Analyzing Collective Migration and Epithelial-Mesenchymal Transitions using Graph Neural Networks

Zhuoyang Lyu

Mentor: Dr. Hyeontae Jeong
Advisor: Prof. Ian Y. Wong

April 30, 2025

Abstract

Collective migration of epithelial cells during epithelial-mesenchymal transition (EMT) underlies key physiological and pathological processes, including tissue development and cancer metastasis. Here, we present a graph-based deep learning framework leveraging Graph Attention Networks (GATs) to predict single-cell migration behaviors from static morphological snapshots and neighborhood interactions. We first validate our approach on self-propelled particle simulations adapted from the Vicsek model, demonstrating near-perfect recovery of individual displacement magnitudes. We then apply the model to time-lapse microscopy of TGF-β1–treated MCF10A monolayers, encoding each cell as a node with handcrafted morphological features and edges defined by spatial proximity. On this experimental dataset, our GAT accurately predicts global average migration speed but exhibits overfitting when tasked with single-cell displacement, two-dimensional movement vectors, and UMAP-embedded trajectory features, highlighting the challenges of single-cell level migration prediction. These findings underscore the promise of attention-based GNNs for capturing collective behaviors while revealing the need for richer temporal contexts and high-dimensional feature embeddings. We discuss avenues for future work, including dynamic graph architectures and deep image-based node representations, toward enabling robust, interpretable models of single-cell migration in complex tissues.

Contents

1 Acknowledgment

I would like to express my deepest gratitude to Prof. Ian Y. Wong for his invaluable guidance, thoughtful supervision, and unwavering support throughout this project and my broader academic journey. I would also like to express my sincere thanks to Dr. Hyeontae Jeong, whose mentorship, and collaboration were instrumental in completing this project. Without his generous help and encouragement, this work would not have been possible. I am also thankful to Professor Stephen Bach for serving as my second reader and for his thoughtful feedback and encouragement. Finally, I would like to acknowledge the funding support from the Brown Undergraduate Teaching and Research Awards (Summer UTRA), which supported the early stage of this research.

2 Introduction

Epithelium, or epithelial tissue, is one of the four fundamental types of animal tissue, forming thin, continuous layers that cover external surfaces and line internal cavities and organs. Epithelial tissues perform a wide array of functions—including protection, secretion, absorption, filtration, diffusion, excretion, and sensory reception—critical for maintaining organ homeostasis and barrier integrity. Structurally, epithelial cells are closely packed and connected by specialized junctions (e.g. tight junctions, adherens junctions, and desmosomes) and rest upon a basement membrane; they exhibit diverse shapes (e.g. squamous, cuboidal, columnar) and arrangements (e.g. simple, stratified, or pseudostratified) that reflect their specific roles. By maintaining selective barriers and regulated transport pathways, epithelia contribute to fluid balance, nutrient exchange, and defense against pathogens.

Epithelial cells dynamically reprogram their adhesive phenotypes to facilitate collective migration during tissue development and wound healing[1, 2, 3, 4]. A well-known example of this phenomenon is the epithelial-mesenchymal transition (EMT) (Fig.1). During EMT, epithelial cells lose cell-cell adhesions and apicobasal polarity, transforming into a mesenchymal phenotype that is more migratory.[1, 3] EMT can be classified into three contextual subtypes: Type 1 EMT underpins embryogenesis and organ development, Type 2 EMT is reactivated in adults for wound healing and tissue regeneration (and, if dysregulated, contributes to fibrosis), and Type 3 EMT facilitates cancer progression and metastasis[1, 5, 6]. During embryonic development, EMT drives fundamental morphogenetic events such as gastrulation and neural crest cell migration, whereas in adult tissues it promotes re-epithelialization and repair following injury[7]. Conversely, chronic or aberrant EMT activation is a hallmark of pathological states, including organ fibrosis and carcinoma dissemination, where transformed cells invade and colonize distant sites. Therefore understanding EMT is essential for advancing regenerative medicine, developing anti-fibrotic strategies, and identifying targeted therapies to inhibit cancer metastasis[1, 4, 8, 9].

Recent research has identified partial or hybrid EMT states, where cells retain some epithelial characteristics while acquiring migratory behaviors.[1, 10, 11] These hybrid states are implicated in collective cell migration, often acting as "leader cells." However, profiling the functional phenotypes of partial EMT states remains challenging since they are transient and likely depend on local cell-cell interactions. Elucidating the dynamics of EMT and accurately predicting cell migration behavior are therefore critical to uncovering the mechanisms underlying cancer invasion and tissue remodeling.

To address these complexities, emerging computational approaches are being employed to model the intricate interactions within cellular environments.[12] Graph Neural Networks (GNNs) [13, 14, 15, 16] represent a powerful class of neural networks specifically designed to process data structured as graphs, a set of vertices or nodes representing entities connected by edges encoding the relationships between entities. By iteratively aggregating information from neighboring

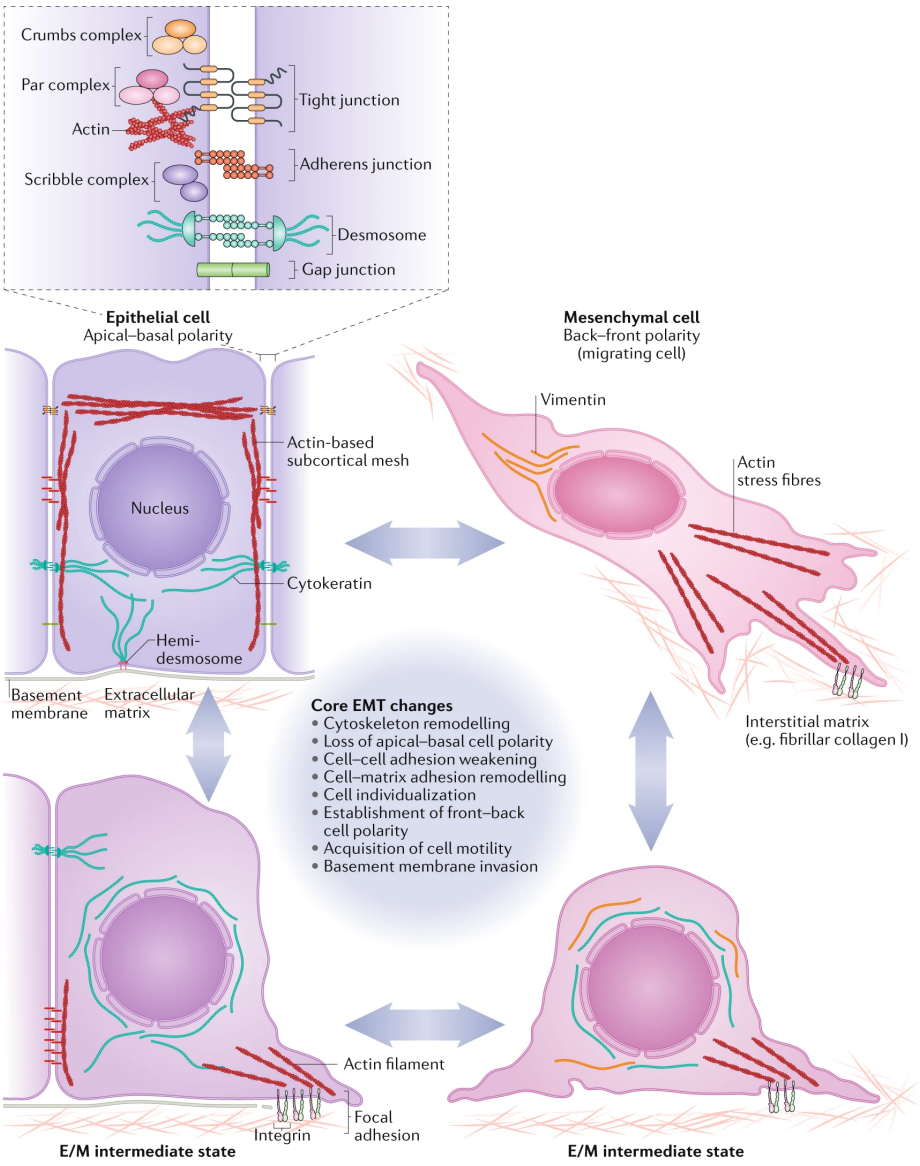

Figure 1: Figure by Yang et al.[1]. The accumulated loss or gain of epithelial/mesenchymal (E/M) characteristics pushes a cell towards various intermediate states (bottom left and right) in a fluid and reversible manner, between a complete epithelial (middle left) and a complete mesenchymal (middle right) state. EMT, epithelial-mesenchymal transition.

nodes through a message-passing process, GNNs capture both local and global structural relationships within the data. This capability makes GNNs particularly suited for studying systems where the topology of interactions plays a pivotal role. We draw an analogy between the unique structure of graph and the specific nature of cell-cell interactions: both the intrinsic properties of individual cells and the relationships between them are indispensable for accurately predicting migration behavior.

In this paper, we explore leveraging graph representation and GNN models to provide new insights into the dynamics of EMT. Our work offers an approach to predicting cell migration behaviors and understanding the underlying processes of tissue remodeling and cancer metastasis.

3 Previous Works

In recent years, the integration of machine learning techniques, particularly Graph Neural Networks (GNNs), into the study of collective cell migration has provided novel insights into the complex dynamics of cellular behavior during processes such as the epithelial-mesenchymal transition (EMT). This section reviews relevant literature that informs and contextualizes our approach to modeling collective cell migration using GNNs, with a focus on key studies that have significantly contributed to this field.

Deep Attention Networks in Collective Cell Migration

The application of deep learning models, especially attention-based networks, has been instrumental in elucidating the rules governing collective motion. An earlier study by Heras et al. [17] demonstrated the efficacy of deep attention networks in modeling collective behavior in zebrafish. Their approach not only achieved predictive accuracy but also provided interpretable models that identified interaction rules among individuals in a collective. This methodological framework has set a precedent for applying similar techniques to understand cellular collectives. A study by LaChance et al.[12] employed deep attention networks to analyze collective cell migration in various tissue types, including epithelial and metastatic breast cancer cells. Their model learned the influence patterns of neighboring cells on a focal cell's movement decisions, revealing distinct attention patterns unique to each tissue type. For instance, endothelial cells exhibited focused attention on immediate forward neighbors, while epithelial cells displayed broader influence patterns. This work underscored the potential of attention mechanisms to uncover interpretable insights into cell-cell interactions during migration.

It is important to note that the modeling in these studies focuses solely on the prediction of turning directions—classifying movements as either left or right. This binary formulation is a simplification that abstracts the continuous nature of directional changes and, while it captures essential aspects of decision-making, it potentially overlooks finer gradations in migratory behavior.

Graph Neural Networks for Inferring Cell Dynamics

GNN has been widely used in biomedical research. Recent research has demonstrated that Multi-Layer Graph Attention Neural Networks (MLGANN) effectively predict drug-target interactions by integrating multi-source chemical and protein data within a graph attention framework[18]. The LINGER framework applies lifelong graph neural network architectures to infer gene regulatory networks from paired single-cell expression and chromatin accessibility data[19]. In single-cell omics, scGNN 2.0 employs graph-based imputation and clustering to recover dropout-prone gene expression profiles and uncover cell-cell communication patterns[20]. In microbial ecology, GNN approaches model community dynamics directly from genomic features to predict steady-state abundances without explicit growth curves[21].

The use of GNNs to predict cell movement from static configurations represents a significant advancement in modeling collective cell behavior. Yang et al.[22] demonstrated that GNNs could predict the ensemble averaged mobility of a cell monolayer from the cell positions. By representing cells as nodes and their interactions as edges, the model effectively captured the structural features indicative of collective movement. This approach highlighted the capability of GNNs to learn complex interaction patterns and predict dynamic behaviors in biological systems. However, the model does not provide detailed single-cell trajectories, which only offer insights into global migration trends and does not address the heterogeneity inherent in individual cell dynamics.

4 Data Collection and Problem Setup

In this study, we develop a computational framework to predict cell migration behavior by integrating both intrinsic cellular features and the influence of the local microenvironment.

Simulation Data Generation

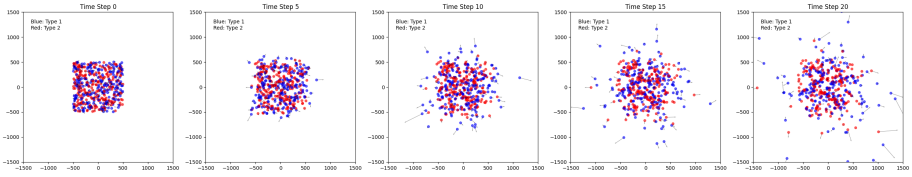

Figure 2: Snapshots of the simulated cell migration. Cells are uniformly initialized in space ($[-500, 500], [-500, 500]$). Each cell is assigned one of two phenotypes to emulate heterogeneous EMT stages. The color of the cell represents its phenotype, blue indicating type 1 with inial speed 25 and red indicating type 2 with inial speed 10. The arrows indicate the velocity vectors of the cell at the time point.

We generated a synthetic dataset by simulating cell migration with a self-propelled particle model adapted from the Vicsek framework[23] (see Methods). Specifically, we uniformly initialized n cells in a square space of side length L, assigning each cell one of two phenotypes to emulate heterogeneous EMT stages. Morphological features were encoded as one-hot phenotype vectors perturbed with Gaussian noise to reflect biological variability. Initial velocities were sampled from normal distributions centered on each phenotype's characteristic speed and oriented in random directions. At each time step, we defined a cell's neighborhood as all cells within a fixed interaction radius r (see graph construction below) and updated its velocity to be a weighted average of its previous velocity and the mean velocity of its neighbors, treating this as the movement over a unit time interval:

$$\mathbf{v}_i(t+1) = (1-\lambda)\,\mathbf{v}_i(t) + \lambda\,\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\mathbf{v}_j(t) + \boldsymbol{\eta}_i(t), \quad \boldsymbol{\eta}_i(t) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_2).$$

Although this averaging interaction is a simplification of true cell-cell forces, it provides a challenging proxy: the model must infer both individual phenotype and neighborhood context to predict movement magnitude accurately. We used this synthetic data to evaluate our model's ability to learn collective behaviors under controlled settings with ground truth available. Figure2 shows the simulated cells and their velocity vectors at different time points.

Experimental Data Collection

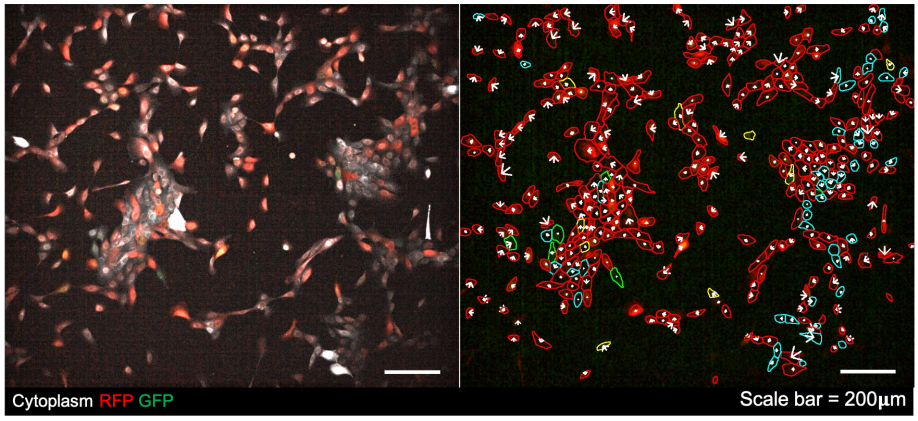

Figure 3: An example snapshot of the cell culture expressing fluorescent reporters color reporters (left) and corresponding segmentation mask (right). In the left panel, white indicates the cytoplasm, while red and green represent the expression of EMT reporters. In the right panel, arrows indicates the direction of the cell movement with size proportional to the cell speed. Scale bar is 200 μm.

Our dataset comprises an collection of morphological and trajectory data obtained from mammary epithelial cells (MCF10A) treated with TGF-β1 cultured in 2D substrate. The treatment, at varying concentrations (0, 1ng/ml, and 5ng/ml) for 1 week before the experiment, induces either partial or complete epithelial-to-mesenchymal transition (EMT). Following treatment, epithelial cells were cultured on the fibronectin coated surface with normal control media to observe their characteristic features of migration and morphologies with the heterogeneous EMT states.

Time-lapse imaging was conducted using live-cell microscopy, capturing images every 15 minutes across a total of 60 frames. From these sequential images, we extracted a comprehensive set of quantitative morphological features by segmenting the cell boundary, assigned color-coded labels to denote the EMT stage (epithelial, mesenchymal, partial EMT, or null), and analyzed migration patterns derived from the tracked cell movements (Fig.3). This detailed dataset forms the basis for our investigation into the dynamics of cell migration and the underlying mechanisms of EMT. (Detail in Methods)

Graph Representation

Leveraging the two-dimensional organization of our cultured cells, we represent each static snapshot of the cell culture as a graph. We encode each cell as a node within a graph, where node attributes encompass quantitative morphological measurements and the cell's current color-coded stage in EMT, as determined via segmentation mask analysis. Edges are constructed between pairs of cells that lie within a predefined radius, effectively capturing the spatial relationships and local interactions among neighboring cells. This graph-based representation allows us to account for both the individual states of cells and the contextual information provided by their immediate surroundings. An example graph representation is shown in Fig.4.

GNN Model

Graph Neural Networks (GNNs) are particularly well-suited for this task, as they can simultaneously learn from the intrinsic features of each cell and the structural dependencies present within the cellular network. By employing this approach, we build models to dynamically predict various movement factors for each cell in the future time frame-such as displacement magnitude, 2D movement vectors, and relative changes in mean distance to neighboring cells—providing a comprehensive, real-time view of the mechanisms driving collective cell migration.

Graph Attention Networks (GAT)[16] were selected as the model architecture for our study due to their inherent ability to capture the complex, non-uniform influence of neighboring cells on the migratory behavior of a focal cell. The core advantage of GAT lies in its attention mechanism, which dynamically

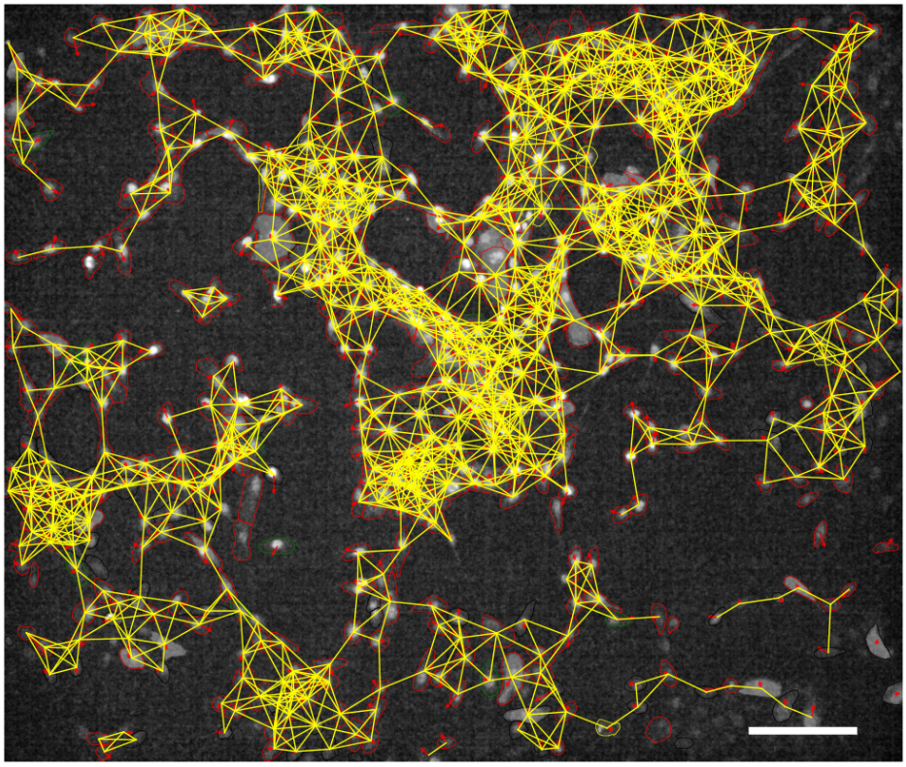

Figure 4: A representative graph representation of the cell layer under control conditions. Yellow lines indicate neighboring connections between cells, defined by a threshold radius. Cell boundaries are visualized with colored outlines corresponding to the reporter expression, and cell velocities are shown as red arrows. Scale bar is 200 μm.

assigns weights to neighboring nodes based on their relative contributions:

$$\mathbf{h}_i^{(l+1)} = \sigma\Big(\sum_{j\in\mathcal{N}_i} \alpha_{ij}\, W\, \mathbf{h}_j^{(l)}\Big), \quad \alpha_{ij} = \frac{\exp\big(\mathrm{LeakyReLU}\big(a^\top [W\mathbf{h}_i^{(l)} \,\|\, W\mathbf{h}_j^{(l)}]\big)\big)}{\sum_{k\in\mathcal{N}_i} \exp\big(\mathrm{LeakyReLU}\big(a^\top [W\mathbf{h}_i^{(l)} \,\|\, W\mathbf{h}_k^{(l)}]\big)\big)}.$$

This is particularly beneficial in our context, where the influence of neighboring cells is expected to vary due to factors such as cell state heterogeneity, varying intercellular distances, and the potential presence of leader cells that can drive collective motion. The attention mechanism in GAT allows the model to learn which neighbors are most informative for predicting a cell's migration. This aligns with the biological intuition that not all neighboring cells exert equal influence. Furthermore, the potential existence of leader cells[1, 11], is naturally accommodated by the GAT framework. The model's ability to highlight influential nodes provides both improved predictive performance and enhanced interpretability, offering deeper insights into the mechanisms driving cell migration.

We designed our integrated framework not only to enhance our understanding of cell migration dynamics but also to provide a robust tool for investigating the interplay between cellular states and their local interactions during EMT.

Prediction of Cell motility in the UMAP space

The prediction of total migration distance can be meaningful in relatively simple systems, such as highly confluent cell layers. However, in our experimental setup with intermediate cell densities, we observed a wide range of migration behaviors. Some cells exhibited directed migration with high persistence, while others displayed random-walk-like motion. These patterns are clearly distinct but cannot be easily characterized using a single metric such as distance or speed.

Thus, here we also targeted to capture diverse migration features, including total distance, mean squared displacement (MSD), neighboring similarity, meandering index, and other features reflecting directional persistence and collective behavior. To better represent this multidimensional behavior, we reduced the migration feature space to a two-dimensional representation using Uniform Manifold Approximation and Projection (UMAP). The UMAP reduction was performed with 10 neighbors, a minimum distance of 0.1, and the Euclidean metric. The resulting 2D UMAP embedding of the migration features did not show clearly separated structures across experimental conditions, highlighting the complexity of our system and the limitations of using simple descriptors (Fig.5A). Nonetheless, we observed a subtle divergence in cell distributions based on TGF-β1 treatment. Cells pre-treated with a high concentration of TGF-β1 (5 ng/ml) tended to cluster at lower values of both UMAP components, whereas control cells were clustered in regions with higher UMAP2 values. The correlation map between UMAP components and migration properties showed a brief description for the data position (Fig.5B). A high UMAP1 value was primarily associated with greater cumulative migration distance, while a high

UMAP2 value was correlated with increased meandering index, reflecting more persistent migration patterns.

To assess whether UMAP space also reflects morphological differences, we overlaid morphological clusters onto the motility UMAP. Morphological features were clustered into four groups using principal component analysis (PCA) (Fig.5C). As shown in Fig. 5D, these clusters exhibited distinct distributions in the UMAP domain. Additionally, kernel density estimation (KDE) for each morphological cluster, even within the control condition alone, revealed localized hotspots, further supporting the relevance of the UMAP space (Fig.5E). Based on these findings, we hypothesize that targeting the 2D UMAP coordinates as prediction outputs for our GAT model could enhance its predictive power by incorporating more nuanced representations of cell motility.

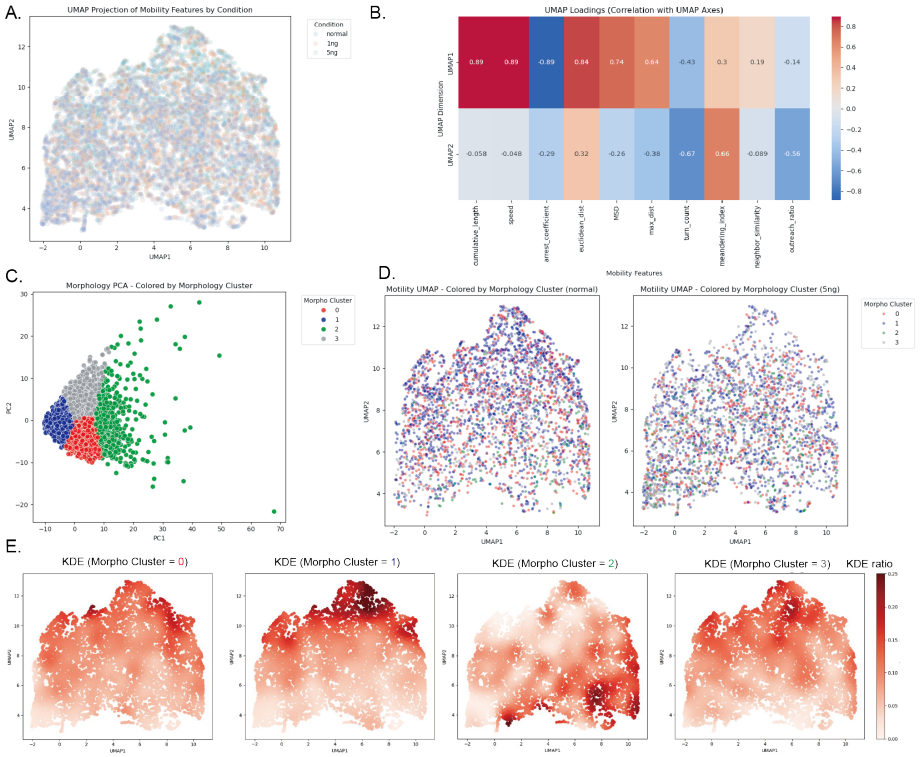

Figure 5: The data distribution in UMAP motility space regarding the condition and cell morphology. (A) UMAP projection of motility features based on the TGF-β1 condition, (B) Correlation map between motility features and UMAP component, (C) Clusters of morphological features of cell in the PC domain, (D) the Distribution of data according to the morphological cluster in control and TGF-β1 5ng/ml condition, (E) KDE plot for each morphological cluster of cells from the control condition

5 Experiments

We assessed the proposed framework through a series of studies. Unless otherwise stated, all models were trained with AdamW optimizer[24] under a mean-squared error objective, dropout rate was set to 0.2, a fixed learning rate of 1×10^{-4}, a batch size of four graphs, and 2000 epochs. This schedule is intentionally set to encourage the revealing of overfitting behaviour, through inspection of loss trajectories. Performance is reported as scatter plots of standardised ground truth versus predicted movement magnitudes (perfect agreement lies on the red dashed diagonal), and the correlation between the true and predicted values in either training or testing set.

5.1 Predicting Migration Magnitude on Simulated Data

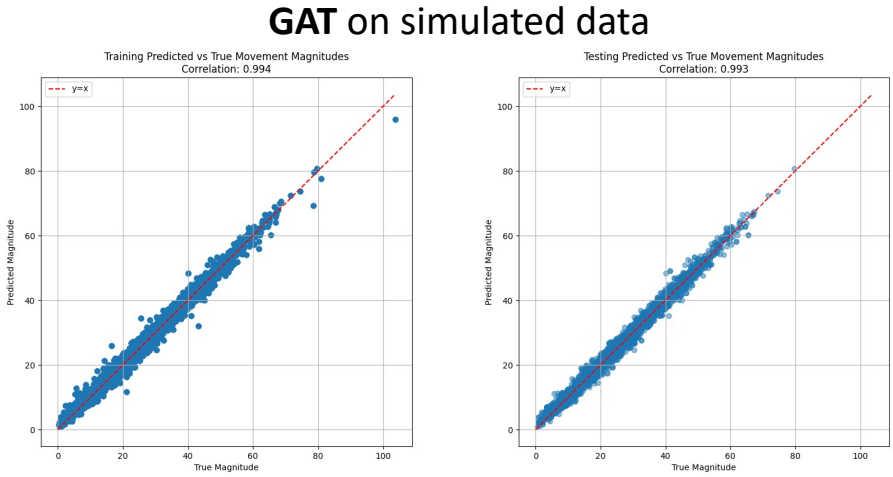

Figure 6: True versus predicted displacement magnitudes for the GAT model trained on self-propelled particle simulations. Both training and testing performance are shown.

To first evaluate the GAT's ability to learn collective behaviors in a controlled setting, we generated synthetic trajectories as described in Section 4.1. We used parameters that more closely match the experimental data we observed: $L = 1000, n = 500, r = 50$, noise level σ of 0.3 (see Methods), $v1 = 50, v2 = 20$ (see Methods). Displacement magnitude at the next timestep calculated from the mean speed is set as the output target:

$$d_i^{t+1} = \|\mathbf{d}_i^{t+1}\|_2 = \|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2$$

for $\mathbf{x}_i^t$ denotes the 2D position vector of cell i at time t. Using only these morphological inputs, we trained a GAT model. As shown in Fig. 6, the model

achieves near-perfect performance on both the training and held-out test sets (correlation > 0.99), indicating that the attention mechanism robustly captures the underlying update rule of the simulation.

5.2 Predicting Average Migration Magnitude

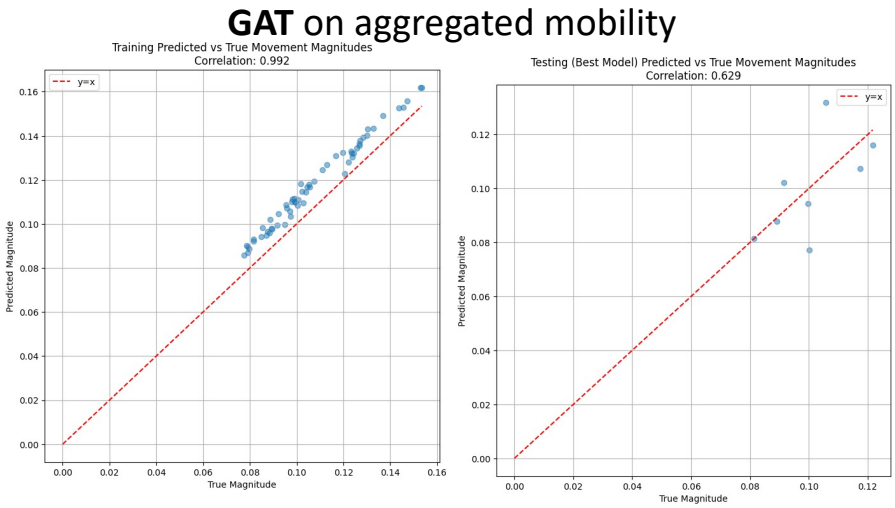

Figure 7: True versus predicted average migration magnitudes from GAT model. Both training and testing performance are shown.

To evaluate the capacity of GATs to predict collective migration, we next replicated the graph-level regression task introduced by Yang et al.[22]. For each cell graph $G = (V, E)$, we defined the target as the mean displacement magnitude $\bar{d} = \frac{1}{|V|} \sum_{i \in V} d_i$, where d_i denotes the migration magnitude of cell i.

To augment our limited experimental dataset, for each graph, we extracted non-overlapping four consecutive frames (one-hour intervals) as the target future frame rather than 12-frame span. We adapted our GAT by inserting a global mean-pooling layer after the final attention block, followed by a linear output head to regress $\bar{d}$. As illustrated in Fig. 7, the model achieved near-perfect agreement on the training graphs (correlation $>$0.99) and retained strong generalization on the held-out test graphs, despite the small number of samples. These results confirm that our GAT model can effectively learn and predict the global mean migration magnitude of cell collectives from static morphological snapshots.

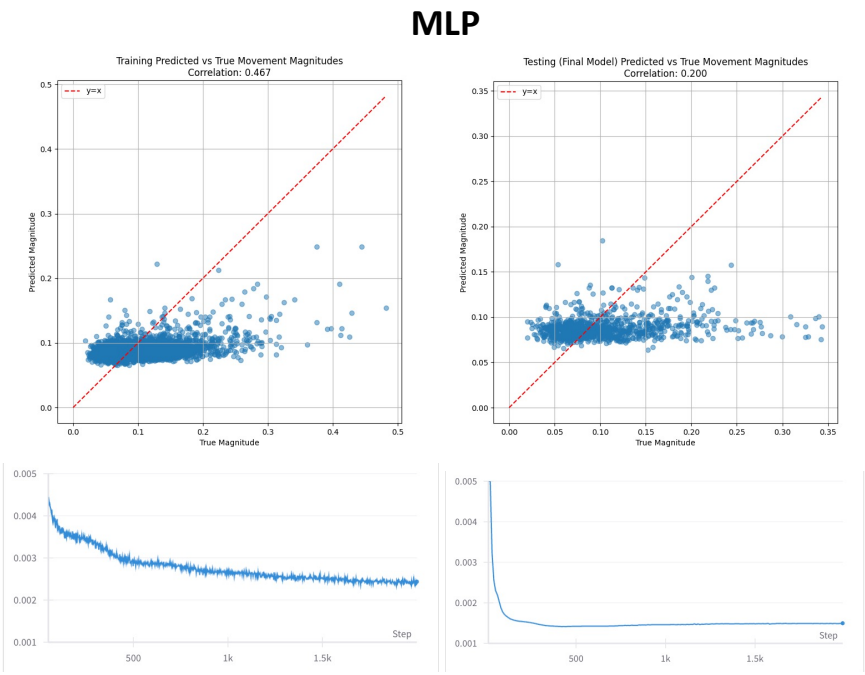

Figure 8: Performance of a MLP model using only morphological features to predict cell migration magnitude. The model fails in modeling movement as it performs poorly during training

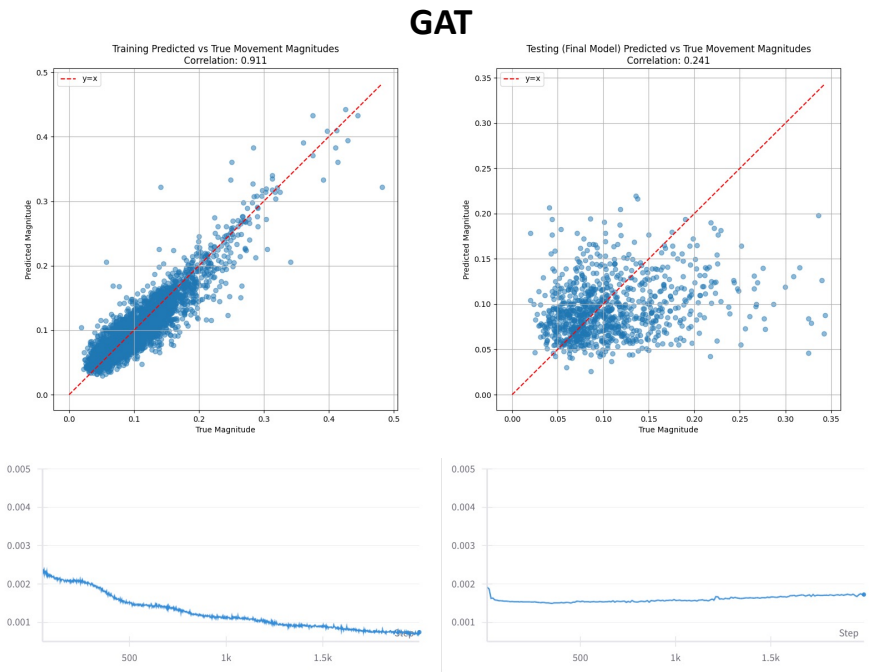

Figure 9: Performance of a GAT model using morphological features as node features to predict cell migration magnitude. The model converges in training but performs poorly during testing

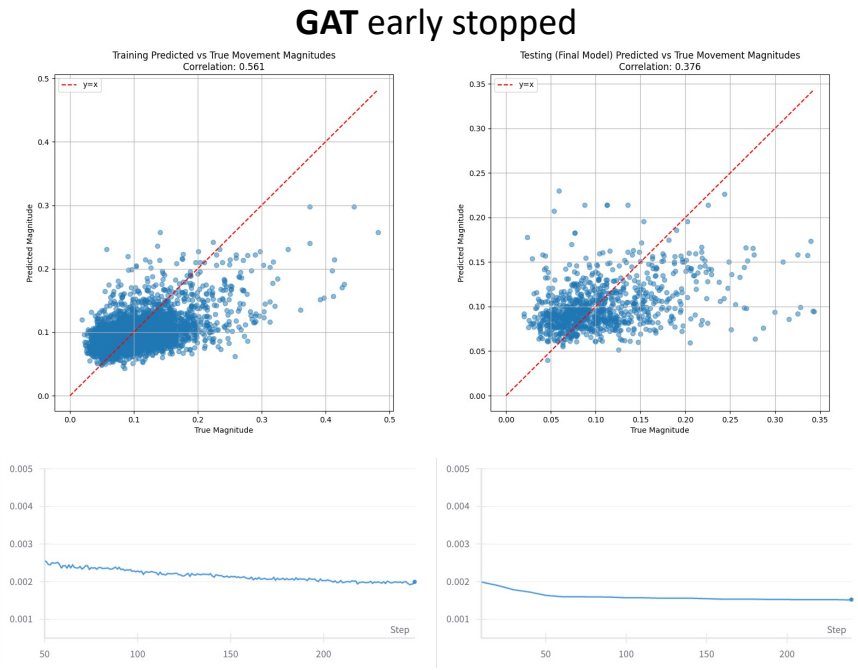

Figure 10: Performance of a GAT model using morphological features as node features to predict cell migration magnitude. The model converges in training but performs poorly during testing

5.3 Predicting Migration Magnitude with Asocial Morphological Features

Moving on to experimental data, we first seek to demonstrate the necessity of considering inter-cellualr social interactions for predicting accurate single-cell level migration magnitude. We format this quesiton as given a model with only the morphological features as input and no graph structure at all to predict the cell migration magnitude as a single scaler (i.e., the distance a cell would travel) over a time span of 12 frames:

$$d_i \;=\; \sum_{t=0}^{12} \left\| \mathbf{x}_i^{t+1} - \mathbf{x}_i^{t} \right\|_2,$$

We trained a Multi-Layer Perceptron (MLP) model that takes only the morphological features of individual cells as input(full list in Methods), without any information about neighboring cells or spatial context. Our results, shown in Figure 8, demonstrates that it failed to learn a predictive mapping from morphological features to movement magnitudes, as it performed poorly not only during testing but also training.

Thus we replaced the MLP with a GAT of similar size (i.e., $\sim$ 100k parameters). After 2000 epochs the model reached an in sample correlation of $r > 0.9$, yet generalisation collapsed on the held-out testing set ($r = 0.24$; Fig. 9). This indicates severe over-fitting. We then used early stopping at the empirical validation optimum (epoch$\sim$ 250) provided a modest gain ($r = 0.38$) but remained inadequate (Fig. 10).

5.4 2D Migration Vector Prediction

To assess the model's ability to capture directional information, we modified the GAT architecture used above so that its final layer outputs the two dimensional displacement vector:

$$\mathbf{d}_i = \sum_{t=0}^{12} |\mathbf{x}_i^{t+1} - \mathbf{x}_i^{t}|$$

for each cell in the next time point. All other model architecture and training hyperparameters were kept identical. For consistency with our previous presentation, we first computed the magnitude of both true and predicted vectors (computed at plotting time) and plotted them against each other (Fig. 11). In the training set, the correlation between true and predicted magnitudes reached 0.515(note this is trained for 2000 epochs), whereas on the held out testing set it fell to 0.221. This decline, more significantly in the training set, relative to scalar magnitude-only prediction (Section 5.3) indicates some additional difficulty in learning directional vector outputs.

Next, we examined the two components separately by plotting true versus predicted values for $\mathbf{d}_i$ (Fig. 12). In training, correlations were 0.536 and 0.497; in testing, 0.149 and 0.218. The similarity of performance across both

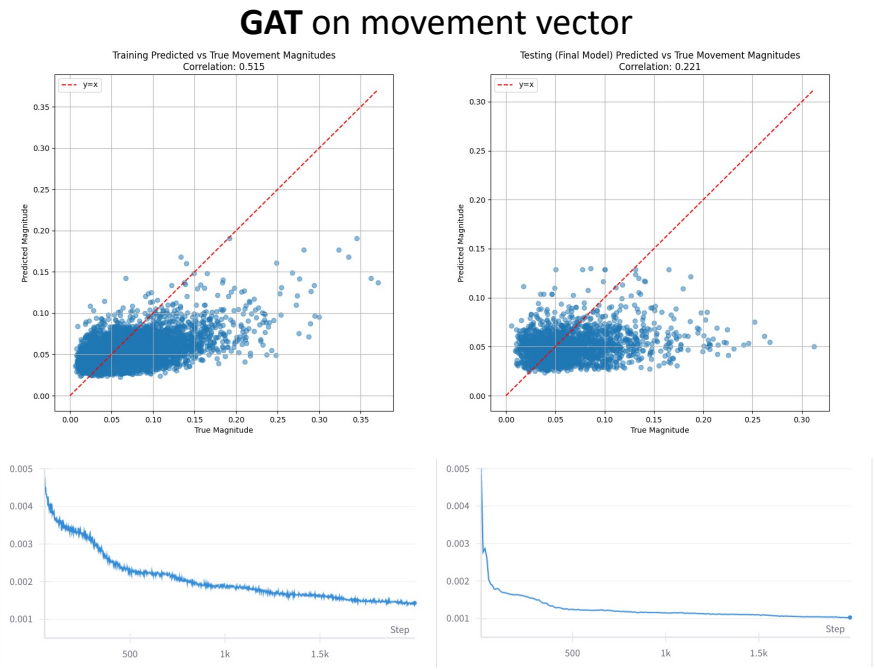

Figure 11: Performance of the GAT model in predicting cell migration vectors. The scatter plot compares the magnitudes of the true and predicted movement vectors computed for visualization.

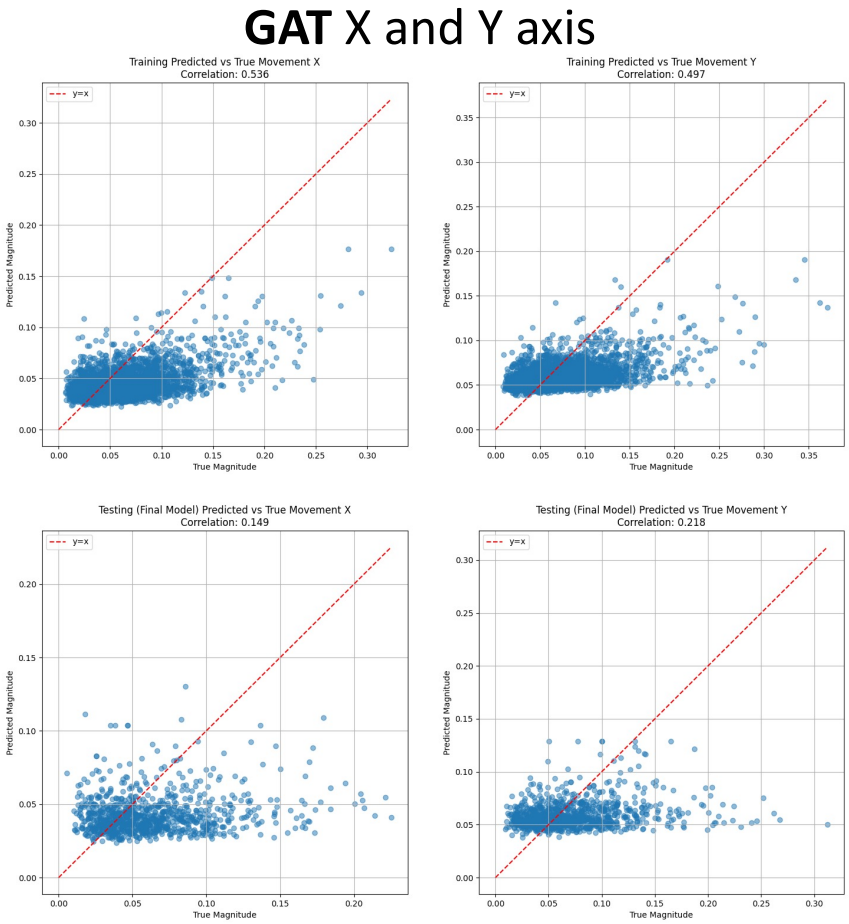

Figure 12: Training and testing performance of the GAT model in predicting cell migration vectors, with X- and Y-axis components shown separately

axes implies that the GAT layers extract a shared representation that supports prediction of each directional component without bias toward either.

Taken together, these results show that predicting the full 2D movement vector is only marginally more challenging than predicting scalar displacement alone. The bulk of feature extraction and relational prediction occurs within the message-passing layers of the GAT, allowing the final linear layer to efficiently map to either magnitude or vector components.

5.5 Predicting UMAP Embeddings of Trajectory Features

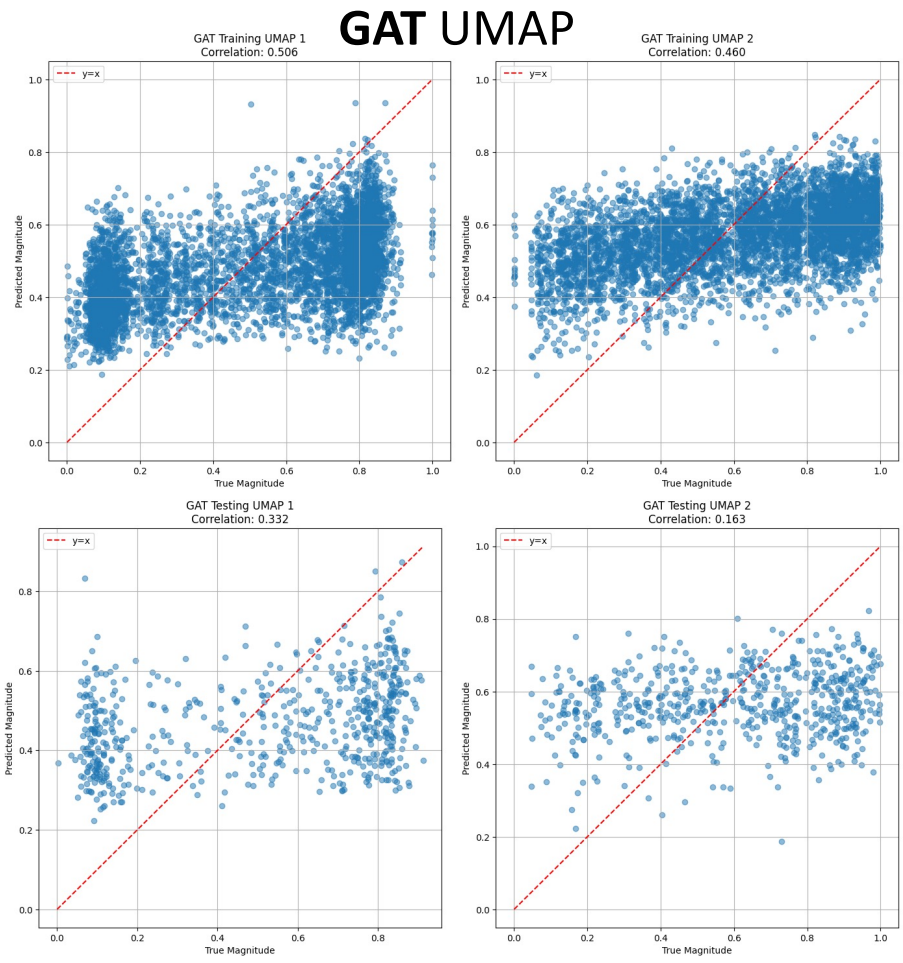

Figure 13: Training and testing performance of the GAT model in predicting 2D UMAP embeddings of cell trajectory features. True versus predicted coordinates are shown separately for each embedding dimension

We next trained the model to predict the two-dimensional UMAP embedding of each cell's trajectory feature vector (cumulative displacement, turn count, maximum displacement, etc.; see Methods). UMAP [25] provides a compact, nonlinear projection that highlights major modes of migratory behavior across cells, and successful prediction would imply that morphology and local graph structure encode these broader trajectory phenotypes.

We kept the model architecture and training hyperparameters identical to Sections above, except that the final linear layer now outputs the pair (u_1, u_2) corresponding to the cell's trajectory UMAP coordinates. As shown in Fig. 13, the model made decent predictions for these embeddings in the training set, achieving correlations about 0.5. However, performance on the held-out test set falls, indicating pronounced overfitting. Early stopping (monitoring validation loss) yielded only marginal improvements in generalization and did not alleviate this gap.

Interestingly, the first embedding dimension has a slightly better results than the second UMAP dimension in both training and testing, suggesting that there are noticable difference between the two dimensions in terms of difficulty in modeling or degree of noisiness. The large drop from training to test performance likely reflects non-generalizable complex mapping from morphology plus neighborhood topology to UMAP space. tTe trade off between training and testing performance persists across various experiments.

6 Discussion

In this study, we have demonstrated the potential and the current limitations of Graph Attention Networks (GATs) for predicting single-cell migration behaviors from static snapshots of epithelial-mesenchymal transition (EMT) cultures. On synthetic self-propelled particle simulations, our GAT model effectively recovered the underlying interaction rules, achieving near perfect generalization on held-out data. When predicting the average migration magnitude on the experimental data, our model again demonstrated promising performance. In contrast, applying the same architecture to experimental microscopy-derived graphs for single-cell level migration prediction led to pronounced overfitting, evident across scalar displacement, 2D vector, and UMAP-embedded trajectory predictions.

We attribute this strong overfitting to several interrelated factors. First, our experimental dataset comprises only a dozen graphs with a few hundred cells each, offering limited coverage of the high-dimensional morphology-topology space; as a result, the model's $\sim 100k$ parameters tend to memorize training examples rather than extract broadly applicable patterns. Second, static single-timepoint graphs and handcrafted morphological metrics fail to capture the dynamic and heterogeneous cues of cell-cell interactions that drive migration, making the input-output mapping intrinsically noisy. Lastly, segmentation and manually defined feature-extraction errors introduce label noise that GATs can inadvertently learn, exacerbating overfitting.

Moreover, it is important to acknowledge that predicting single-cell migration from static snapshots is a difficult task. Biological variability, stochastic fluctuations in cell behavior, and unmeasured microenvironmental factors such as extracellular matrix heterogeneity, paracrine signaling, and mechanical forces contribute substantial noise and randomness at the level of individual cells. These uncontrolled variables place an upper bound on the achievable predictive accuracy of any static-graph approach.

To advance this line of inquiry, future work should focus on improving data representation method that captures more nuanced, yet pivotal information implicitly. For example include longer temporal context of input, using temporal graph and corresponding model architecture. Another potential method is to use more high-dimensional node features, such as learned embeddings from raw image patches or segmentation mask crops via pretrained convolutional image feature extractors, to capture contextually relevant cues. Systematic exploration of simulation parameters and noise regimes may also clarify the boundaries within which GATs can generalize. Addressing these challenges will be essential for harnessing GNNs' full potential to reveal mechanistic insights into collective migration during EMT and, ultimately, for developing predictive tools of practical relevance to tissue remodeling and cancer metastasis.

To push this work forward, our future studies will enhance data representations to encode richer, subtler, but biologically critical information. Incorporating longer temporal context via dynamic graph constructions and time-aware architectures (e.g. temporal GNN) will allow the model to leverage sequential cell-tracking cues. Likewise, enriching node attributes with high-dimensional embeddings, potentially derived from raw image patches or segmentation mask crops using pretrained vision model, could capture both individual morphological and contextual signals that static metrics might miss. By systematically exploring simulation parameters and noise levels, we can better identify the conditions under which GATs generalize effectively. Tackling these challenges will be key to unlocking GNNs' full potential for uncovering the mechanistic drivers of collective migration in EMT and for building predictive tools with genuine clinical and engineering impact.

7 Methods

7.1 Cell Culture

Human mammary epithelial cells(MCF10A) transfected with Z-CAD dual sensors expressing d2GFP and dsRed were generous gift from M.J. Toneff and J. M. Rosen. The Z-CAD dual sensor is compsed with the destabilized GFP containing ZEB1 3' UTR and RFP on the E-cadherin(CDH1) promoter. The MCF10A cell line was cultured in growth medium made with DMEM / F12 (Invitrogen, 11965-118) with horse serum 5% (Invitrogen, 16050-122), 20ng/ml human epidermal growth factor (R&D Systems, 236-EG), 10 μg/mL insulin (Sigma Aldrich, I-1882), 0.5 μg/mL hydrocortisone (Sigma Aldrich, H-0888),

100 ng/mL cholera toxin (Sigma Aldrich, C-8052), and 1% penicillin streptomycin (Invitrogen No. 15070-063).

MCF10A cells were subjected to epithelial-mesenchymal transition(EMT) induction using various concentrations of recombinant TGF-β1 (R&D Systems, 240-B002) in growth media. Cells were cultured for 7 days in T25 tissue culture dishes. To maintain cell-cell contact under normal conditions, cells were seeded at a density of 5×10^5 per dish. For EMT induction with TGF-β1, cells were seeded at a lower density of 2×10^5 cells per dish in media supplemented with 1ng/ml and 5ng/ml of recombinant human TGF-β1. This reduction of cell density was intended to minimize cell-cell junctions and enhance the EMT process. To assess the total number of cells and analyze the cytoplasmic morphology, we also pretreated the MCF10A with 1μ M CellTracker$^{\text{TM}}$ Deep Red dye (Invitrogen, C34565) prior to seeding them in the glass bottom 96 well, the imaging plate in our experiments. Additionally, to track the single-cell trajectories through nuclear movements, cells were treated with 5μg/ml of Hoechst 33342 just before starting imaging.

7.2 Cell boundary segmentation

Cell segmentation was conducted using the CellPose algorithm [26, 27], a deep learning-based open-source tool designed to generalize across various cell types and experimental setups. We firstly employed the pre-trained model Cyto3, the super generalist model, available within the CellPose software package (version 3.0.7) to track the cytoplasm channel of cell images. Before loading images into the software, cell images containing separate cytoplasmic and nuclear channels were merged into an RGB format using ImageJ software. The cytoplasmic channel was assigned to the red channel, while the nucleus channel was allocated to the blue channel of the RGB image. Then, we used two color channels for better segmentation of images the main segmented channel was red channel and the blue channel was added as an optional channel, with the 40 pixel as a target diameter. Post-segmentation, results were manually reviewed and corrected for any segmentation errors, such as under-segmentation or touching cells, to ensure the accuracy of the morphological data extracted for subsequent analysis.

7.3 Cell shape feature extraction

The shape feature extraction were calculated from the segmented cell mask by adapting methodologies in Bhaskar, Dhananjay et al.[28]. In addition to the geometrical analysis, we also utilized the zernike moments to get rotational invariant features of cell boundary.

7.4 Morphological and Trajectory Features

The morphological features we used are the following:

- Area: $A = \sum_{(x,y)\in \text{mask}} 1$

- Perimeter (polygon): $P = \sum_{i=1}^{n} \left\| \mathbf{p}_i - \mathbf{p}_{i+1} \right\|, \quad \mathbf{p}_{n+1} = \mathbf{p}_1$
- Ellipse Eccentricity: $e = \sqrt{1 - \left(\frac{b}{a}\right)^2}$, where a and b are semi-major and semi-minor axes
- Ellipse Major Axis Length: $2a$
- Ellipse Minor Axis Length: $2b$
- Ellipse Variance: $\frac{\sigma}{\mu}$, normalized standard deviation of distance from fitted ellipse
- Compactness: $\text{Compactness} = \frac{\sqrt{\frac{4A}{\pi}}}{\text{Feret}_{\max}}$
- Elongation: $\text{Elongation} = 1 - \frac{\text{Feret}_{\min}}{\text{Feret}_{\max}}$
- Circularity: $\text{Circularity} = \frac{4\pi A}{P^2}$, where P is the perimeter
- Extent: $\text{Extent} = \frac{\text{Area}}{\text{Bounding Box Area}}$
- Skewness: $\text{Skewness}_x = \frac{\frac{1}{n}\sum (x_i - \bar{x})^3}{\left(\frac{1}{n}\sum (x_i - \bar{x})^2\right)^{3/2}}$, and similarly for y
- Kurtosis: $\text{Kurtosis}_x = \frac{\frac{1}{n}\sum (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum (x_i - \bar{x})^2\right)^2}$, and similarly for y

Let $\mathbf{x}_i^t$ denote the 2D position of cell i at time t. The trajectory features we used are the following:

- Euclidean Distance

$$d_{\text{euclidean}} = \|\mathbf{x}^{\text{final}} - \mathbf{x}^{\text{start}}\|$$

- Cumulative Length

$$L_{\text{cumulative}} = \sum_{t=0}^{\text{final-1}} \|\mathbf{x}^{t+1} - \mathbf{x}^{t}\|$$

- Meandering Index

$$\text{Meandering Index} = \frac{d_{\text{euclidean}}}{L_{\text{cumulative}}}$$

- Outreach Ratio

$$\text{Outreach Ratio} = \frac{\max\left(\|\mathbf{x}^{t+1} - \mathbf{x}^{t}\|\right)}{L_{\text{cumulative}}}$$

- Mean Squared Displacement (MSD)

$$\text{MSD} = \frac{1}{|\text{final}|} \sum_{t=0}^{\text{final}-1} \|\mathbf{x}^{t+1} - \mathbf{x}^{t}\|^2$$

- Maximum Distance

$$\text{Max Distance} = \max_{t} \|\mathbf{x}^{t+1} - \mathbf{x}^{t}\|$$

- Arrest Coefficient

$$\text{Arrest Coefficient} = \frac{\text{count}\{i : \|\mathbf{x}^{t+1} - \mathbf{x}^{t}\| < \delta\}}{t}$$

- Turn Count

$$\text{Turn Count} = \text{count}\left\{i : \theta_i > 60^\circ\right\}, \theta_i = \arccos \frac{(\mathbf{x}^{t} - \mathbf{x}^{t-1}) \cdot (\mathbf{x}^{t+1} - \mathbf{x}^{t})}{\|\mathbf{x}^{t} - \mathbf{x}^{t-1}\| \, \|\mathbf{x}^{t+1} - \mathbf{x}^{t}\|}.$$

- Neighbor Similarity

$$\text{Neighbor Similarity} = \frac{1}{t} \sum_{i=1}^{t} \cos\theta_i = \frac{\vec{v}_i \cdot \vec{v}_{\text{neighbor},i}}{\|\vec{v}_i\| \|\vec{v}_{\text{neighbor},i}\|}$$

7.5 Simulation Data Generation

Let $L > 0$ denote the side length of a square domain, $v_1, v_2 > 0$ the speeds of the two cell types, $n \in \mathbb{N}$ the number of cells, $\sigma > 0$ the noise standard deviation, and $r > 0$ the neighborhood radius. We fist assign each cell an initial location by sample

$$\mathbf{x}_i \sim \mathcal{U}([0, L]^2), \quad i = 1, \ldots, n,$$

and assign each cell a phenotype

$$\tau_i \sim \text{Bernoulli}(0.5), \quad v_{\tau_i} = \begin{cases} v_1, & \tau_i = 1, \\ v_2, & \tau_i = 0. \end{cases}$$

Each cell has an initial velocity by sampling directions $\theta_i \sim \mathcal{U}([0, 2\pi])$ and set

$$\mathbf{v}_i = v_{\tau_i} \begin{pmatrix} \cos\theta_i \\ \sin\theta_i \end{pmatrix} + \boldsymbol{\eta}_i, \quad \boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, v_{\tau_i}\sigma^2 I_2)$$

Each cells' morphological features are simulated as: for one-hot encoding $\mathbf{e}_1 = (1, 0)^\mathsf{T}$, $\mathbf{e}_0 = (0, 1)^\mathsf{T}$, let

$$\mathbf{m}_i = \mathbf{e}_{\tau_i} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_2).$$

Neighbors of cell i is defined as

$$\mathcal{N}_i = \big\{\, j : \|\mathbf{x}_i - \mathbf{x}_j\| < r \big\},$$

In this simulation we assume that cell's migration behavior is determined by an averge of its own velocity and the mean of its neighbors' velocity. The mean velocity of the neighbors is

$$\bar{\mathbf{v}}_{\mathcal{N}_i} = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbf{v}_j.$$

Each cells' target movement vector in the future frame after a unit time, is then defined as

$$\mathbf{d}_i = \tfrac{1}{2}\big(\mathbf{v}_i + \bar{\mathbf{v}}_{\mathcal{N}_i}\big) + \boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_2).$$

The graph representationis constructed in the same way as the the rest of the work with experimental data, with a neighborhood radius equal to r, and the target is the magnitude of the target movement vector defined above, to stay consistent with the migration magnitude experiment.

7.6 Graph Representation

Empirically, we set the neighborhood radius to 75 μm based on the measured mean cell radius and intercellular spacing, which yielded an average node degree of 8.06. For each graph, the target time point was chosen as the 12th frame ($\sim$3 hours later), reducing stochastic variability by integrating cell displacements over a long-enough interval. To ensure independent samples, we extracted a single, non-overlapping graph per recording, resulting in 12 total graphs: 8 for training, 2 for validation, and 2 for testing. On average, each graph contained approximately 416 nodes and 3,352 edges.

7.7 Neural Network Training

All models were customized from their original descriptions and implemented in Python using `PyTorch` [29] and `PyTorch Geometric` [30]. Training was carried out on a single GPU hosted by Brown University's Center for Computation and Visualization (CCV) cluster.

References

[1] Jing Yang et al. "Guidelines and definitions for research on epithelial–mesenchymal transition". In: *Nature Reviews Molecular Cell Biology* 21.6 (2020), pp. 341–352. DOI: 10.1038/s41580-020-0237-9. URL: https://doi.org/10.1038/s41580-020-0237-9.

[2] Jean Paul Thiery et al. "Epithelial-Mesenchymal Transitions in Development and Disease". In: *Cell* 139.5 (2009), pp. 871–890. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.11.007. URL: https://www.sciencedirect.com/science/article/pii/S0092867409014196.

[3] M. Angela Nieto. "Epithelial Plasticity: A Common Theme in Embryonic and Cancer Cells". In: *Science* 342.6159 (2013), p. 1234850. ISSN: 0036-8075. DOI: 10.1126/science.1234850.

[4] Elena Scarpa and Roberto Mayor. "Collective cell migration in development". In: *Journal of Cell Biology* 212.2 (2016), pp. 143–155. ISSN: 0021-9525. DOI: 10.1083/jcb.201508047.

[5] Raghu Kalluri and Robert A. Weinberg. "The basics of epithelial-mesenchymal transition". In: *Journal of Clinical Investigation* 119.6 (2009), pp. 1420–1428. DOI: 10.1172/JCI39104.

[6] Samy Lamouille, Jian Xu, and Rik Derynck. "Molecular mechanisms of epithelial-mesenchymal transition". In: *Nature Reviews Molecular Cell Biology* 15.3 (2014), pp. 178–196. DOI: 10.1038/nrm3758.

[7] Jean-Paul Thiery et al. "Epithelial-mesenchymal transitions in development and disease". In: *Cell* 139.5 (2009), pp. 871–890. DOI: 10.1016/j.cell.2009.11.007.

[8] Kevin J. Cheung and Andrew J. Ewald. "A collective route to metastasis: Seeding by tumor cell clusters". In: *Science* 352.6282 (2016), pp. 167–169. ISSN: 0036-8075. DOI: 10.1126/science.aaf6546.

[9] Mohit Kumar Jolly et al. "Implications of the hybrid epithelial/mesenchymal phenotype in metastasis". In: *Frontiers in Oncology* 5 (2015), p. 155. ISSN: 2234-943X. DOI: 10.3389/fonc.2015.00155.

[10] Ievgenia Pastushenko et al. "Identification of the tumour transition states occurring during EMT". In: *Nature* 556.7702 (2018), pp. 463–468. ISSN: 0028-0836. DOI: 10.1038/s41586-018-0040-3.

[11] Peter Friedl and Darren Gilmour. "Collective cell migration in morphogenesis, regeneration and cancer". In: *Nature Reviews Molecular Cell Biology* 10.7 (2009), pp. 445–457. ISSN: 1471-0072. DOI: 10.1038/nrm2720.

[12] James LaChance et al. "Learning the rules of collective cell migration using deep attention networks". In: *PLoS Computational Biology* 18.4 (2022), e1009293. DOI: 10.1371/journal.pcbi.1009293.

[13] Franco Scarselli et al. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. ISSN: 1045-9227. DOI: 10.1109/TNN.2008.2005605.

[14] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *International Conference on Learning Representations (ICLR)*. 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[15] William L. Hamilton, Rex Ying, and Jure Leskovec. "Inductive Representation Learning on Large Graphs". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. URL: https://papers.nips.cc/paper/6703-inductive-representation-learning-on-large-graphs.

[16] Petar Veličković et al. "Graph Attention Networks". In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://openreview.net/forum?id=rJXMpikCZ.

[17] Francisco J. H. Heras et al. "Deep attention networks reveal the rules of collective motion in zebrafish". In: *PLOS Computational Biology* 15.9 (2019), e1007354. DOI: 10.1371/journal.pcbi.1007354. URL: https://doi.org/10.1371/journal.pcbi.1007354.

[18] Qianwen Lu, Zhiheng Zhou, and Qi Wang. "Multi-layer graph attention neural networks for accurate drug-target interaction mapping". In: *Scientific Reports* 14 (2024), p. 26119. DOI: 10.1038/s41598-024-75742-1.

[19] Qiuyue Yuan and Zhana Duren. "Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data". In: *Nature Biotechnology* 43 (2025), pp. 247–257. DOI: 10.1038/s41587-024-02182-7.

[20] Haocheng Gu et al. "scGNN 2.0: a graph neural network tool for imputation and clustering of single-cell RNA-Seq data". In: *Bioinformatics* 38.23 (2022), pp. 5322–5325.

[21] Albane Ruaud et al. *Modelling Microbial Communities with Graph Neural Networks*. 2024. URL: https://openreview.net/forum?id=cXs5md5wAq.

[22] Haiqian Yang et al. "Learning collective cell migratory dynamics from a static snapshot with graph neural networks". In: *PRX Life* 2.4 (2024), p. 043010.

[23] Tamas Vicsek et al. "Novel type of phase transition in a system of self-driven particles". In: (2006). DOI: 10.1103/PhysRevLett.75.1226. arXiv: cond-mat/0611743 [cond-mat.stat-mech]. URL: https://arxiv.org/abs/cond-mat/0611743.

[24] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *arXiv preprint arXiv:1711.05101* (2017). arXiv: 1711.05101 [cs.LG]. URL: https://arxiv.org/abs/1711.05101.

[25] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML]. URL: https://arxiv.org/abs/1802.03426.

[26] Carsen Stringer et al. "Cellpose: a generalist algorithm for cellular segmentation". In: *Nature methods* 18.1 (2021), pp. 100–106.

[27] Carsen Stringer and Marius Pachitariu. "Cellpose3: one-click image restoration for improved cellular segmentation". In: *Nature Methods* (2025), pp. 1–8.

[28] Dhananjay Bhaskar et al. "A methodology for morphological feature extraction and unsupervised cell classification". In: *bioRxiv* (2019). DOI: 10.1101/623793. eprint: https://www.biorxiv.org/content/early/2019/04/30/623793.full.pdf. URL: https://www.biorxiv.org/content/early/2019/04/30/623793.

[29] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *arXiv preprint arXiv:1912.01703* (2019). DOI: 10.48550/arXiv.1912.01703.

[30] Matthias Fey and Jan Eric Lenssen. "Fast Graph Representation Learning with PyTorch Geometric". In: *Proceedings of the ICLR 2019 Workshop on Representation Learning on Graphs and Manifolds*. New Orleans, LA, USA, 2019. URL: https://arxiv.org/abs/1903.02428.