Classifying epithelial-mesenchymal transition states in single cell cancer data using large language models

Shi Pan[1], Eloise Withnell[1], Maria Secrier[1,*]

[1]UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK

* To whom correspondence should be addressed (m.secrier@ucl.ac.uk)

Abstract

Epithelial–mesenchymal plasticity plays a significant role in various biological processes including tumour progression and chemoresistance. However, the expression programmes underlying the epithelial–mesenchymal transition (EMT) in cancer are diverse, and accurately defining the EMT status of tumour cells remains a challenging task. In this study, we employed a pre-trained single-cell large language model (LLM) to develop an EMT-language model (EMT-LM) that allows us to capture discrete states within the EMT continuum in single cell cancer data. In capturing EMT states, we achieved an average Area Under the Receiver Operating Characteristic curve (AUROC) of 90% across multiple cancer types. We propose a new metric, ADESI, to aid the biological interpretability of our model, and derive EMT signatures liked with energy metabolism and motility reprogramming underlying these state switches. We further employ our model to explore the emergence of EMT states in spatial transcriptomics data, uncovering hybrid EMT niches with contrasting potential for antitumour immunity or immune evasion. Our study provides a proof of concept that LLMs can be applied to characterise cell states in single cell data, and proposes a generalisable framework to predict EMT in single cell RNA-seq that can be adapted and expanded to characterise other cellular states.

MAIN

The epithelial-to-mesenchymal transition (EMT) is a crucial process in cancer progression, enabling epithelial cancer cells to acquire mesenchymal properties, thereby enhancing their migratory and invasive capabilities, ultimately facilitating tumour metastasis and posing significant challenges for therapeutic interventions[1,2]. Increasingly, EMT is recognized not as a binary event but as a continuum of hybrid states that the cells explore as they transition from a fully epithelial to a fully mesenchymal phenotype[3]. This inherent plasticity challenges the traditional on-off model of classical epithelial and mesenchymal markers employed in EMT analyses such as E-cadherin, vimentin, or fibronectin, and suggests a more complex model of cancer metastasis[4]. Previous studies from our group and others have demonstrated that intermediate states within the EMT continuum can be captured successfully, are relatively

stable and have potential clinical relevance[5-7]. However, the EMT trajectories taken by cancer cells vary depending on the cancer tissue and various external stimuli[8]. This complicates the identification of hybrid states in the process in a manner that is generalisable across cancer types, and limits our ability to explore these states outside a very tightly controlled in vitro/in vivo setting.

Single-cell RNA sequencing (scRNA-seq) has emerged as a critical tool for dissecting the complexities of biological systems at the resolution of individual cells, significantly advancing our understanding of tissue biology. When it comes to characterising EMT, scRNA-seq has offered nuanced insights into the transcriptional landscapes that underpin the dynamic interplay between epithelial and mesenchymal states[8,9]. However, transferring this knowledge to other systems, in particular in order to explore EMT across human cancer tissues at larger scales, remains limited. Since EMT programmes exhibit a high degree of variability and non-linearity[10], linear models that attempt to capture the full complexity of the transition likely struggle at identifying key EMT genes. This implies the need for extensive, high-quality data, and demands models capable of efficiently understanding systems-wide dependencies.

Generally speaking, capturing cell *state* (e.g. EMT, stemness, senescence etc) is challenging and the field is still finding its way around this complex question. On the other hand, significant advancements have been brought by AI, particularly foundation models, to the challenge of capturing cell *type* (e.g. T cells, B cells, cancer cells etc). Methods like scBERT[11], scGPT[12], Cell2Sentence[13], GenePT[14] or scFoundation[15] have leveraged large-scale language models such as BERT, GPT and other transformer-based architectures[16-18] to capture both short- as well as long-range dependencies underlying cellular phenotypes in single cell data. This offers a promising avenue not only for identifying cell subgroups but also for enhancing our understanding of their regulation in different contexts and conditions.

Owing to the self-supervised nature of large-scale single cell language models and their training on extensive, unlabelled single cell RNA-seq (scRNA-seq) data, we speculate that they have assimilated, to some extent, the ubiquitous patterns underlying cell type and function, e.g. naïve T cells versus exhausted T cells versus cancer cells etc. These could potentially be adapted to capture cell state. In this study, we exploit this concept to develop EMT-LM, a generalisable classifier of EMT states in single cell cancer data. We achieve a 90% average accuracy rate in predicting five distinct EMT states in single cells across breast, lung, prostate and ovarian cancers. Furthermore, the predictions of our model were highly correlated with observed EMT programmes in independent cancer cell lines and spatial transcriptomics experiments - suggesting our model captures common underlying circuitry for different cancers at multiple EMT time points. Finally, we use spatially profiled cancer tissue to gain new insights into the EMT states captured by our model. Specifically, we show distinct interplay between these states and the tumour microenvironment, with some hybrid states localised in niches that favour either immune evasion or antitumour immunity. Our study underscores the capability of LLMs to predict cell state in single cell data, as well as providing a bespoke EMT model that can be applied in a range of biological scenarios.

RESULTS

EMT-LM, a large language model that predicts EMT states in single cell cancer data

To facilitate a better understanding of EMT, we explored whether a pre-trained language model can predict EMT status from single cell RNA sequencing data. To define the ground truth transcriptional states along the EMT continuum in different cancer tissues, we employed data from Cook and Vaderhyden[8]. This study encompasses single cell RNA sequencing from breast (MCF7), lung (A549), prostate (DU145) and ovarian (OVCA420) cancer cell lines undergoing EMT, profiled at 0 days, 8 hours, 1 day, 3 days and 7 days after stimulation with TGF-β. We labelled cells sequenced at the 0-day sampling timepoint as epithelial (E), and ulterior states as hybrid E/M (EM1 at 8 hours, EM2 at 1 day, EM3 at 3 days) and mesenchymal (M, 7 days) (**Figure 1a**). Thus, this experimental setup provides a clean reference for the transcriptional profiles of cells occupying distinct states within the EMT continuum.

We utilised scBERT[11], a language model-based framework, as backbone to train a binary classification model for each EMT time point (**Figure 1b**). Learning from the experimentally measured longitudinal profiles of single cells undergoing EMT transformation under TGF-β stimulation, our model builds on the original single cell large language model (scBERT), integrating filtering layers and a fusion strategy to enhance the model's performance and Parameter-Efficient Fine-Tuning (PEFT)[19] to reduce computational demands (**Figure 1b**). A model was trained for each of the five states. In the second phase of training, we integrated these models through a fusion net to obtain the final training results (**Figure 1c**). This process effectively shifts the initial unsupervised learning approach into a supervised one, enabling model optimization via gradient descent. Our model not only provides a classifier for distinct EMT states in single cell data, but also allows the exploration of the gene regulatory programmes underlying these states through a new gene ranking approach, ADESI (**Figure 1d**). In the subsequent sections, we demonstrate that EMT-LM can both recapitulate known biology as well as uncover new biology in external datasets.

EMT-LM behind the scenes - gradual learning of EMT states

Before analysing the performance of the model, it is important to understand the patterns that the model is learning when attempting to classify EMT states. To this end, we explored the changes of the learned latent space. **Figure 2a** uses principal component analysis (PCA) to illustrate the five embedding spaces during the two proposed training phases: the original data distribution, the feature space of the data after being processed by the original scBERT[11], the feature space generated by scBERT after PEFT, the feature space of the proposed scMultiNet, and finally, the feature space resulting from the EMT-LM fusion. The evolution of the model's learned representations shows that the modifications to the original backbone model progressively improve the detection of EMT patterns. In the original gene expression distribution (**Figure 2a** - raw data), cells profiled at different EMT timepoints cluster together and the separation is only driven by a cancer cell line effect (A549, DU145, MCF7, and

OVCA420) - thus PCA is not sufficient to capture EMT states from single cell data, as would be expected. The scBERT model usually provides a good embedding space for various biological problems without further fine-tuning. Yet it fails to elucidate salient features pertinent to EMT in our visualisation (**Figure 2a** - scBERT). The cell line batch effect is reduced; however, cells from distinct EMT sampling times remain extensively intermixed. This highlights a limitation of such pre-trained models when directly inferring specific biological tasks without fine-tuning: the pre-trained model may learn a robust feature space, but this space may not be explicitly related to the biological process of interest. Thus, the original pre-trained scBERT model, while powerful, was not ideally suited to distinguish EMT states. The embedding space from our scMultiNet approach demonstrates more distinct boundaries and tighter clustering between categories (**Figure 2a** - scMultiNet). This continuous spatial transformation demonstrates that the model has captured key feature signals related to EMT. Finally, our fused EMT-LM model shows a distinct separation of all EMT states (**Figure 2a** - EMT-LM).

Figure 2b uses PHATE[20] to further demonstrate how the changes in the embedding space align with an EMT continuum. In our final EMT-LM model, the baseline, unstimulated epithelial state appears farther away from the gradually transforming cancer cell space, as expected (**Figure 2b**, 1st panel). When removing the epithelial cells, the stimulated cell populations naturally arrange themselves from right to left (blue to orange) on an EMT continuum in the latent space of the EMT-LM model (**Figure 2b**, 2nd panel). Remarkably, although the model was trained using discrete labels <u>without explicit temporal information</u>, it successfully captured the sequential nature of the EMT process, suggesting its ability to learn the underlying dynamics from the data.

We applied Diffusion Pseudotime (DPT)[21] to the embedding space derived from our final model to infer cell state transition sequences in the form of a 'pseudotime' across the single cell data points and elucidate EMT trajectories (**Figure 2b**, 4th panel). The reconstructed EMT trajectory (**Figure 2b**, 5th panel) aligns with the real sampling times in the experiment from early (8h) to late (7d) time points, as expected, demonstrating that our model captures a distinct and faithful chronology of cell state transformation.

Interestingly, the 8h (EM1) time point appears most distinct, with samples branching into diverse but more closely aligned trajectories at subsequent timepoints. This might imply that the early 8h time point of the E/M transformation could potentially capture a state of maximal plasticity which then converges into more stable intermediate or mesenchymal-like states at later points in time.

EMT-LM model performance

We trained our EMT-LM model on 80% of randomly sampled single cells from the dataset and tested its performance in classifying EMT states on the remaining 20% randomly selected samples. After two training phases, our model achieved 90% averaged AUROC performance across all EMT states (**Figure 3a**). Furthermore, it outperformed widely utilised machine

learning algorithms such as K-nearest neighbours (KNN), which presented an average AUROC of 0.57, Random Forest with average AUROC of 0.75, and AdaBoost with AUROC of 0.83 (**Figure 3a**). Thus, using large language models considerably improves the performance of EMT state classification over standard methods.

In the EMT-LM model, the "0 days" state (untreated), as the category closest to a pure epithelial state, had the highest AUROC value of 0.93, followed closely by the "7 days" mesenchymal category (according to Cook and Vanderhyden[8]) with an AUROC of 0.91. Thus, our model can accurately recognize the two extremes within the EMT continuum. In addition, our model demonstrates considerable accuracy in classifying the hybrid E/M states EM1-EM3 captured at 8 hours, 1 day and 3 days, with AUROCs between 0.88-0.91 (Figure 3a).

To assess to what extent the performance of our model depends on the EMT programme of the cell of origin, the test set was divided to evaluate performance in different cell lines separately (Figure 3c-f). The best performance was observed in the breast and ovarian cancer cell lines MCF7 and OVCA420, respectively, with average AUROCs of 0.92 (Figure 3c-d), followed by the prostate cancer cell line DU145 with average AUROC of 0.90 (Figures 3e) and the lung cancer cell line A549 with average AUROC of 0.89 (Figure 3f). Overall, these performances suggest that our model is able to capture robust common patterns of EMT transformation across multiple cancer tissues.

Validation in inhibitor and stimulus removal data

We validated the model using additional experimental results conducted as part of the original study by Cook and Vanderhyden[8]: the addition of the EMT inhibitor LY364947 (a small molecule inhibitor of TGFBR1), and the removal of EMT-inducing stimuli. In both cases, we expected to see a reversion towards a more epithelial state, at least to a certain extent. The inhibitor cannot guarantee a full EMT inhibition, and therefore we would not expect a complete reversion to the 0-day state. However, our model does confirm that EMT is significantly reduced (Wilcoxon rank-sum test p-value <0.001) in the cell lines after treatment with the inhibitor (**Figure 4a-b**) and after removal of the TGF-β stimulus (**Figure 4c**, **Extended Data Figure 2a**). We observe expected EMT state changes in the removal data, most prevalently as a drop to an EM3 hybrid state at 8 hours and 1 day post-treatment, with the majority of cells reaching an epithelial state at 3 days (**Figure 4d**, **Extended Data Figure 2b**).

These results suggest that our model has not simply learnt a batch effect linked with the sampling of cells at different time points but rather is able to capture patterns underlying the true biology of these transitions.

Validation in external datasets

To further validate our model, we applied it to two external datasets where EMT progression has been tracked in vitro: MCF10A cells treated with TGF-β, sequenced at multiple time points, from Paul et al[22] and a pooled single-cell MCF10A and HuMEC cells undergoing spontaneous

or TGF-induced EMT from McFaline-Figueroa et al[9] (**Figure 5a-b**, **Extended Data Figure 1**).

In the MCF10A cells from Paul et al[22], the predicted values from our model showed a consistent increase, particularly across the T0-T4 time points, with time points 5-7 appearing to reach an EMT steady state endpoint (**Figure 5a-b**). Furthermore, the EMT-LM scores capture the EMT transition in a single metric that outperforms classical epithelial, partial EMT (pEMT) and mesenchymal markers commonly employed in the literature (**Figure 5b**). In the McFaline-Figueroa et al[9] experiment, it is noteworthy that our model predicts an increase in EMT in the TGF-β-stimulated, but not in the spontaneous model (**Extended Data Figure 1b**) - confirming that we are able to specifically capture TGF-β-driven EMT transformation. **Extended Data Figure 1b** also suggest that the level of EMT transformation achieved with TGF-β stimulation is higher than that in the spontaneous model described in the study. Overall, our model demonstrates generalisation capabilities across multiple datasets.

It is important to highlight that the model-predicted scores exhibit different scales across datasets. We interpret this as a reflection of the complexity of the EMT process. Although cells undergo EMT following stimulation, the specific timing of entry, the duration of this process, and the real-time length of EMT are all aspects that lack clear methods and frameworks for analysis and understanding. This phenomenon is observable only when the model predicts across different datasets, rather than merely clustering within a single dataset - which no other available method for EMT quantification is able to do currently.

We next sought to evaluate the model in an external EMT inhibition dataset from Cassier et al[23]. This dataset includes spatial transcriptomic slides from the endometrial tumours of two individuals before and after treatment with the EMT inhibitor Netrin-1. This inhibitor is expected to suppress EMT and is suggested by the authors to induce a reversion towards a more epithelial state. In both cases our model predicts a decrease in EMT after treatment with the inhibitor (**Figure 5c**), which is visible throughout the slides (**Figure 5d**). The fact that our model is able to capture this in endometrial cancer, which is not a cancer type it was originally trained on, suggests that it is learning an EMT programme that may generalise beyond breast, prostate, lung and ovarian cancers.

Gene programmes underlying TGF-β-induced EMT state transitions in cancer

Having ascertained that our EMT-LM model is capable of making reasonable predictions of EMT state, we next sought to understand the gene regulation patterns that the model has learned, which could help explain the driver events underlying EMT state switches in cancer. Traditional models rely on attention scoring to prioritise features from the model that could explain the model's behaviour. However, since LLMs are highly sensitive to small changes in gene expression, these prioritised features might not always translate in easily interpretable and actionable biology. Instead, approaches that can help extract important signals that also reflect major changes in gene expression can guide the identification of the master regulators of a specific programme amongst the myriad of small modulators. In this work, in order to explore

key EMT determinants autonomously and enhance interpretability, we leveraged traditional attention scores to develop the Attention-Driven Expression Significance Index (ADESI, see Methods). Specifically, we reweigh the attention scores based on the expression of the respective genes, which helps prioritise the most biologically meaningful changes underlying the described EMT states. **Figure 6a** illustrates the expression dynamics across EMT states for the genes deemed by ADESI to have the most significant impact on the prediction outcomes. Distinct patterns of gene expression regulation linked to the different epithelial, hybrid and mesenchymal states are clearly visible in the heat map, confirming that the model is capable of capturing relevant regulatory programmes in the data.

Thus, our ADESI metric can be used to derive gene signatures that help capture the different EMT states described. However, the number of genes included in each signature can vary, potentially leading to different outcomes in downstream analyses. To ensure the robustness of these gene signatures for subsequent analyses, we explored how varying the number of genes in the signature affects our ability to separate EMT states in the original dataset. We progressively increased the number of genes returned from either the raw attention metrics or ADESI score and calculated the mean Cohen's d score for each gene list within the original dataset (**Figure 6b**). The analysis indicates that a subset of 200 genes provides sufficient information to reliably distinguish the original EMT states in the dataset. Comparing two methods of scoring the signature, Gene Set Variation Analysis (GSVA) and scaled gene set scoring (see Methods), revealed that GSVA is superior in transferring patterns learnt from the model. Furthermore, the ADESI score demonstrates higher accuracy in predictions compared to the raw attention scores.

To further confirm the validity of the 200-gene signature (**Supplementary Table 1**), we applied it to the two EMT inhibitor datasets from Cook and Vanderhyden[8] and Cassier et al[23] which we had employed in previous experiments as well. To reiterate, we would expect to see a reduction in EMT upon inhibitor treatment and a reversion towards an epithelial phenotype in both datasets. Indeed, when applying the epithelial (E) signature derived from the ADESI metric, we observed an increase in this signature upon EMT inhibition (**Figure 6c-d**), suggesting a shift towards an epithelial state and demonstrating the biological relevance of this gene signature. Applying the 200-gene signatures for each EMT state (E, EM1-3, M) in the Netrin-1 inhibitor spatial transcriptomic dataset provided a more granular understanding of how these cell states are distributed throughout the tumour and how they shift upon treatment (**Figure 6e**). After EMT inhibition, we observe an increase in the E state in both Patient 1 and Patient 2, as expected and reported by the original study too. We also see a remarkable increase in the EM1 state in Patient 2 and generally clearer shifts to lower hybrid states compared to Patient 1 – suggesting that the success of Netrin therapy is highly dependent on the extent of EMT transformation in the original tumour.

Having verified the robustness of the gene programmes underlying distinct EMT states, we next sought to understand the key pathways upregulated for each EMT time point (**Supplementary Table 2**). At the 8h time point (EM1), the most significant gene sets related to oxidoreduction and ATP synthesis, including genes *COX6B1, COX5A* and *COX5B*,

271 components of the electron transport chain[24]. These genes underscore the essential role of
272 mitochondria in the metabolic reprogramming that accompanies EMT[25], likely due to a
273 heightened demand for ATP as cells increase their motility and invasiveness[26]. The elevation
274 of transcription regulators (e.g. *SNRPB*, *SNRPC*) aligns with the extensive transcriptional
275 reprogramming cells undergo during EMT[27]. Knockdown of *SNRPB* has been linked to reduced
276 EMT in non-small cell lung cancer[28]. Structural genes such as *TMSB4X*, which is involved in
277 actin polymerization, were also significantly involved, reflecting the cytoskeletal
278 reorganisation required for cell motility[29].

279 At the 1-day time point (EM2), we observed continued metabolic reprogramming, particularly
280 involving oxidoreduction-driven active transmembrane transport. This suggests an ongoing
281 need for efficient electron transfer and ATP generation. There was also enrichment in
282 extracellular exosome and vesicle activity. Extensive research has suggested that extracellular
283 vesicles play a crucial role in EMT plasticity[30]. Additionally, there was enrichment in the
284 supramolecular fibre organisation gene set, including genes such as *ARPC2, ARPC3,* and
285 *ACTG1*, which are involved in the formation and organisation of actin filaments, further
286 supporting cell structure and motility[31].

287 Notably, the key mesenchymal marker *VIM* was upregulated at the 3-day time point (EM3),
288 suggesting this hybrid state has some features of mesenchymal cells. Besides this, the top genes
289 were primarily related to structural integrity and cell migration, including genes involved in
290 the structural constituent of the cytoskeleton and the cell leading edge, such as *ACTB* and
291 *TUBA1B*. There was also continued enrichment of extracellular vesicles.

292 By the 7-day time point (M), we observed an enrichment of cadherin binding, mediated through
293 genes like *ITGB1* – which has been shown to be a critical effector in promoting metastasis[32].
294 This suggests that this state combines the adhesive properties of epithelial cells with the
295 migratory and invasive capabilities of mesenchymal cells.

296 Our overall observations also indicated high levels of ribosomal activity. EMT and ribosomal
297 biogenesis have been interconnected in recent studies[33], suggesting a coordinated regulation of
298 cellular plasticity and protein synthesis. In our study, we opted not to regress out ribosomal
299 biogenesis genes before employing the scBERT model, which is pre-trained on unscaled
300 scRNA-seq data with no filtering, however we note that this may play a role in the emphasis
301 of ribosomal biogenesis pathways in the top genes.

302 **The EMT-LM signatures capture distinct biological niches within tumours**

303 Beyond unveiling key checkpoints during cell state switches, the EMT gene signatures derived
304 above allow us to apply the knowledge learnt from the model to explore new aspects of the
305 biology underlying EMT in cancer. Since these states are expected to co-exist to a certain extent
306 within the tumour as cells gradually progress towards more advanced malignancy, we asked
307 whether we could investigate their emergence within individual tumours and how it is
308 influenced by the tumour microenvironment. To achieve this, we used a dataset consisting of

309 minutely profiled breast tumours via spatial transcriptomics, which provides a tissue-wide view
310 of the expression programmes active throughout a tumour, also capturing immune and stromal
311 components of the microenvironment (**Figure 7a**). We employed our previously developed
312 method, SpottedPy[34], to identify cellular niches where the tumour cells occupy distinct
313 epithelial, hybrid EM1-3 and mesenchymal states, and we interrogated their interplay with the
314 immune microenvironment (**Figure 7a**).

315 The early hybrid niches EM1 and EM2 displayed the strongest association with other cell
316 populations in the TME (**Figure 7b-c**). Cancer-associated fibroblasts, particularly of the
317 inflammatory type (iCAFs), tumour-associated macrophages (TAMs) and exhausted LAG3+
318 CD8 T-cells were significantly associated with the EM1 state, suggestive of an
319 immunosuppressive phenotype[35] (**Figure 7b**). A macrophage niche linked with antitumour
320 immunity (CXCL10[36]) and reduced inflammation (EGR[37], APOE[38]) was associated with the
321 EM2 state, potentially highlighting this cancer state as more amenable to immune recognition
322 and killing (**Figure 7c**). The epithelial niche appeared closely linked with proliferative tumour
323 and T cell regions, as would be expected, providing further validation for the gene lists derived
324 from the attention scores (**Figure 7e-f**). We additionally observed a gradient of decreasing
325 interactions from the EM2 to the M state (**Figure 7c-e**). The weaker associations seen for the
326 EM3 and M niches are indicative of a colder, more immune evasive environment, as might be
327 required at more advanced stages of transformation (**Figure 7e-f**). Overall, these results suggest
328 that the EMT niches develop within distinct microenvironments and emphasises the value of
329 treating EMT as discrete states.

330

331 **DISCUSSION**

332 In this work, we introduce EMT-LM, a language model capable of identifying multiple EMT
333 states in single cell cancer data. We demonstrate superior performance compared to traditional
334 unsupervised and supervised approaches, as well as to the use of classical EMT markers to
335 identify such states. We further employ this model to explain regulatory programmes
336 underlying EMT switches, demonstrating enhanced model interpretability. Finally, applying
337 the EMT signatures defined by our model to spatially profiled tumours unveils biological
338 niches with distinct tumour microenvironment composition, suggesting unique and shared
339 adaptation strategies as cancer cells progress towards a mesenchymal state.

340 Unlike traditional methods, this study introduces a pre-trained foundation model that is able to
341 capture complex dependencies in pathway regulation, focusing on more than 16,000 genes.
342 The resulting model achieves classification results and transferability surpassing baseline
343 models such as KNN, decision trees or gradient boosting. We suggest our approach diverges
344 significantly from traditional methods by incorporating language models, enabling our model
345 to not only focus on a few upregulated or downregulated genes but also to delve deeper into
346 the subtle interconnections between these changes. We highlight that the ADESI method used
347 to explain gene contributions can accurately capture expected EMT patterns, as demonstrated
348 in the sensitivity analysis and downstream validation in inhibitor datasets. This shows that

ADESI scores can be used to explore the underlying biology and therefore suggests that LLMs can be expanded not just for prediction problems, but also to gain deeper biological insights.

Importantly, it was not guaranteed that it would be possible to predict EMT states with high accuracy at different time points, considering the process could be solely continuous, without stable intermediate states. This is an ongoing question in the field, with conflicting studies suggesting epithelial-mesenchymal plasticity is either continuous or not[39-44]. Our results suggest that there is a significant transcriptional change between these time points that is sufficiently recognisable to suggest the existence of some relatively stable intermediate states. Furthermore, our method can output both a continuous score and discrete states, offering a new way to describe cell states. While most methods focus on one aspect, this approach acknowledges that the majority of discrete states inherently have aspects of continuity. This dual capability provides a more comprehensive understanding of cell states.

It is worth noting that EMT-LM is more generalisable across distinct datasets and cancer types compared to traditional approaches. To our knowledge, this is the only supervised method for classifying EMT to date. Thus, compared to unsupervised pseudotime approaches, the EMT state identification by our model is not biased by the ability of a new cohort to capture a broad range of EMT states. For instance, if all cells in a newly analysed dataset were found in the same EMT state (e.g. EM1), our model would be perfectly capable of assigning just this state, whereas a pseudotime model would try to enforce a continuum of states thereby ending up potentially incorrectly classifying at least 2-3 states across a small range of phenotypic variation that is not strong enough to push the cells into a different state. In addition to classifying states in new data, EMT-LM also provides a regression score which allows the transfer of continuous (pseudotime-like) labels from one dataset to another, "pseudotime transfer learning". This approach enables the reuse of pseudotime information, enhancing the efficiency and applicability of pseudotime-like analysis across different datasets.

Single-cell language models (scLLMs) show potential for various biological applications, yet significant obstacles remain. For instance, pre-trained models often encode mixed signals from diverse biological processes. When targeting a complex biological process that lacks a single dominant signal (e.g. EMT in this study), classical single-cell language models may struggle to deliver accurate inferences. Thus, this paper introduces PEFT and two training stages to adapt the feature space of the original pre-trained model. The proposed EMT-LM capitalises on the strengths of pre-trained models while also independently optimising and designing for specific biological processes. This method broadens the application of single-cell pre-trained language models. It is important to note that both full-parameter fine-tuning and Partially-Extracted Feature Tuning (PEFT)[19] can effectively modify the feature space. However, PEFT is more efficient in terms of it requiring less data and fewer computational resources for comparable outcomes. Furthermore, the PEFT algorithm applied in the proposed pipeline specifically targets low-rank adaptation of the original mapping, ensuring that the model retains its understanding of the pre-trained feature space and thereby enhancing its generalizability.

Our model was built based on transcriptional profiles of cancer cells undergoing EMT under TGF-β stimulation, and thus only captures internal changes under this specific stimulus and in

the absence of any tumour microenvironment factors. While TGF-β is amongst the most common EMT drivers and more likely to be generalisable across different cancer types, future work will focus on expanding the capability of our model to capture a broader array of EMT programmes profiled under different stimuli.

Furthermore, our model has been trained on data representing the five distinct EMT states independently, without knowledge of the temporal connection between them. This was done by design, as hybrid E/M states are generally believed to be parallel rather than subsequent events in tumour evolution and thus different datasets might just capture a single such state. Our model allows for this flexibility, and the EM1-EM2-EM3 states can be considered as equally possible end points a priori. In the future, models that incorporate a temporal component should be explored in order to understand if they more accurately represent EMT switches in cancer.

Cook and Vanderhyden emphasise the plasticity of EMT, which is highly dependent on context, both tissue and stimulus. They, and others[9,39], have rightfully pointed out that identifying a unifying programme for EMT is highly challenging. However, they do suggest some higher order logic may exist and that it may be captured with more complex methods. Our model does not contradict their results, but rather corroborates that it may be possible to identify elements of this higher order logic, i.e. to capture some generic processes underlying EMT - within the limits of similar tissue and stimulus contexts to the ones from the training data. Notably, our EMT-LM model does not simply recapitulate classical EMT drivers. There would be no need for a model that simply captures what is already well characterised in the literature. Instead, it seems to capture broader RNA metabolism, cellular energetics and motility features that accompany EMT switches which may be more generalisable across different cancer types and can act as a proxy readout for this complex process. In the future, it will be important to investigate how these higher-order processes feed into EMT regulation and identify the key bottlenecks in these pathways.

Our analysis demonstrates the potential of using pre-trained large language models to understand complex regulatory programmes such as those underlying EMT. We show that LLMs offer additional advantages to analysing scRNA-seq data compared to standard approaches, by capturing novel relationships between genes and cell states that may be clinically relevant. We further provide a flexible and expandable core model, scMultiNet, which can be adapted to develop models that capture a variety of cellular programmes in single cell data from healthy human cohorts as well as in disease.

METHODS

Training dataset

We sourced gene expression data from Cook and Vanderhyden[8], made available at: https://drive.google.com/drive/folders/1SIEIf7UswTv_0S6TypYsaRzMcfkfsgji (GSE147405). The dataset encompasses 11,119 cells profiled through scRNA-seq from A549 (lung), DU145 (prostate), MCF7 (breast) and OVCA420 (ovarian) cancer cell lines, subjected to TGF-β induced EMT, paired with EMT sampling times ranging from 0 days (epithelial state), 8 hours, 1 day, 3 days (hybrid states labelled EM1, EM2, EM3, respectively) to 7 days (closer to a fully mesenchymal state). These cell lines maintained an epithelial morphology and have been demonstrated to undergo EMT in previous studies. TGF-β was selected as an inducing factor due to its proven efficacy in promoting EMT across various cell lines. Morphological changes consistent with EMT in each cell line have been documented in Cook and Vanderhyden[8]. In our training, we experimented with models trained on all cancer types and models excluding specific cancer types to evaluate the model's performance.

Learning discrete cell states with a PEFT single cell foundation model

In this study, the proposed EMT-LM consists of 5 expert models developed based on scBERT[11] and a fusion model to construct the final outcome. The proposed pipeline includes two training stages where training stage I aims to modify the embedding space and to bridge the gap in dataset sizes between pre-training and the focused EMT problem. We name the expert models 'scMultiNet'. Training stage II aims to train a fusion net to produce the final outcome. The gene expression profile data was filtered as required by scBERT and encoded into 200-dimensional vector spaces for each expert network (scMultiNet-E, scMultiNet-EM1, scMultiNet-EM2, scMultiNet-EM3, or scMultiNet-M). The selected pre-trained foundational model (scBERT or other potential single-cell language model in the future) is adept at analysing single-cell RNA sequencing (scRNA-seq) data, aiming to generalise cell state predictions across various tissues and patient cohorts. As a pre-trained single-cell language model, the scBERT model underwent extensive pre-training on a corpus of approximately 100,000 scRNA-seq samples with gene2vec and a vocabulary of around 16,000 genes. The proposed method follows all the preprocessing steps in scBERT to make sure the proposed model can assign positional significance to genes.

We name our expert models 'scMultiNet' as we employ a simple multiplication layer consisting of fine-tuned language models. The proposed model considers the attention vector of the language model as a weight of raw gene expression profile data. We employed Low-Rank Adaptation (LORA), a Parameter Efficient Fine-Tuning (PEFT) method, in our training stage I to adapt the embedding space language model's architecture for interpreting single-cell gene expression profiles (**Figure 1**). PEFT enhanced the model's adaptability to the targeted dataset while minimising the risk of overfitting. The design of scMultiNet ensured the model's focus remained on genes relevant to the EMT process, filtering out those with low expression

463 or tangential relevance. Genes exhibiting minimal expression were removed through
464 normalisation and selective filtering to mitigate their potential to skew the model's learning
465 outcomes.

466 For the target tasks, we employed a Multilayer Perceptron (MLP) as classifier or regressor. In
467 all training stages, we utilised a one-dimensional convolution for dimensionality reduction and
468 a three-layer neural network to transform gene feature vectors into probabilistic cell type
469 identities. Training stage I applied a binary classification network for each category that was
470 regulated by a learning rate of 1e-4 and a cosine step scheduler. Subsequently, we concatenate
471 the feature vectors extracted from binary classification models and feed into the fusion network
472 in training phase II. Feature vectors from the binary models were concatenated, and a
473 convolutional layer was incorporated to streamline the dimensionality of the attention map.
474 Three linear layers were deployed for the final classification task.

475 **Model validation**

476 Validation of the EMT-LM model's performance was carried out by randomly splitting the
477 original data into training (80%) and testing (20%) sets, focusing on assessing the model's
478 accuracy. The model's feature space was visualised using Principal Component Analysis (PCA)
479 with Scanpy and Potential of Heat-diffusion for Affinity-based Transition Embedding
480 (PHATE) with its original implementation from the Krishnaswamy lab[20] to compare the
481 changes before and after training, demonstrating that the training process effectively captured
482 the key information related to EMT.

483 During training, we randomised the order of all sampling time points and discretized them into
484 one-hot encodings. By not introducing temporal information during training, if the learned
485 embedding space exhibited clear time-series associations, it would indicate that the model had
486 learned the correct information (results shown in **Figure 2**).

487 To further validate the robustness and broad applicability of our approach, we tested our model
488 on two unseen single-cell datasets based on the same TGF-β stimulation but with different
489 experimental designs. These datasets are available for download from GSE194019 and
490 GSE114687. The model's performance was analysed by comparing the categories provided by
491 the datasets with the continuous scores inferred by the model (results shown in **Figure 3**).

492 **Comparison with traditional machine learning methods**

493 The evaluation process involved an initial dimensionality reduction via PCA, followed by
494 training with machine learning algorithms (KNN, Random Forest and Adaboost from SciPy).
495 A unified assessment using the average Receiver Operating Characteristic (ROC) curve was
496 employed to report the classification performance of the proposed model.

Attention-Driven Expression Significance Index (ADESI) scoring

We introduce the Attention-Driven Expression Significance Index (ADESI), a novel method designed to enhance the interpretability of complex biological datasets. Traditional single-cell language models use attention visualisation to demonstrate model interpretability and versatility. However, this may fail to capture changes between states that would be considered impactful enough to be biologically meaningful. ADESI addresses this issue by leveraging the scMultiNet architecture and Z-scores to provide biologically relevant gene weightings. Initially, mirroring traditional methods, we extracted attention scores from each expert model relevant to the designated category. These scores were then multiplied by the original expression levels of the respective genes to derive the scMultiNet output. The average score for each gene was calculated in the final classified category and the top K genes were prioritised. This design is used to aid researchers in understanding the genes that the model's attention mechanism focuses on, which could be potentially employed as markers of these states.

Spatial transcriptomics analysis

Breast cancer 10x Genomics Visium slides were obtained from Barkley et al[45] (slides 0-2), from the 10x Genomics website (https://www.10xgenomics.com/) (slides 3-5) and Wu et al[46] (slides 6-12). We combined the three spatial transcriptomics datasets into a common anndata Python format for analysis. We analysed a total of 32,845 spatially profiled spots, and retained spots if they exhibited at least 100 genes with at least 1 count in a cell, had more than 250 counts per spot and less than 20% of total counts for a cell which are mitochondrial. Endometrial carcinoma Visium slides were obtained from Cassier et al[23]. Filtered matrices were loaded, merged per patient and spots with fewer than 1,000 detected genes were removed. Pre-processing and normalisation were conducted using the ScanPy package. To more precisely identify tumour cells we employed the copy number inference tool STARCH[47] and only kept spots that showed evidence of copy number changes, which are likely to be tumour specific. SpottedPy[34] was used to identify EMT niches and assess the spatial relationships.

Gene signature analysis

We used Gene Set Variation Analysis (GSVA) calculated using the GSEApy package[48] to score the tumour samples with the EMT-LM gene signatures. We categorised the high attention genes based on their upregulation or downregulation in the initial dataset for each time point. We calculated a composite score by subtracting the GSVA score of downregulated genes from that of the upregulated genes. The scaled score for comparison was calculated by subtracting the Z-score of downregulated genes from the Z-score of the upregulated genes. Cohen's d score was used for assessing how well the gene sets distinguish each cluster. Cohen's d was computed by assessing the difference in means between the groups, standardised by their pooled variability. G:profiler[49] was used for gene set enrichment analysis.

Model building, statistical analysis and data visualisation

The EMT-LM model was built with PyTorch for transformer blocks, and pytorch-lightning for efficient training. The analyses of the results were conducted in Python and R. Groups were compared using a two-sided Student's t-test, Wilcoxon rank-sum test or ANOVA as required. Correlations were calculated using Spearman's rank order correlation. Graphs were generated using the Seaborn and Matplotlib Python packages.

DATA AVAILABILITY

The results published here are based upon publicly available data made available at the Gene Expression Omnibus (GEO) by Cook et al[8] (GSE147405), Paul et al[22] (GSE194019) and McFaline-Figueroa et al[9] (GSE114687). The breast cancer spatial transcriptomics data employed in the study were downloaded from Barkley et al[45] (GSE203612), Wu et al[46] and https://www.10xgenomics.com/. The endometrial cancer spatial transcriptomic slides were downloaded from Cassier et al[23] (GSE225691).

CODE AVAILABILITY

The EMT-LM package can be found at the following GitHub repository: https://github.com/secrierlab/EMT-LM. The scMultiNet implementation can be accessed within this repository at: EMT-LM/scLLM/Models/scMultiNet. All the code developed for the purpose of this analysis is deposited at: EMT-LM/Experiment.

FIGURES

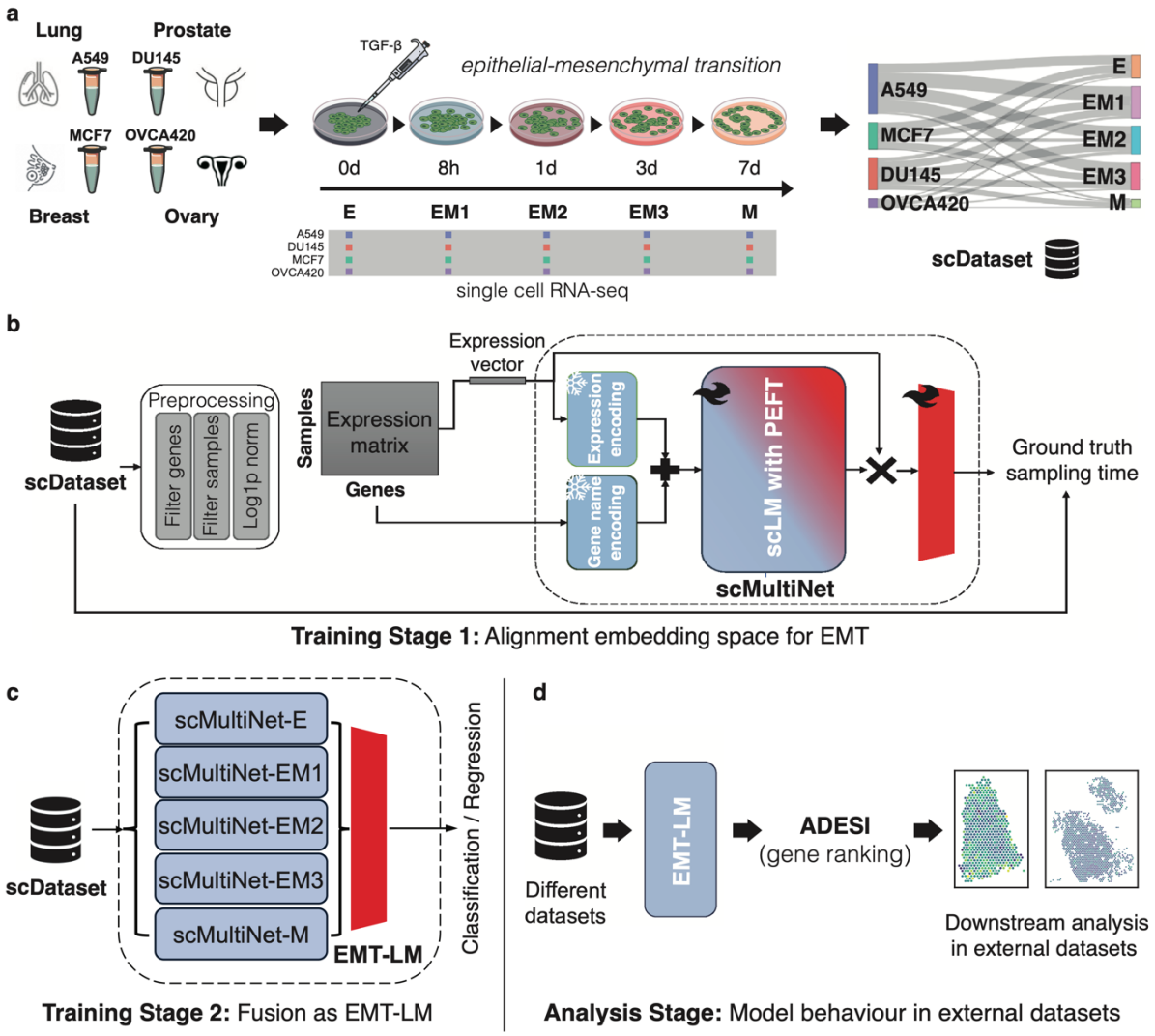


Figure 1: EMT classification pipeline in scRNA-seq data. **(a)** scRNA-seq profiles of lung (A549), prostate (DU145), breast (MCF7) and ovarian (OVCA420) cell lines before EMT induction with TGF-β (0 days - 0d) and 8 hours (8h), 1 day (1d), 3 days (3d) and 7 days (7d) after stimulus, respectively, are employed as a ground truth dataset for EMT transformation. The corresponding epithelial (E), hybrid (EM1-3) and mesenchymal (M) states of the sequenced cells at different time points are also shown. Illustrations generated with bioicons.com. **(b)** The first training phase involves binary classification training for each EMT timepoint using the Parameter-Efficient Fine-Tuning (PEFT) method. This allows the original pre-trained model's feature space to better represent EMT. **(c)** The final model is obtained by fusing the five binary classification models for the individual states (E, EM1-3, M). **(d)** For downstream analysis, we use ADESI to rank genes that contribute significantly to the model based on attention scores and gene expression. This gene list is subsequently subjected to various downstream analyses.

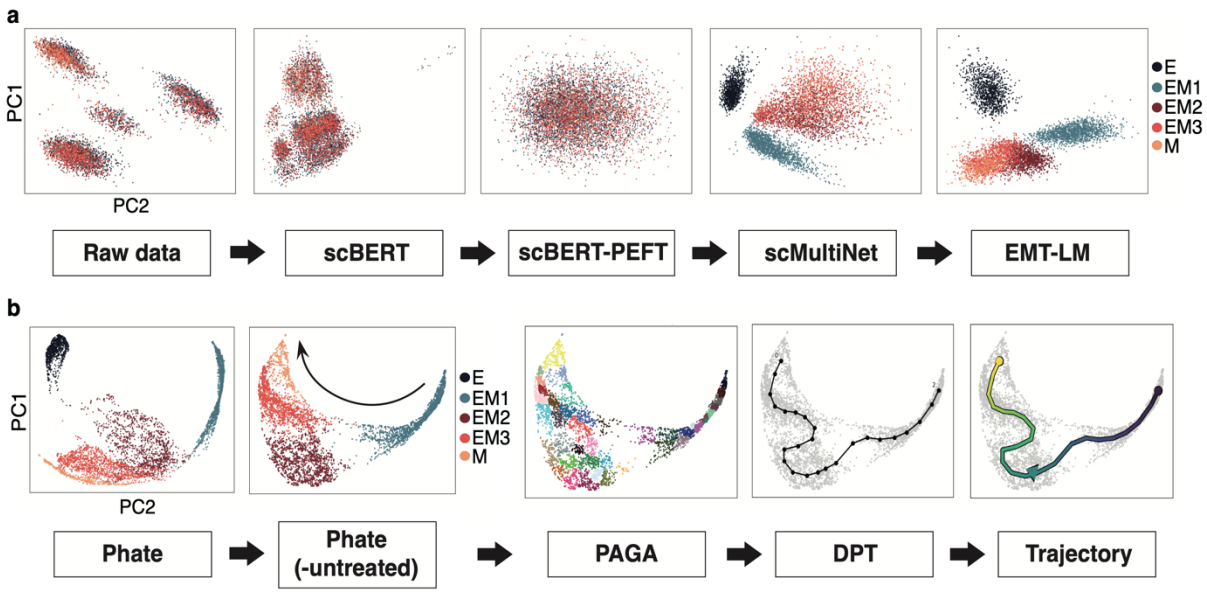


Figure 2: EMT embedding space visualisation and trajectory analysis. (a) The changes in embedding space are shown through PCA projections from various models, highlighting the transformation of raw data into refined clusters following the application of scBERT, PEFT, scMultiNet and EMT-LM. Every point corresponds to a single cell that has been profiled. The five colours indicate the ground truth EMT states. **(b)** PHATE visualisation emphasises the EMT trajectories captured by the embedding space of the final EMT-LM model. From left to right: visualisation of all sampled data points using PHATE, only samples after TGF-β stimulation (removing day 0, arrow indicates the time course progression), PAGA, DPT and trajectory reconstruction. The PAGA step delineates the data into distinct clusters, which are then sequentially ordered using DPT to infer developmental trajectories, culminating in a trajectory plot that captures the dynamic journey of cell fate decisions.

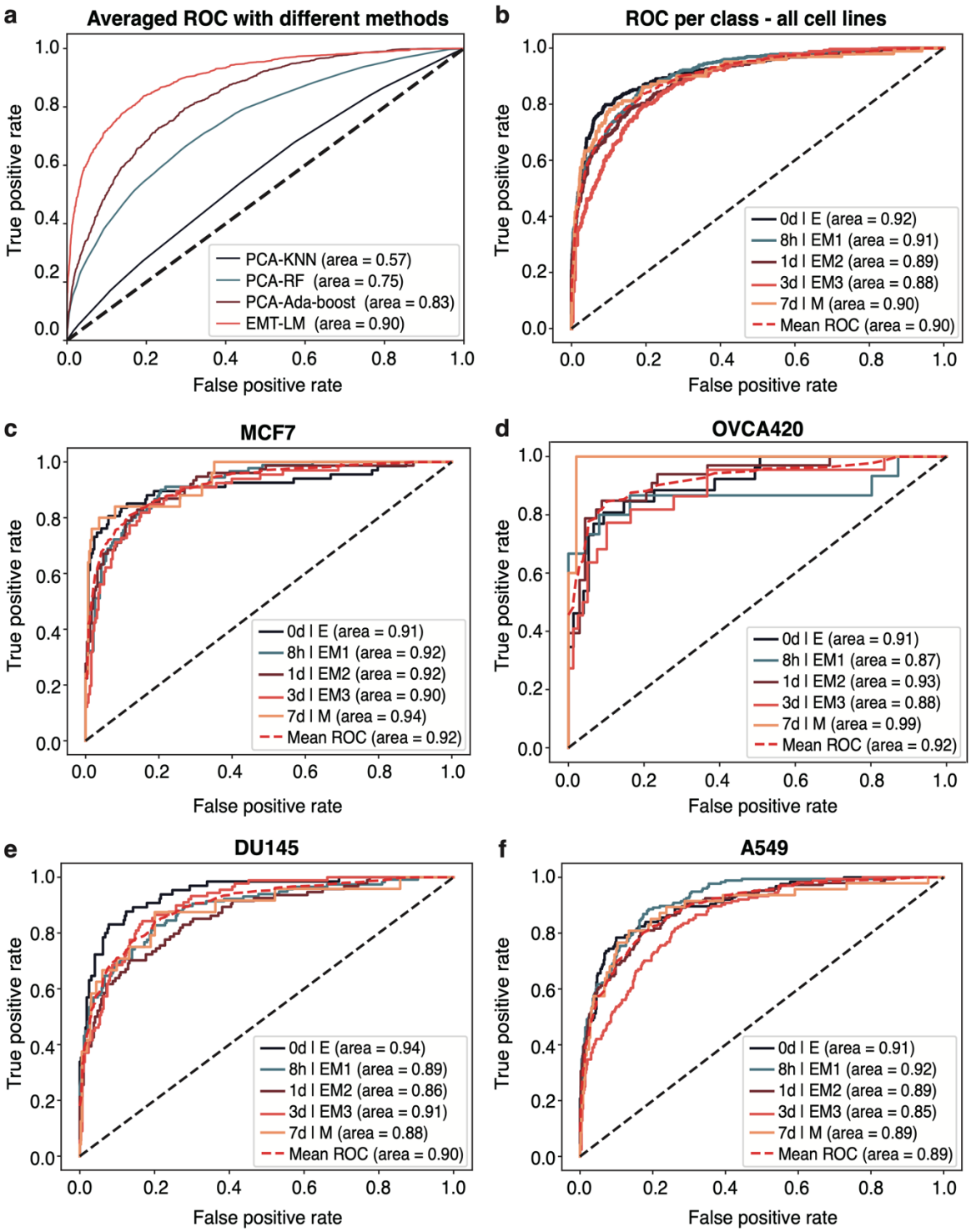


Figure 3: Model validation and comparison in different settings. (a) The averaged Receiver Operating Characteristic (ROC) curves for the proposed EMT-LM method alongside classical machine learning approaches. KNN = K-Nearest Neighbours; RF = Random Forest. **(b)** ROC curve for each class in the validation dataset with AUROC demonstrating the model's discrimination capacity for each class, and a mean AUROC indicating overall model performance. **(c-f)** ROC analysis for different cell lines, namely MCF7, OVCAR420, DU145 and A549, illustrating the model's predictive performance for each EMT state in distinct cancer types. Solid lines represent individual classes and dashed lines depict the mean ROC.

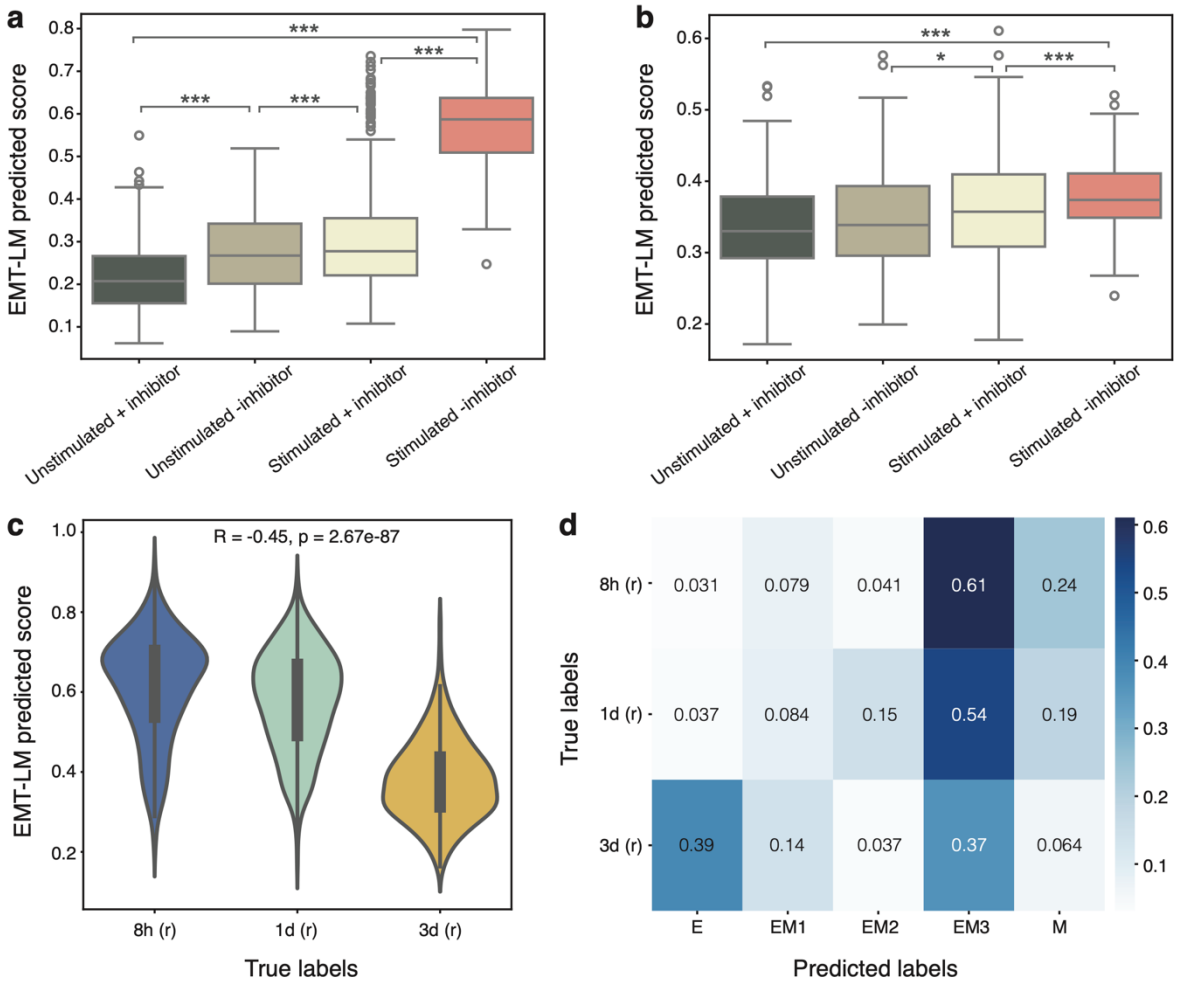


Figure 4: Model validation in EMT inhibitor and stimulus removal data in the original study. (a-b) Predicted EMT-LM scores in A548 (a) and MCF7 (b) cell lines that are either unstimulated or stimulated with TGF-β, and either treated (+inhibitor) and untreated (-inhibitor) with a TGFBR1 inhibitor. (*** p-value <0.001, ** <0.01, *<0.1). EMT scores are significantly lower after the addition of the inhibitor and in the unstimulated versus stimulated condition, as expected. **(c)** EMT-LM score distribution for the TGF-β stimulus removal data across all cancer cell lines, sequenced at 8h, 1 day and 3 days post-removal, respectively. The scores decrease with increasing time after removal, as expected. The Spearman correlation coefficient (R) and p-value are reported. **(d)** Confusion matrix of predicted cell states for the cells with TGF-β removal (averaged across all cancer cell lines).

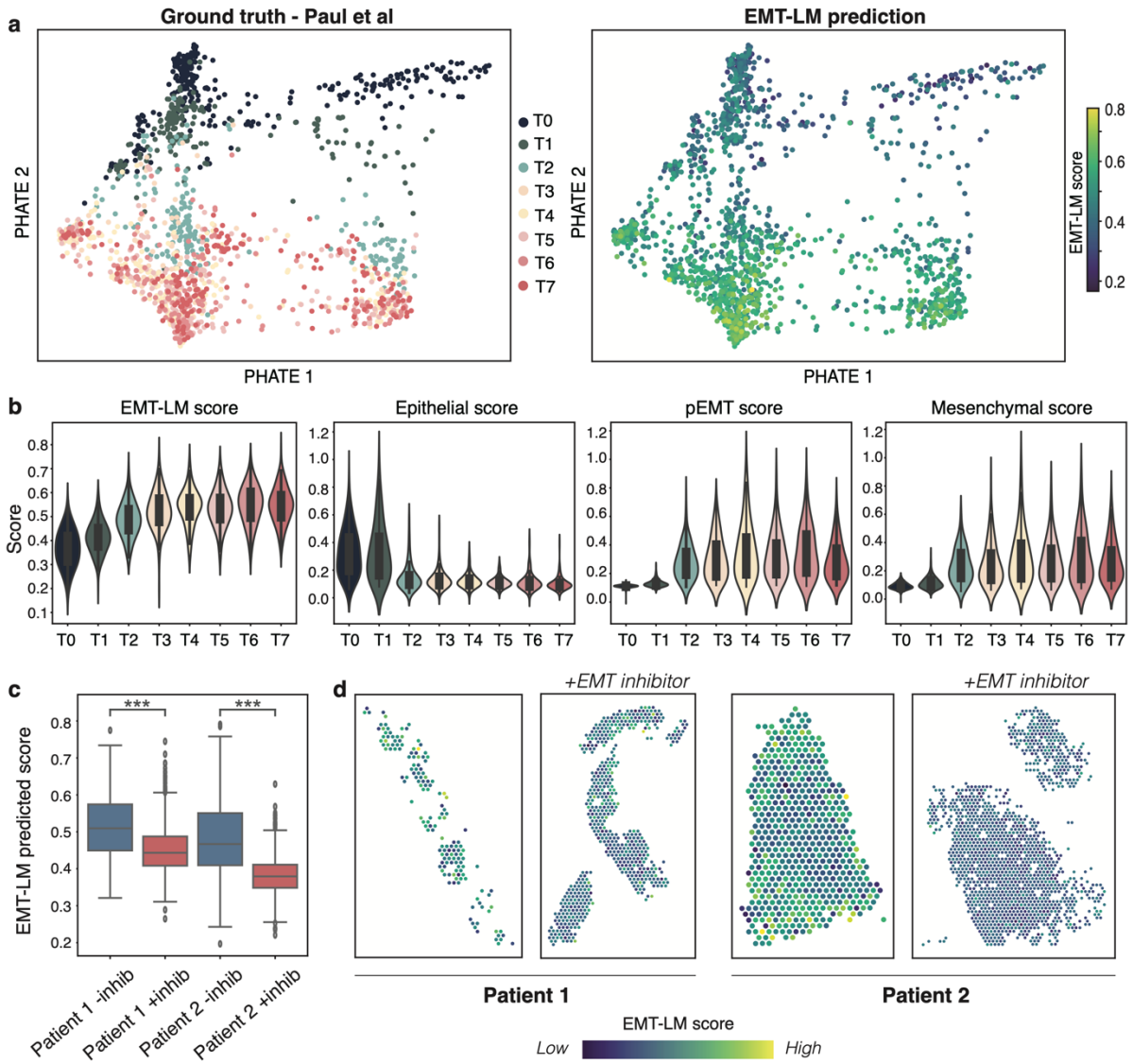


Figure 5: Validation of the EMT-LM model in external datasets. (a-b) Model validation in the Paul et al. dataset: **(a)** PHATE visualisation of individual cell projections coloured by ground truth sampling time (left panel, time points measured T1-T7) and scores predicted by EMT-LM (right panel, higher values indicate a more mesenchymal state); **(b)** Distribution of EMT-LM scores predicted for individual cells at the 7 time points alongside classic epithelial, partial EMT (pEMT) and mesenchymal scores from the literature. **(c-d)** Model validation in an external EMT inhibitor (Netrin-1) dataset from Cassier et al[23]: **(c)** Score distribution across endometrial spatial transcriptomic slides before (-inhib) and after (+inhib) treatement with the EMT inhibitor; **(d)** Visualisation of EMT-LM predicted scores acroos the spatially profiled slides before and after EMT inbition in Patient 1 (left) and Patient 2 (right).

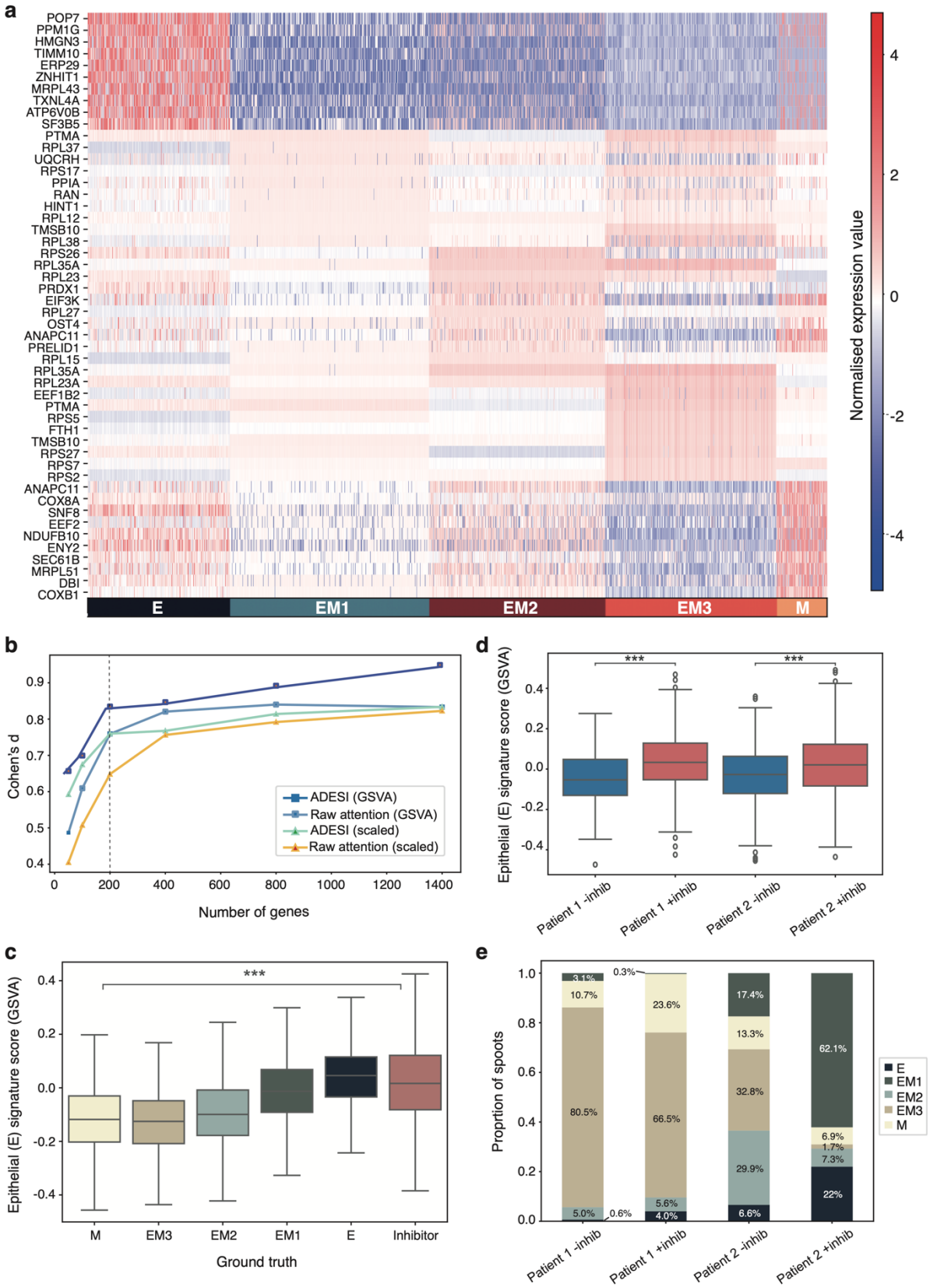


Figure 6: EMT gene signature inference and sensitivity analysis. (a) Expression heat map of key genes selected by the model using ADESI scoring across different categories, with a gradient from blue to red indicating increasing expression. The colour bar at the bottom

distinguishes the EMT state for each profiled sample, offering an intuitive representation of which genes are important in predicting various EMT states. **(b)** Sensitivity analysis plot highlighting the effect of increasing the gene list size and alternating both methods to rank genes (raw attention vs. ADESI) and methods to score the gene list (scaled versus GSVA) on distinguishing the EMT states in the Cook and Vanderhyden data. The dashed line highlights the elbow point. **(c)** Epithelial gene list validation in the Cook and Vanderhyden EMT inhibition data (*** $p<0.001$). The epithelial signature scores highest in the epithelial (E) cells at baseline and after the addition of the inhibitor, as expected. The signature decreases progressively from the EM1 to the M state. **(d)** Validation of the epithelial gene signature in the EMT inhibitor (Netrin-1) dataset from Cassier et al (*** $p<0.001$). An increase in the epithelial signature is seen in both patients after the treatment with the inhibitor (+inhib). (e) EMT states assigned based on the ADESI-derived gene signatures in the Netrin-1 EMT inhibitor dataset from Cassier et al. The fraction of 10x Visium tumour spots occupying each state in the spatial transcriptomics slides from both patients before (-inhib) and after (+inhib) the treatment with the inhibitor is shown.

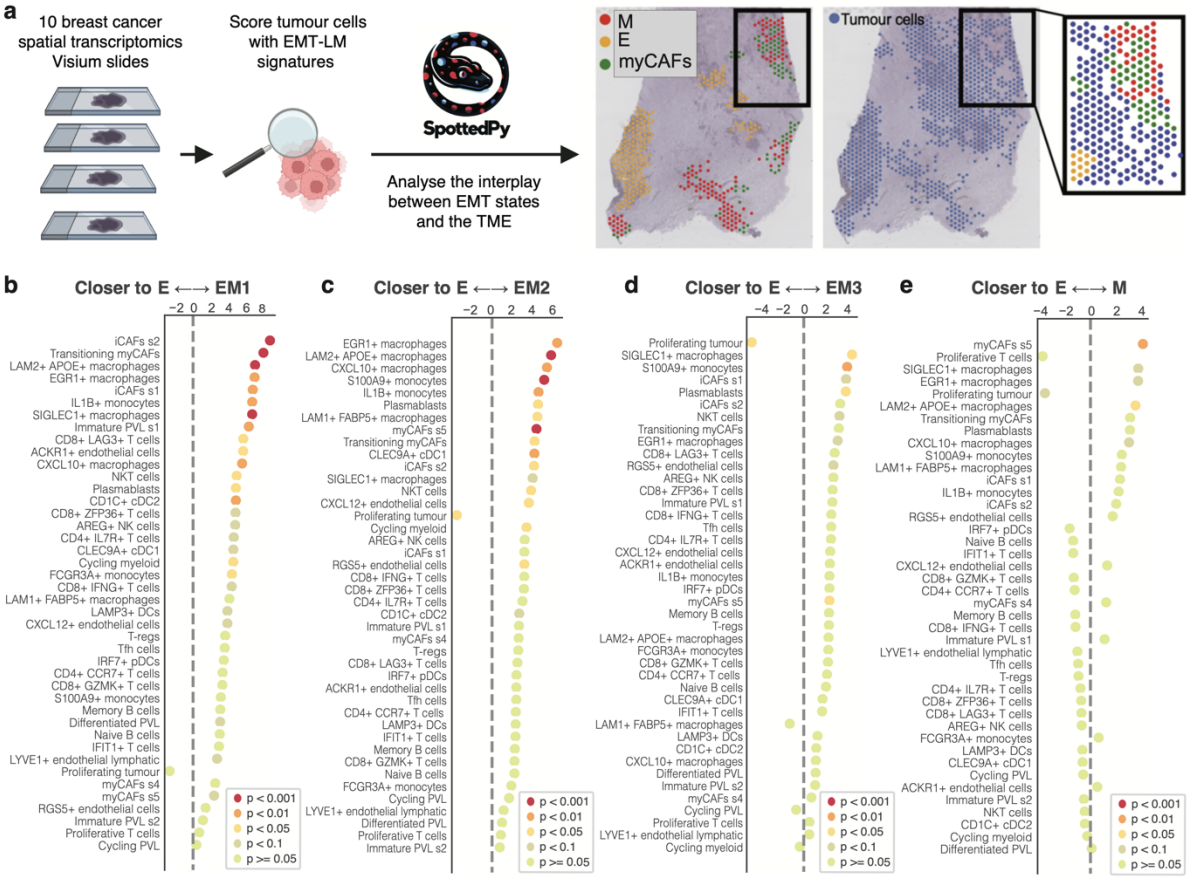
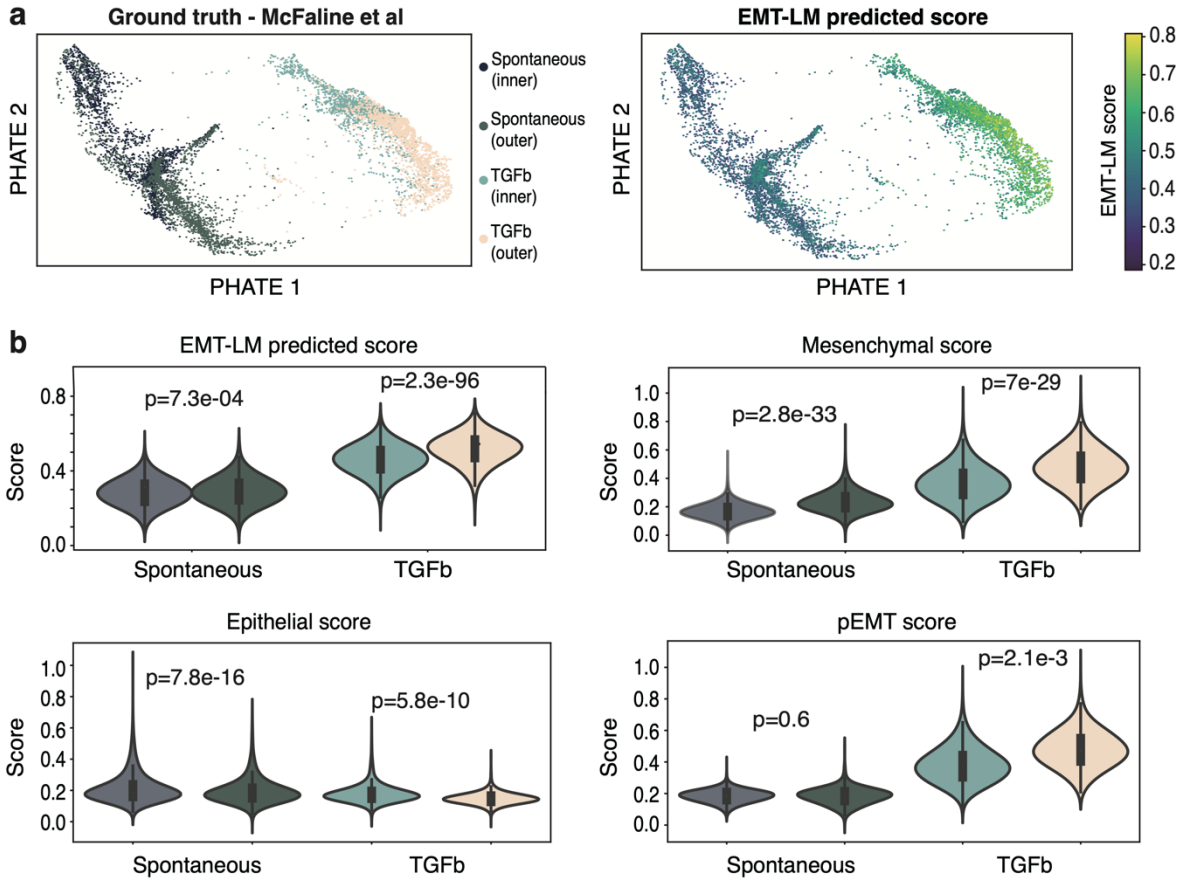
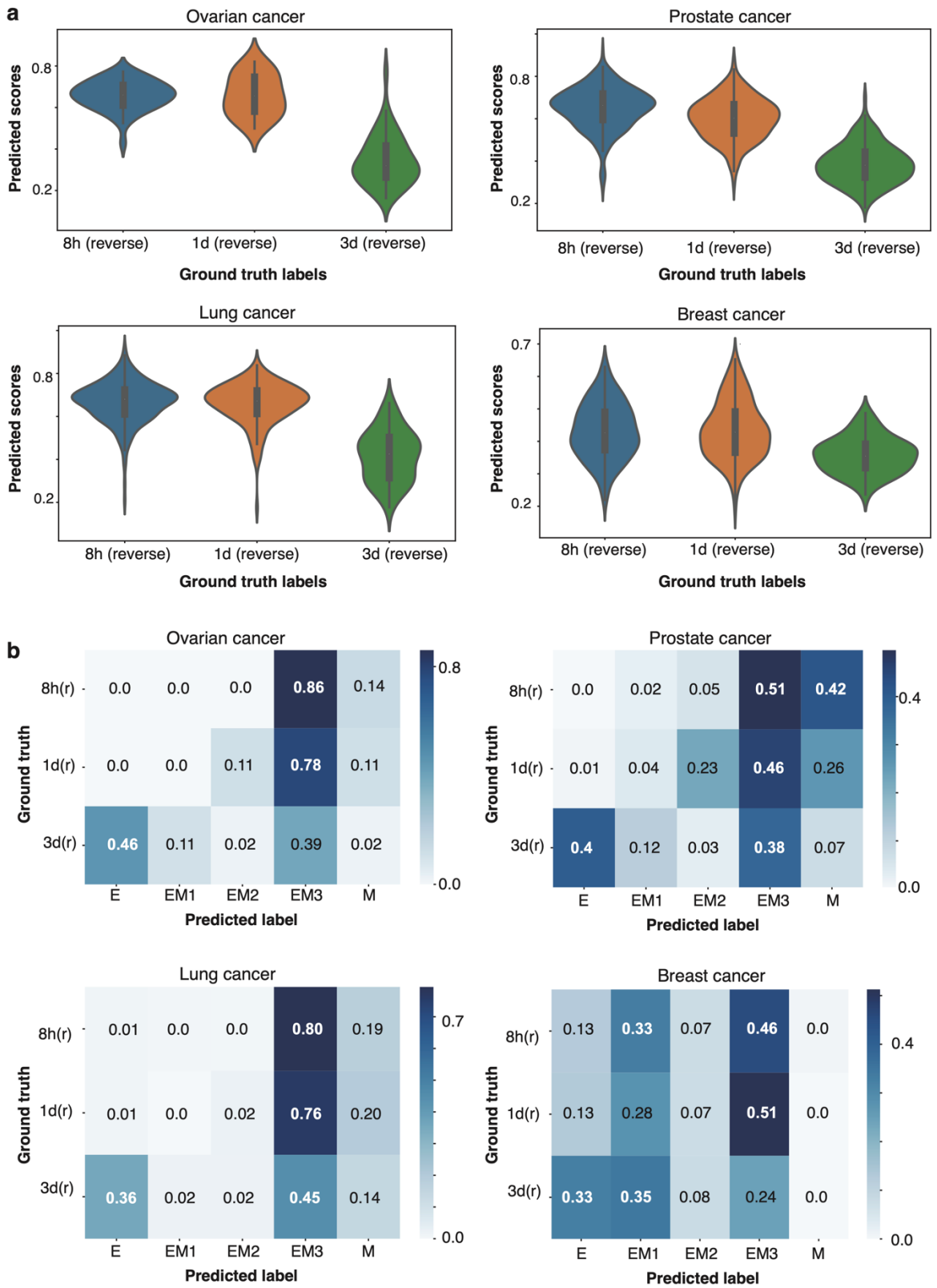


Figure 7: Interplay between EMT niches, immune and stromal cells across spatially profiled breast cancer slides. (a) Workflow describing the analysis of spatial relationships across 10 breast cancer spatial transcriptomic slides using SpottedPy, allowing the identification of EMT niches and their interplay with immune/stromal cell populations. (b-e) Distances between immune/stromal cell populations and EMT niches (EM1, EM2, EM3 and M niches are compared to E niches) in breast cancer spatial transcriptomic data. Negative values indicate cell populations closer to the E state, positive values indicate cell populations found closer to the EM1-EM3 and M states, respectively. The colours of the circles indicate the statistical significance of the association between the EMT niche and the respective cell population.

EXTENDED DATA



Extended Data Figure 1: Validation of the EMT-LM model in the McFaline-Figueroa et al dataset. (a) Left: PHATE visualisation of individual cell projections coloured by EMT stimulus (spontaneous or TGF-β-induced) and ground truth sampling time point (inner – before EMT transformation; outer – after EMT transformation). Right: Scores predicted by EMT-LM, with higher values indicating a more mesenchymal state). **(b)** Distribution of EMT-LM scores predicted for individual cells in the two stimulus conditions and before/after the EMT induction (inner-outer), alongside classic epithelial, partial EMT (pEMT) and mesenchymal scores from the literature. Violin plot colours are the same as in panel a.



Extended Data Figure 2: EMT-LM predictions in the TGF-β stimulus removal data in every cancer cell line. (a) EMT-LM predicted scores for ovarian, prostate, lung and breast cancer cell lines upon reversal of the EMT process, measured at 8 hours, 1 day and 3 days after stimulus removal. **(b)** The confusion matrices for different cancer types, contrasting ground truth categories and categories predicted by EMT-LM.

666 **SUPPLEMENTARY TABLE CAPTIONS**

667 **Supplementary Table 1.** The top 200 up-/down-regulated genes in each EMT state, as
668 uncovered by the ADESI metric and robustness analysis.

669 **Supplementary Table 2.** Pathway enrichment analysis results for the EMT gene signatures.
670 **(a)** The 0d (E) gene signature. **(b)** The 8h (EM1) gene signature. **(c)** The 1d (EM2) gene
671 signature. **(d)** The 3d (EM3) gene signature. **(e)** The 7d (M) gene signature.

672

673

674

AUTHOR CONTRIBUTIONS

MS and SP designed the experiment. MS supervised the analyses. SP developed and tested the EMT-LM model and the ADESI score. EW prepared the training and validation datasets, performed the gene sensitivity analysis, pathway enrichment and spatial transcriptomics analyses. All authors wrote and approved the manuscript.

ACKNOWLEDGEMENTS

MS and SP were supported by a UKRI Future Leaders Fellowship (MR/T042184/1). EW was supported by a studentship award from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme in Health Data Science (218529/Z/19/Z). Work in MS's lab was supported by a BBSRC equipment grant (BB/R01356X/1) and a Wellcome Institutional Strategic Support Fund (204841/Z/16/Z).

The illustrations in Figure 1 were generated using the icon libraries available at https://bioicons.com/.

ETHICS DECLARATIONS

This study employs only publicly available data. All data comply with ethical regulations, with approval and informed consent for collection and sharing already obtained by the relevant consortia.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

1 Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. *Nat Rev Mol Cell Biol* **20**, 69-84 (2019). https://doi.org/10.1038/s41580-018-0080-4

2 Lu, W. & Kang, Y. Epithelial-Mesenchymal Plasticity in Cancer Progression and Metastasis. *Dev Cell* **49**, 361-374 (2019). https://doi.org/10.1016/j.devcel.2019.04.010

3 Sinha, D., Saha, P., Samanta, A. & Bishayee, A. Emerging Concepts of Hybrid Epithelial-to-Mesenchymal Transition in Cancer Progression. *Biomolecules* **10** (2020). https://doi.org/10.3390/biom10111561

4 Haerinck, J., Goossens, S. & Berx, G. The epithelial-mesenchymal plasticity landscape: principles of design and mechanisms of regulation. *Nat Rev Genet* **24**, 590-609 (2023). https://doi.org/10.1038/s41576-023-00601-0

5 Malagoli Tagliazucchi, G., Wiecek, A. J., Withnell, E. & Secrier, M. Genomic and microenvironmental heterogeneity shaping epithelial-to-mesenchymal trajectories in cancer. *Nat Commun* **14**, 789 (2023). https://doi.org/10.1038/s41467-023-36439-7

6 Brown, M. S. *et al.* Phenotypic heterogeneity driven by plasticity of the intermediate EMT state governs disease progression and metastasis in breast cancer. *Sci Adv* **8**, eabj8002 (2022). https://doi.org/10.1126/sciadv.abj8002

7 Kroger, C. *et al.* Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc Natl Acad Sci U S A* **116**, 7353-7362 (2019). https://doi.org/10.1073/pnas.1812876116

8 Cook, D. P. & Vanderhyden, B. C. Context specificity of the EMT transcriptional response. *Nat Commun* **11**, 2142 (2020). https://doi.org/10.1038/s41467-020-16066-2

9 McFaline-Figueroa, J. L. *et al.* A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat Genet* **51**, 1389-1398 (2019). https://doi.org/10.1038/s41588-019-0489-5

10 Celia-Terrassa, T. *et al.* Hysteresis control of epithelial-mesenchymal transition dynamics conveys a distinct program with enhanced metastatic ability. *Nat Commun* **9**, 5005 (2018). https://doi.org/10.1038/s41467-018-07538-7

11 Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* **4**, 852-+ (2022). https://doi.org/10.1038/s42256-022-00534-z

12 Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* (2024). https://doi.org/10.1038/s41592-024-02201-0

13 Levine, D. *et al.* Cell2Sentence: Teaching Large Language Models the Language of Biology. *bioRxiv*, 2023.2009.2011.557287 (2024). https://doi.org/10.1101/2023.09.11.557287

14 Chen, Y. & Zou, J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. *bioRxiv* (2024). https://doi.org/10.1101/2023.10.16.562533

15 Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481-1491 (2024). https://doi.org/10.1038/s41592-024-02305-7

16 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 (2018). <https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D>.

17 Liu, X. *et al.* GPT understands, too. *AI Open* (2023). https://doi.org/https://doi.org/10.1016/j.aiopen.2023.08.012

18 Gu, A. & Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752 (2023). <https://ui.adsabs.harvard.edu/abs/2023arXiv231200752G>.

19 Ding, N. *et al.* Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat Mach Intell* **5**, 220-235 (2023). https://doi.org/10.1038/s42256-023-00626-4

20 Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* **37**, 1482-1492 (2019). https://doi.org/10.1038/s41587-019-0336-3

21 Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845-848 (2016). https://doi.org/10.1038/nmeth.3971

22 Paul, I. *et al.* Parallelized multidimensional analytic framework applied to mammary epithelial cells uncovers regulatory principles in EMT. *Nat Commun* **14**, 688 (2023). https://doi.org/10.1038/s41467-023-36122-x

23 Cassier, P. A. *et al.* Netrin-1 blockade inhibits tumour growth and EMT features in endometrial cancer. *Nature* **620**, 409-416 (2023). https://doi.org/10.1038/s41586-023-06367-z

24 Abdulhag, U. N. *et al.* Mitochondrial complex IV deficiency, caused by mutated COX6B1, is associated with encephalomyopathy, hydrocephalus and cardiomyopathy. *Eur J Hum Genet* **23**, 159-164 (2015). https://doi.org/10.1038/ejhg.2014.85

25 Guerra, F. *et al.* Mitochondrial Dysfunction: A Novel Potential Driver of Epithelial-to-Mesenchymal Transition in Cancer. *Front Oncol* **7**, 295 (2017). https://doi.org/10.3389/fonc.2017.00295

26 Cao, Y. *et al.* An emerging master inducer and regulator for epithelial-mesenchymal transition and tumor metastasis: extracellular and intracellular ATP and its molecular functions and therapeutic potential. *Cancer Cell Int* **23**, 20 (2023). https://doi.org/10.1186/s12935-023-02859-0

27 Debnath, P., Huirem, R. S., Dutta, P. & Palchaudhuri, S. Epithelial-mesenchymal transition and its transcription factors. *Biosci Rep* **42** (2022). https://doi.org/10.1042/BSR20211754

28 Liu, N. *et al.* SNRPB promotes the tumorigenic potential of NSCLC in part by regulating RAB26. *Cell Death Dis* **10**, 667 (2019). https://doi.org/10.1038/s41419-019-1929-y

29 Chi, L. H. *et al.* Global Proteomics-based Identification and Validation of Thymosin Beta-4 X-Linked as a Prognostic Marker for Head and Neck Squamous Cell Carcinoma. *Sci Rep* **7**, 9031 (2017). https://doi.org/10.1038/s41598-017-09539-w

30 D'Angelo, E. *et al.* Intrinsic and Extrinsic Modulators of the Epithelial to Mesenchymal Transition: Driving the Fate of Tumor Microenvironment. *Front Oncol* **10**, 1122 (2020). https://doi.org/10.3389/fonc.2020.01122

31 Zheng, S. *et al.* Role and mechanism of actin-related protein 2/3 complex signaling in cancer invasion and metastasis: A review. *Medicine (Baltimore)* **102**, e33158 (2023). https://doi.org/10.1097/MD.0000000000033158

32 Xu, Z. *et al.* Integrin beta1 is a critical effector in promoting metastasis and chemo-resistance of esophageal squamous cell carcinoma. *Am J Cancer Res* **7**, 531-542 (2017).

33 Prakash, V. *et al.* Ribosome biogenesis during cell cycle arrest fuels EMT in development and disease. *Nat Commun* **10**, 2110 (2019). https://doi.org/10.1038/s41467-019-10100-8

34 Withnell, E. & Secrier, M. SpottedPy quantifies relationships between spatial transcriptomic hotspots and uncovers new environmental cues of epithelial-mesenchymal plasticity in cancer. *bioRxiv*, 2023.2012.2020.572627 (2023). https://doi.org/10.1101/2023.12.20.572627

35 Zhang, C. *et al.* CAFs orchestrates tumor immune microenvironment-A new target in cancer therapy? *Front Pharmacol* **14**, 1113378 (2023). https://doi.org/10.3389/fphar.2023.1113378

36 Ardighieri, L. *et al.* Infiltration by CXCL10 Secreting Macrophages Is Associated With Antitumor Immunity and Response to Therapy in Ovarian Cancer Subtypes. *Front Immunol* **12**, 690201 (2021). https://doi.org/10.3389/fimmu.2021.690201

37 Trizzino, M. *et al.* EGR1 is a gatekeeper of inflammatory enhancers in human macrophages. *Sci Adv* **7** (2021). https://doi.org/10.1126/sciadv.aaz8836

38 Baitsch, D. *et al.* Apolipoprotein E induces antiinflammatory phenotype in macrophages. *Arterioscler Thromb Vasc Biol* **31**, 1160-1168 (2011). https://doi.org/10.1161/ATVBAHA.111.222745

39 Yang, J. *et al.* Guidelines and definitions for research on epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol* **21**, 341-352 (2020). https://doi.org/10.1038/s41580-020-0237-9

40 Li, D. *et al.* Heterogeneity and plasticity of epithelial-mesenchymal transition (EMT) in cancer metastasis: Focusing on partial EMT and regulatory mechanisms. *Cell Prolif* **56**, e13423 (2023). https://doi.org/10.1111/cpr.13423

41 Jolly, M. K. *et al.* Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget* **7**, 27067-27084 (2016). https://doi.org/10.18632/oncotarget.8166

42 Bakir, B., Chiarella, A. M., Pitarresi, J. R. & Rustgi, A. K. EMT, MET, Plasticity, and Tumor Metastasis. *Trends Cell Biol* **30**, 764-776 (2020). https://doi.org/10.1016/j.tcb.2020.07.003

43 Aiello, N. M. *et al.* EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. *Dev Cell* **45**, 681-695 e684 (2018). https://doi.org/10.1016/j.devcel.2018.05.027

44 Sha, Y. *et al.* Intermediate cell states in epithelial-to-mesenchymal transition. *Phys Biol* **16**, 021001 (2019). https://doi.org/10.1088/1478-3975/aaf928

45 Barkley, D. *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat Genet* **54**, 1192-1201 (2022). https://doi.org/10.1038/s41588-022-01141-9

46 Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**, 1334-1347 (2021). https://doi.org/10.1038/s41588-021-00911-1

47 Elyanow, R., Zeira, R., Land, M. & Raphael, B. J. STARCH: copy number and clone inference from spatial transcriptomics data. *Phys Biol* **18**, 035001 (2021). https://doi.org/10.1088/1478-3975/abbe99

48 Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39** (2023). https://doi.org/10.1093/bioinformatics/btac757

49 Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**, W193-200 (2007). https://doi.org/10.1093/nar/gkm226