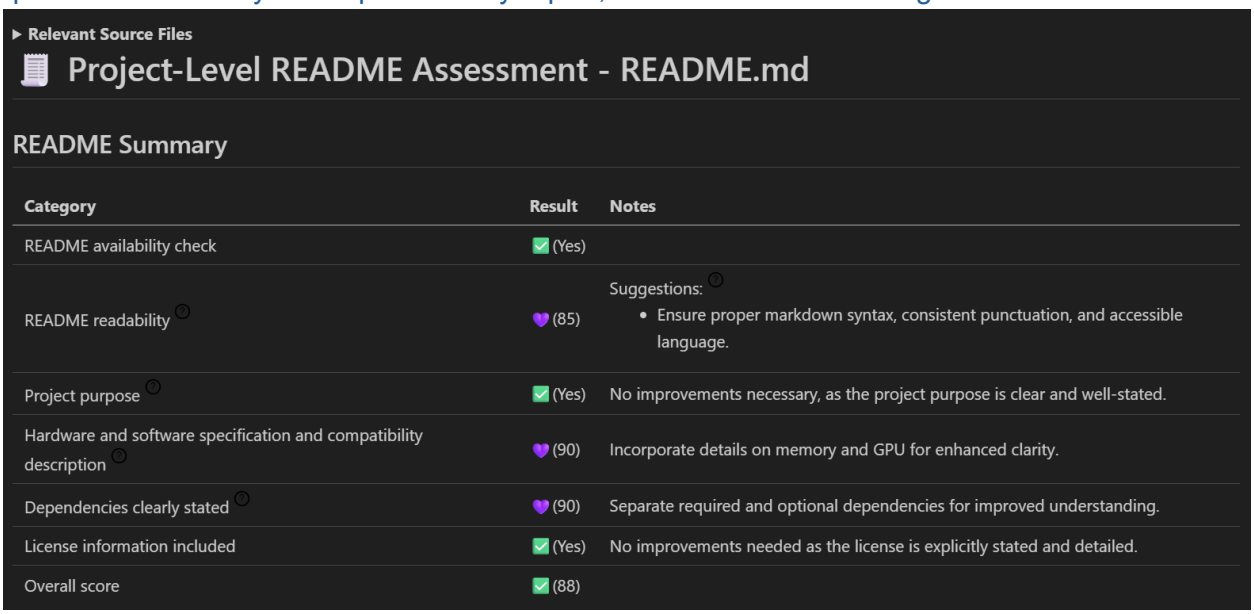
2. Regarding over-processed figures, we have revised the affected figures (including Figure 4e) by directly presenting the original results without additional processing, thereby improving accuracy and interpretability.

3. Regarding the use of healthy scRNA-seq datasets as controls, we agree that such analyses can provide important negative controls. We therefore leveraged large-scale healthy scRNA-seq datasets from the Human Cell Atlas (ERP136281) to simulate paired tissue conditions, including stomach–lymph node and bladder–lymph node. When applying EmitGCL to these healthy tissue pairings, we did not identify any early metastatic cell populations or metastatic precursor states (see Figures below: **a**. stomach–lymph node and **b**. bladder–lymph). These results support the specificity of EmitGCL and suggest that the detected signals in cancer datasets are unlikely to arise from normal tissue variation alone.



Reviewer #2 (Remarks on code availability):
The code is now available in GitHub with some level of documentation. The tutorial available needs improvement (e.g. contains absolute paths which are not portable to other users).

Response: Thanks. We have implemented additional user experience improvements to the GitHub repository. In particular, we refined the documentation and removed non-portable elements (e.g., absolute paths) to improve usability across different user environments. In addition, we employed an in-house AI agent to systematically evaluate the codebase and generate a quantitative usability and reproducibility report, as summarized in the figure below.



▸ Relevant Source Files

Project-Level README Assessment - README.md

README Summary

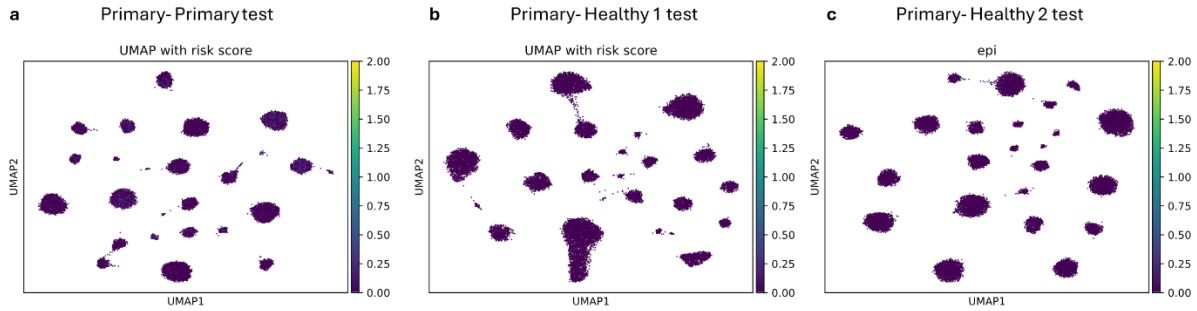
Category	Result	Notes
README availability check	✅ (Yes)	
README readability	💜 (85)	Suggestions: • Ensure proper markdown syntax, consistent punctuation, and accessible language.
Project purpose	✅ (Yes)	No improvements necessary, as the project purpose is clear and well-stated.
Hardware and software specification and compatibility description	💜 (90)	Incorporate details on memory and GPU for enhanced clarity.
Dependencies clearly stated	💜 (90)	Separate required and optional dependencies for improved understanding.
License information included	✅ (Yes)	No improvements needed as the license is explicitly stated and detailed.
Overall score	✅ (88)	

Reviewer #3 (Remarks to the Author):
Two of the concerns I had in my previous round report have not been carefully addressed by this revision. Please see below for details. (1) My concern: Insufficient and unclear benchmarking results. In my previous round report, I suggested the inclusion of unmatched/mismatched pairs for understanding the false positive behavior of EmitGCL. The authors dismissed this suggestion in their response letter and replied “Accordingly, mismatched or healthy pairs are not an appropriate baseline for estimating false- or true-positive rates in the intended use case.” While I am not exactly sure what “the intended use case” is in authors’ mind, I imagine one unavoidable use case is to detect whether a biopsy samples from a patient’s primary tumor contains metastatic precursors, and another is to detect whether a potential metastatic site with no visible tumor actually contains occult metastatic cells. Both primary-healthy and primary-primary test cases are ideal for understanding the method’s false positive rate when calling occult metastatic cells. Moreover, these cases are not artificial. It is possible that a suspected metastatic site is actually the primary site of a separate tumor or simply a malign one. Thus, I do not understand why the authors dismissed my initial comment.

Response: Thanks. In this revision, we performed additional benchmarking using biologically realistic unmatched and mismatched scenarios based on the publicly available scRNA-seq dataset GSE200972, which profiles spatially distinct regions of primary lung tumors from the same

patients. Each patient contributes two tumor regions sampled from different locations within the same primary tumor, together with matched adjacent normal lung tissues, providing a well-defined negative-control setting.

For the primary–healthy pairs, the primary tumor samples were paired with their matched adjacent normal lung tissues from the same patient. Specifically, we evaluated the combinations of primary lesion 1 with normal tissue 1 and primary lesion 2 with normal tissue 2. As shown in the Figure below, (**a**) and (**b**) EmitGCL did not identify metastatic precursor cells or occult metastatic cells in either primary–healthy setting. In addition, to model the scenario in which a suspected metastatic lesion represents an independent primary tumor, we paired the two primary tumor lesions from the same patient. As shown in the Figure below, (**c**) EmitGCL did not identify metastatic precursor cells or occult metastatic cells in this primary–primary comparison.



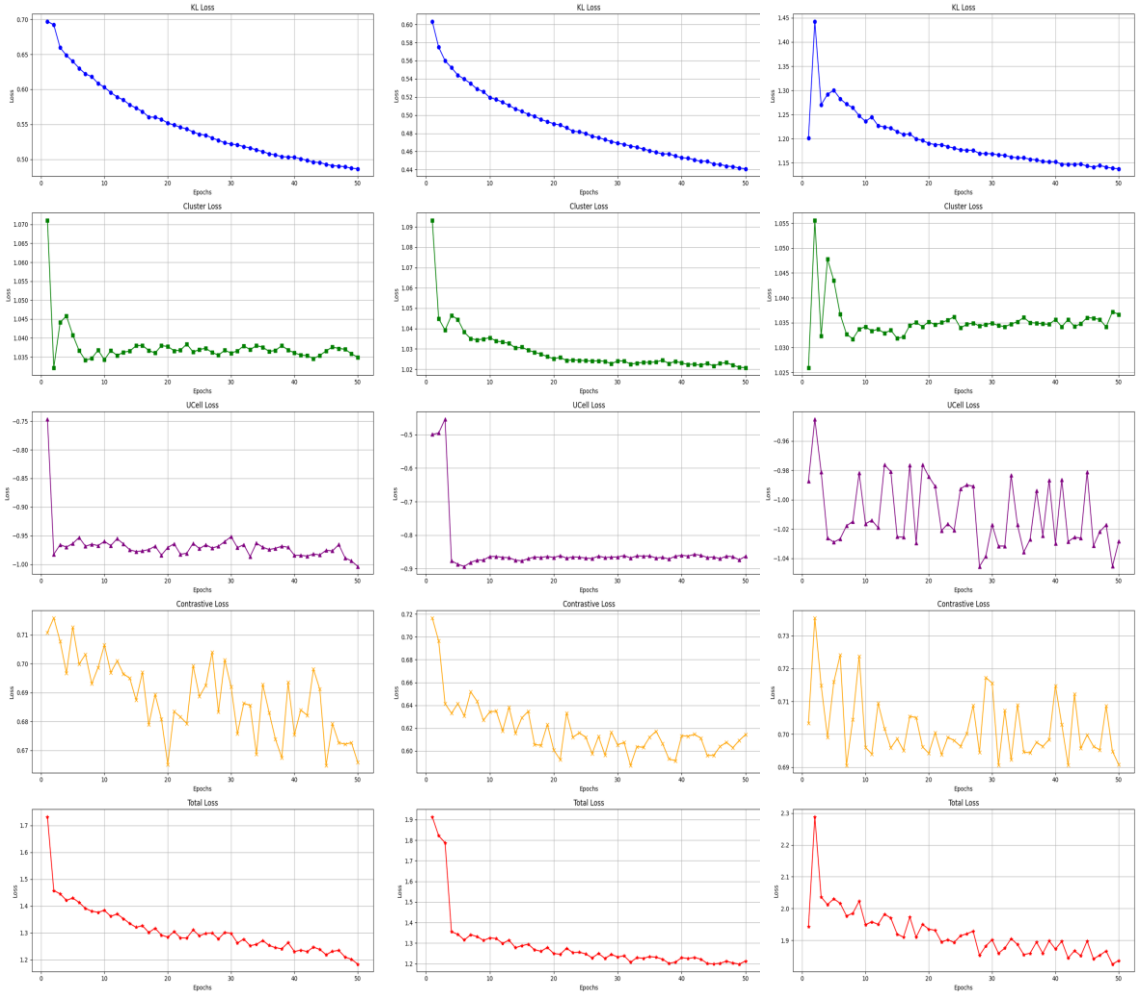
These results indicate that EmitGCL does not spuriously call metastatic precursors or occult metastatic cells in unmatched or mismatched negative-control scenarios, directly addressing the reviewer's concern regarding false positive behavior.

(2) My concern about model training objective in my previous round report "In equation (16), the authors proposed to train the weights by using a multi-task loss function that has four terms. While each of the four terms makes sense, I am concerned that they are just summed up without proper weights." In their response letter, the authors declared that "Each loss component is computed as the mean over its contributing elements within each mini- batch, which places the four terms on comparable scales." and thus regarded weighing the four terms as unnecessary. However, this claim was not supported by the results reported in Supplementary Fig. 3d. For example, the range of the contrastive loss term is almost ten times that of the KL loss term for HN272. In addition, I cannot understand the reasoning behind the weight search grid reported in Supplementary Fig. 3e. The current weights are all ones, and the search grid in the figure set possible weights for some terms to be all below 0.2 and for some others to be all above 5, essentially making the search grid uninformative for balancing the four terms. I encourage the authors to take this issue more seriously, as there four terms are intrinsically different and there is no obvious reason for assigning them equal weights.

Response: We thank the reviewer for the suggestion regarding the weighting of the multi-task loss terms. We therefore addressed this concern through both theoretical justification and empirical validation, as detailed below.

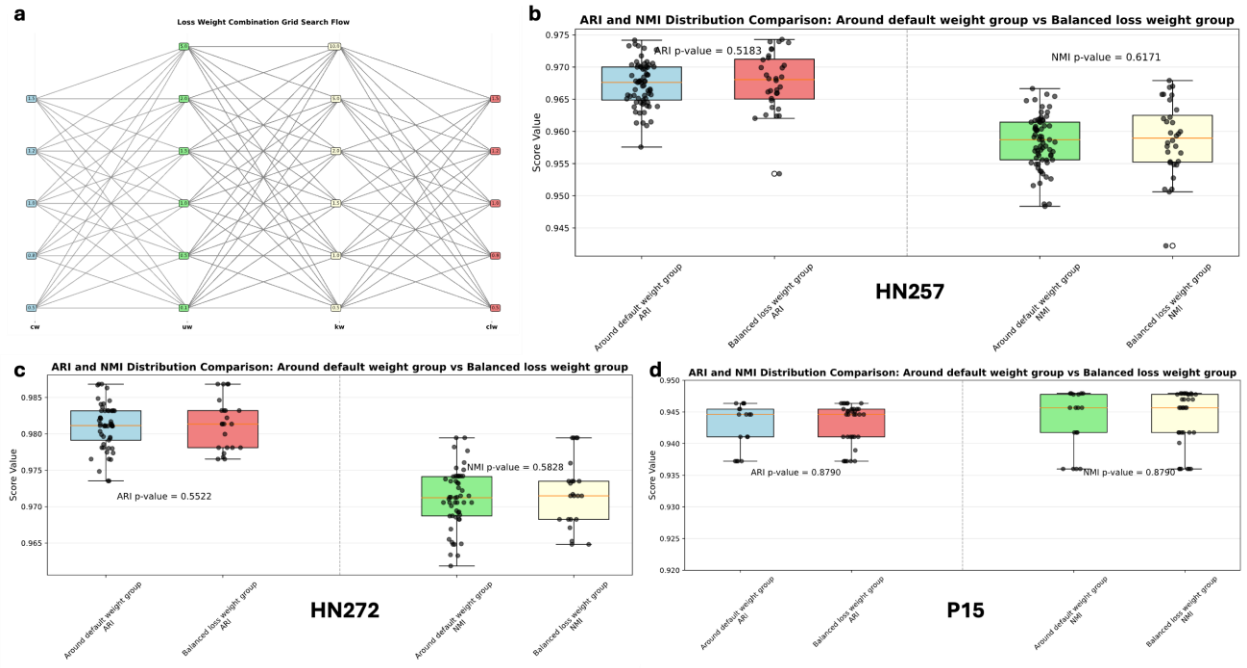
1. At the theoretical level, model training was performed using the AdamW optimizer, which normalizes parameter updates by the historical second moments of gradients. As shown by Loshchilov and Hutter (ICLR 2019), AdamW reduces sensitivity to raw loss magnitudes by adaptively rescaling gradients, such that loss terms with larger numerical ranges do not necessarily dominate the optimization process. In addition, the decoupled weight decay in AdamW prevents implicit coupling between regularization strength and gradient scaling, further stabilizing multi-objective optimization. Under this optimization scheme, the relative influence of different loss terms is governed primarily by their gradient dynamics rather than their absolute loss values, providing a principled explanation for the observed robustness to loss-weight variation. Based on these theoretical considerations, we initially expected the model to exhibit limited sensitivity to moderate variations in loss weights. **Meanwhile, we fully agree with the reviewer that empirical verification is essential.** In the revised manuscript, we conducted systematic experiments to directly assess whether AdamW indeed mitigates sensitivity to loss-weight configurations in our multi-task setting. These results confirm that model performance remains stable across a broad range of weight perturbations, thereby empirically supporting the theoretical rationale described above.

2. To more thoroughly examine the effect of loss weighting, we expanded the set of loss-weight combinations beyond the default equal-weight setting. Specifically, we evaluated an enlarged grid of parameter combinations comprising two categories: (i) configurations with varied loss-term weightings, including perturbations around the default weights and schemes designed to balance the contributions of different loss components, (ii) configurations in which all loss weights remained within the same order of magnitude. As an illustrative example, we present the loss curves for a representative balanced configuration, as shown in Figure below, demonstrating that the individual loss terms converge to comparable numerical ranges during training and indicating effective balancing across objectives.



Concretely, instead of treating all weight combinations equally, we grouped configurations according to the relative magnitudes of the resulting loss terms during training. Configurations were categorized based on whether the maximum loss term exceeded the minimum loss term by more than a fivefold difference. This criterion allowed us to distinguish relatively balanced settings from those with more pronounced loss-scale disparities. The evaluated loss-weight combinations included contrastive loss weights (0.05, 0.1, 0.2, 1.0), classification loss weights (0.05, 0.1, 0.2, 1.0), KL-divergence weights (1.0, 2.0, 5.0, 10.0), and regularization weights (1.0, 2.0, 5.0, 10.0), as summarized in (**a**).

All configurations were evaluated on the same three datasets used in the main analysis. For each group, we compared the resulting tumor state assignments with those obtained under the default equal-weight setting by computing the ARI and NMI. As shown in the Figure (**b–d**), both ARI and NMI remain consistently high across datasets HN257 (**b**), HN272 (**c**), and P15 (**d**), indicating that even configurations with up to fivefold differences in loss magnitudes do not substantially alter the learned representations or state assignments. These results have been included in **Supplementary Fig. 2**.



Reviewer #4 (Remarks on code availability):

N/A.

Reviewer #5 (Remarks to the Author): Clinical expert in breast cancer and biomarkers; replaces Reviewer #4

This manuscript introduces EmitGCL, a deep-learning framework that uses knowledge-aware graph contrastive learning on paired single-cell omics data to predict future cancer metastasis. The model demonstrates strong performance in identifying rare metastatic cell types, proposes new biomarkers, and validates a key transcription factor via computational and experimental approaches. While the study addresses an important clinical problem and shows technical sophistication, major limitations in clinical translatability, model generalizability, and mechanistic depth prevent it from meeting the high standards of this journal. Substantial additional validation is required before publication can be recommended. Issues that need to be addressed:

1. The study cohort is underpowered, with only 28 patients spanning six highly heterogeneous cancer types. This limited sample size is inadequate for reliably training a complex deep learning model, raising concerns about overfitting and generalizability. Substantial expansion of patient numbers within each cancer type is required to validate model robustness.

Response: We fully agree that patient-level supervised classification using a small number of samples would indeed risk overfitting. In fact, this was precisely one of the key methodological challenges we encountered during study design. To address this limitation, our framework was intentionally developed as a cell-level representation learning model rather than a patient-level supervised classifier. Each individual sample is processed independently, and no patient-level labels are provided during training. The model operates in an unsupervised (i.e., contrastive and

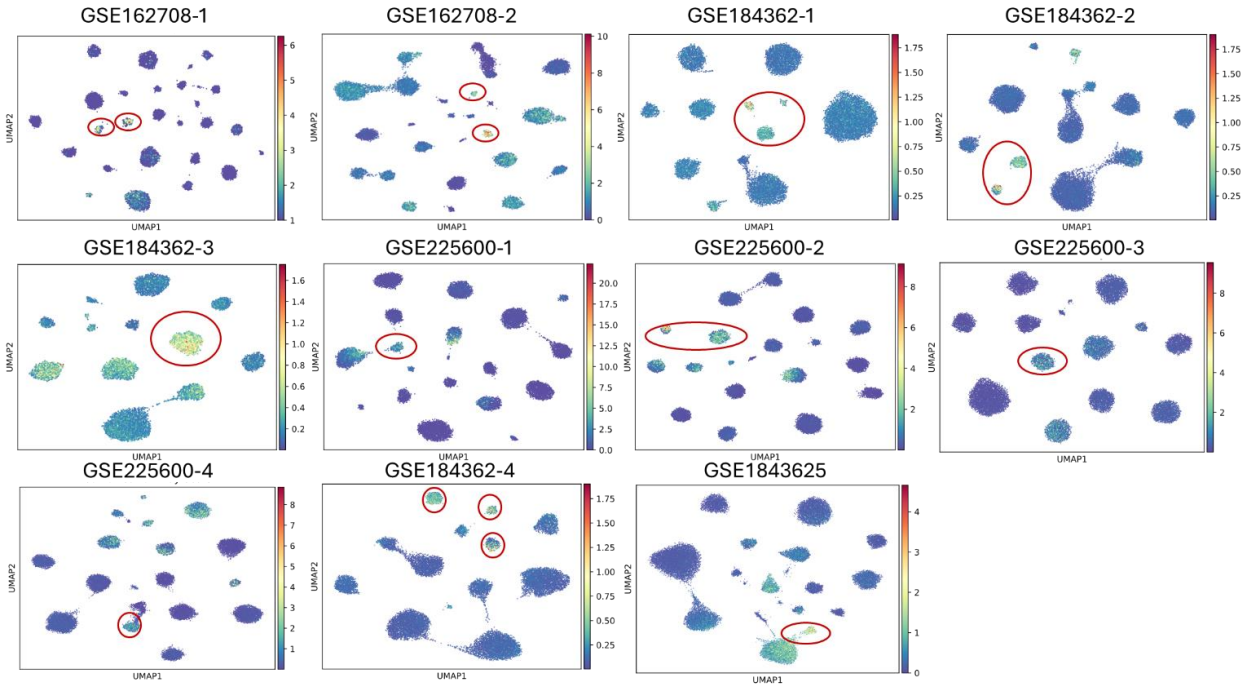
structural) learning paradigm, thereby reducing reliance on sample-level supervision and mitigating overfitting risks associated with limited cohort sizes. Importantly, we evaluated generalizability by applying the same trained model, without parameter re-tuning, to previously unseen samples. Across independent datasets, tumor state assignments remained consistent and biologically interpretable, supporting the robustness and transferability of the learned representations. We further note that our primary mechanistic analyses focused on breast cancer, where paired primary–metastatic datasets are relatively more abundant. This allowed us to perform patient-level stratification analyses (e.g., metastasis occurrence and future metastasis risk). In contrast, for cancer types with limited paired data, we intentionally refrained from making cancer-type specific biomarker claims. Instead, we focused on identifying mechanisms consistently observed across available breast cancer cohorts and validated these as potential markers for future metastasis.

Meanwhile, as demonstrated in the additional experiments conducted in response to Reviewer #1 (Comment 1) and Reviewer #3 (Comment 1), the model exhibits a low false-positive rate and high reproducibility across independent evaluations. These results further support the robustness and generalizability of our framework.

Finally, to further address concerns regarding sample size and generalizability, we performed an additional systematic search of publicly available datasets and identified 11 paired datasets that met our inclusion criteria:

Dataset ID	Cancer type	Site
GSE162708	Pancreas	Pancreas -> liver (2)
GSE184362	Papillary thyroid carcinoma (PTC)	PTC -> Lymph node (5)
GSE225600	Breast	Breast -> Lymph node (4)

We applied EmitGCL to these newly identified datasets without modifying model parameters. Across all cohorts, tumor state assignments remained consistent, and key performance metrics were stable. These results provide additional evidence that the robustness of the model is intrinsic to its design and is not sensitive to changes in sample composition.



2. The clinical validation relies predominantly on a prediction case from a single patient's lymph node data. This isolated success may be coincidental and does not demonstrate general applicability. It is essential to include predictions from multiple independent patients to systematically assess the model's reproducibility and clinical utility.

Response: We thank the reviewer for this important comment. Unfortunately, currently available public datasets do not contain additional paired single-cell datasets with longitudinal metastasis timing information, which limits the possibility of expanding validation at the single-cell paired level. To address this limitation, we incorporated an independent large-scale bulk RNA-seq cohort comprising 420 patients with detailed clinical follow-up data. Importantly, this cohort includes longitudinal records of metastasis occurrence and timing, enabling stratification of patients into those who subsequently developed metastasis and those who remained metastasis-free. Using the biomarkers identified by our model, we evaluated their predictive performance in this independent cohort (**Manuscript Fig. 4f**). The results demonstrate that our framework effectively distinguishes patients who later developed metastasis from those who did not, thereby providing population-level validation of the model's clinical relevance.

3. For the proposed biomarkers HSP90AA1 and HSP90AB1, the study lacks functional mechanistic investigation. Correlation data alone are insufficient to establish their causal roles in metastasis. Functional experiments directly testing their impact on migration, invasion, or other metastatic phenotypes are necessary.

Response: To functionally assess the role of the HSP90 isoforms HSP90AA1 and HSP90AB1 predicted by EmitGCL, we performed scratch-wound migration assays in EO771 breast cancer cells treated with the HSP90 inhibitor STA-9090 (ganetespib), which targets both HSP90 family

members including HSP90AA1 and HSP90AB1 by competing with the ATP binding site thereby blocking Hsp90 interaction with its co-chaperones and preventing Hsp90 from properly folding its client proteins. Treatment with 10 nM STA-9090 for 30 hours had minimal effect on cell migration, whereas 100 nM significantly reduced wound closure (**Manuscript Fig. 4d**).

To control for potential cytotoxic effects, parallel viability assays were performed under identical treatment conditions (10 and 100 nM, 30 hours). Treatment with 100 nM STA-9090 resulted in a modest decrease in cell viability (~10%). After normalizing wound closure measurements to viability, HSP90 inhibition remained significantly associated with a reduction in migratory capacity, indicating that the observed effect is not attributable to reduced cell viability (**Manuscript Fig. 4e**).

These results demonstrate that pharmacological inhibition of HSP90 suppresses breast cancer cell migration, providing functional evidence that HSP90AA1 and HSP90AB1 contribute to metastatic phenotypes. This finding is consistent with prior studies (PMIDs: 32236631, 33672529, 37958410) and supports the mechanistic relevance of EmitGCL-predicted metastasis-associated biomarkers.

4. The observed phenotype cannot be strictly attributed to YY1 loss without demonstrating that re-introducing wild-type YY1 into knockout cells restores migratory capacity, thus establishing definitive causality.

Response: We provide orthogonal and complementary lines of evidence supporting a causal role of Yy1 in breast cancer metastasis. Specifically, loss-of-function experiments demonstrated that YY1-knockout cells exhibit reduced migratory capacity, while gain-of-function experiments showed that EO771 cells with stable Yy1 overexpression display significantly enhanced lung metastatic colonization in vivo following tail-vein injection into C57BL/6 mice (**Manuscript Fig. 5j, k and Supplementary Fig. 9e**). Together, these results consistently indicate that Yy1 directly regulates metastasis and metastasis-associated phenotypes.

We fully agree that a rescue experiment reintroducing wild-type YY1 would provide the most direct demonstration of causality. However, YY1-knockout cells were generated using a pooled CRISPR/Cas9 strategy rather than single-cell clonal selection. As a result, individual YY1-KO clones were not isolated or preserved, and the heterogeneous edited population could not be used for rescue experiments that require a clone-defined genetic background.

5. To study the function of YY1 in mouse models, additional experiments should be conducted, such as comparing the lung colonization of YY1 overexpressing cells with control cells, to provide stronger support for its role as a therapeutic target.

Response: Thanks. To directly compare lung colonization between Yy1-overexpressing and control cells, we performed additional in vivo experiments to assess the role of Yy1 in metastatic colonization. EO771 breast cancer cells were engineered to stably express firefly luciferase and either an empty vector (control) or a Yy1 overexpression construct. Successful Yy1 overexpression was confirmed by western blot (Supplementary **Fig. 9e**). In addition, control and

Yy1-overexpressing cells were introduced into C57BL/6 mice via tail-vein injection (n = 6 per group). Lung metastasis was assessed 15 days post-injection using in vivo bioluminescence imaging with an IVIS Lumina X5 system. Mice injected with Yy1-overexpressing cells exhibited significantly higher metastatic burden in the lungs compared to control mice, as demonstrated by total photon flux measurements within the thoracic region (**Manuscript Fig. 5j and k**). Statistical comparisons were performed using Welch's t-test. These results demonstrate that Yy1 overexpression enhances in vivo lung metastatic colonization, providing functional evidence supporting Yy1 as a driver of breast cancer metastasis and a potential therapeutic target.

6. All analyses are based on retrospective data with known outcomes, creating risk of "answer leakage." The model may simply identify features of already-metastatic patients rather than genuinely predicting future metastasis. Prospective validation on samples from treatment-naive, non-metastatic patients with long-term follow-up is crucial.

Response: EmitGCL is not trained using patient-level outcome labels, nor does it directly optimize future metastasis endpoints. Instead, the framework performs unsupervised, cell-state level inference using only matched primary and metastatic site information, contrastive learning objectives, and biological priors. No future clinical outcomes are provided to the model during training, minimizing the risk of answer leakage.

A key aspect of our approach is the identification of occult metastatic cells, cell populations that were clinically undetectable at the time of sampling but later confirmed to be metastatic. These cells represent a biologically meaningful proxy for "future metastasis" rather than retrospective outcome labeling. To further mitigate concerns of retrospective bias, we performed multiple independent validations, including false-positive evaluation in non-metastatic patients, leave-one-dataset-out analyses, and biomarker validation across five independent cohorts comprising 420 patients (**Manuscript Fig. 4f**). Together, these analyses support that EmitGCL captures metastasis-associated cellular programs rather than merely recapitulating known outcomes.

We acknowledge that prospective validation will be essential for clinical translation and have clarified this point as a limitation and future direction in the revised manuscript.

Reviewer #5 (Remarks on code availability):
Coding is not my area of expertise. Please prioritize opinions from experts in code-related fields.

Response: Please refer our response to Reviewer #2 regarding code availability.

Reviewer 2:

The authors present a substantial amount of work to address the reviewers' comments. My outstanding points have been addressed, including code availability and documentation and a normal control. However, I still would like to mention figure presentation: although they have addressed in part my previous point on Figure 4e (now Figure 4g), this panel was simply the most obvious example of a more systematic underlying problem that runs across all figures. The issues highlighted in that panel are still reflected in several other panels (e.g., Figure 2e, f, g, h; Figure 3b, c; Figure 4c, g, j). As we have seen in previous rounds of revision, the post-processing of these panels introduces mistakes and omits genuinely informative aspects of the plots, such as axis ranges (apart from specific cases like UMAP representations, where this is appropriate). I would recommend that the authors work on this aspect prior to publication.

Response: We thank the Reviewer and are glad the previous points have been addressed. We agree that post-processing should not introduce errors or omit informative features. For Figure 2e–h, Figure 3b,c, and Figure 4c,g,j, we have re-exported the panels directly from the original plotting software with minimal manual editing, restored the full axis ranges and tick labels, and verified all values against the Source Data file. UMAP panels were left unchanged, as axis ranges are not informative there. We thank the Reviewer for helping us improve the figures prior to publication.