Computer Methods and Programs in Biomedicine 255 (2024) 108337

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

A self-supervised embedding of cell migration features for behavior discovery over cell populations

Miguel Molina-Moreno [a,b,*], Iván González-Díaz [a], Ralf Mikut [c], Fernando Díaz-de-María [a]

[a] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, Leganés, 28911, Spain
[b] Department of Immunobiology, Yale University, Amistad Street Building, 10 Amistad St, New Haven, 06520, USA
[c] Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz, 1, Eggenstein-Leopoldshafen, 76344, Baden-Württemberg, Germany

ARTICLE INFO

Keywords:
Cell migration
Self-supervised learning
Recurrent Neural Networks
Behavior discovery

ABSTRACT

Background and objective: Recent studies point out that the dynamics and interaction of cell populations within their environment are related to several biological processes in immunology. Hence, single-cell analysis in immunology now relies on spatial omics. Moreover, recent literature suggests that immunology scenarios are hierarchically organized, including unknown cell behaviors appearing in different proportions across some observable control and therapy groups. These dynamic behaviors play a crucial role in identifying the causes of processes such as inflammation, aging, and fighting off pathogens or cancerous cells. In this work, we use a self-supervised learning approach to discover these behaviors associated with cell dynamics in an immunology scenario.

Materials and methods: Specifically, we study the different responses of control group and therapy groups in a scenario involving inflammation due to infarct, with a focus on neutrophil migration within blood vessels. Starting from a set of hand-crafted spatio-temporal features, we use a recurrent neural network to generate embeddings that properly describe the dynamics of the migration processes. The network is trained using a novel multi-task contrastive loss that, on the one hand, models the hierarchical structure of our scenario (groups-behaviors-samples) and, on the other, ensures temporal consistency within the embedding, enforcing that subsequent temporal samples obtained from a given cell stay close in the latent space.

Results: Our experimental results demonstrate that the resulting embeddings improve the separability of cell behaviors and log-likelihood of the therapies, when compared to the hand-crafted feature extraction and recent methods from the state of the art, even with dimensionality reduction (16 vs. 21 hand-crafted features).

Conclusions: Our approach enables single-cell analyses at a population level, being able to automatically discover shared behaviors among different groups. This, in turn, enables the prediction of the therapy effectiveness based on their proportions within a study group.

1. Introduction

Spatial omics, which studies the cell dynamics and interaction with their environment, has become an essential tool in immunology. This technique allows for the discovery and description of cell behaviors in scenarios of inflammation, aging or cancer [1,2]. In these scenarios, a cell behavior is defined as a way of conduct of a group of cells that share some morpho-kinetic properties, e.g., small ellipsoid cells that do not move with the blood flow, or large sessile cells that move close to the blood vessel wall. In this way, cell behaviors are related to cells' phenotypes and migration patterns, including volume and shape changes (cells have the ability to change their phenotypes to migrate [1]), trajectories (velocity and type of motion can also represent a specific behavior), and their interaction with external cues (such as blood vessels, dendritic cells, etc.) [2].

Specifically, this work focuses on recent biological studies that have pointed out the existence of shared cell behaviors in application scenarios within the field of *in vivo* microscopy. For example, immune cells that migrate within a blood vessel in the presence of inflammation (or within a tumor area in the case of cancer) show behavior proportions that depend on the group which they belong to (e.g. wild-type group, a gene knockout treatment or a drug treatment) [3]. This particular setup occurs in many biological processes [4], and many

* Corresponding author at: Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, Leganés, 28911, Spain.
E-mail addresses: migmolin@ing.uc3m.es (M. Molina-Moreno), igdiaz@ing.uc3m.es (I. González-Díaz), ralf.mikut@kit.edu (R. Mikut), fdiaz@ing.uc3m.es (F. Díaz-de-María).

https://doi.org/10.1016/j.cmpb.2024.108337
Received 20 November 2023; Received in revised form 26 April 2024; Accepted 17 July 2024
Available online 19 July 2024

fields of science, such as: topic modeling in document analysis, where word frequencies depend on the topic of the document [5]; or image retrieval, where visual word frequencies play a similar role [6]. In our target scenario, we consider that each video capture (3D+time) belongs to a given group of biological interest (control groups, or groups receiving a specific treatment), and contains a population of samples (cells) exhibiting different behaviors in proportions that are particular and representative of its group. Our goal is to discover these unknown behaviors and enable a subsequent analysis with the aim of supporting the development and assessment of novel therapies.

In the last years, biologists have generally addressed this task of cell behavior analysis by using commercial software, such as Imaris (Bitplane, South Windsor, CT, USA) or Fiji [7], for assisted hand-crafted feature extraction, and algorithms such as t-SNE [8] or UMAP [9] to subsequently obtain a low-dimensional representation of the data that can be easily analyzed with visual supporting tools. The identification or discovery of latent behaviors is performed either manually, by observing the data distributions, or automatically, using clustering algorithms that discover groups within the data according to a given similarity measure. However, this approach has several limitations:

1. It usually requires manual supervision at some stages of the process, thus limiting the number of samples (cells) that can be analyzed in the studies.
2. The manually-designed feature set may contain irrelevant and redundant features, which can affect the results of the clustering process. Although these features may be common features for comparison of research among different laboratories and experiments, this may lead to a biased analysis, unless the set of hand-crafted features is carefully designed and curated for each specific application. Furthermore, usual dimensionality reduction algorithms (e.g. UMAP or t-SNE) may hide some important data relationships for single-cell analysis [10].
3. Due to potential flaws in the segmentation and tracking processes (which can also happen with automatic approaches) there are cells whose features do not show temporal consistency, i.e., samples coming from the same cell in consecutive instants can exhibit significantly different feature values. Furthermore, although adjacent temporal instants from the same cell are typically similar and correspond to the same behavior, in many biological scenarios, behavior transitions are also frequent in cells (during the time span of the capture a cell can change its behavior, especially with large temporal sequences) [3].

To overcome these limitations, improved data representations should be developed to disentangle and reveal the latent factors that explain the dynamic properties of the data. In particular, the discipline of representation learning proposes the use of reduced-dimensionality data embeddings [11], with the aim of providing more appropriate features to describe the input [12–14].

This paper proposes a *self-supervised learning approach over cell populations* that, starting from an initial set of hand-crafted spatio-temporal features, learns to generate embeddings that encode the dynamics of the cell migration process. To the best of our knowledge, this is the first time that a self-supervised deep architecture has been applied to the discovery of shared cell behaviors across control and therapy groups. Specifically, the main contributions of this work can be summarized as follows:

- *Dynamic representation of cell behaviors*. Our approach integrates a bidirectional Long Short-Term Memory (LSTM) network to encode sequences of low-level features representing the behavior of a cell within a temporal window. As shown in the experiments, the resulting embedding successfully identifies redundant and irrelevant features in the initial set of features, achieving a significant dimensionality reduction, while keeping or even improving the performance of the behavior discovery task and preserving the explainability of the discovered behaviors.
- *Modeling of the hierarchical scenario (groups-cells-samples)*. We propose a multi-task contrastive loss that allows us to train the model in a self-supervised mode and aligns with the hierarchical nature of the data, comparing samples according to the per-group proportions of cell behaviors. In this manner, sample similarity is considered at population level (biological group), rather than cell level, in accordance with the fact that captures from different groups contain cells belonging to a shared set of behaviors. Our proposed multi-task contrastive loss is related to the well-known supervised contrastive loss [15], but it differs in that it does not analyze sample-to-sample similarities. Instead, it focuses on sample-to-population similarities, drawing inspiration from very recent advances in bag-based losses designed for weakly supervised or noisy scenarios [16]. In this way, our approach can adapt to a hierarchical arrangement of groups, behaviors and cells. Furthermore, it can also serves as a means of regularizing the embedding generation, a task that has been traditionally posed as a prediction problem (given one part of a sequence, learning to predict the subsequent one) [17], which does not suit prone-to-outliers scenarios, as the one presented in this paper.
- *Attainment of stable but transitory behaviors*. Moreover, the multi-task loss includes a second term that imposes temporal coherence to the behaviors in the cell trajectories. It does so by enforcing that subsequent temporal instances of a cell remain close in the latent space, while still allowing for cell behavior transitions during the time span of the captures.

The rest of this paper is organized as follows. Section 2 reviews the related literature. In Section 3 we first provide a description of our scenario and employed dataset, then a description of our method, and, lastly, detailed descriptions of each of its main components in the subsequent subsections. Section 4 describes performance metrics, discusses the influence of the hyperparameters of the method and presents the experimental results that support our method in comparison with the state of the art, in terms of: performance in dimensionality reduction, separability and model fitting. Then, Section 5 discusses our results from a biological point of view, putting emphasis on the validity of transitions, explainability and error analysis. And finally, Section 6 summarizes our conclusions and outlines future lines of research.

2. Related work

In this section, we discuss the most relevant previous works on unsupervised sequence modeling and self-supervised learning methods.

2.1. Unsupervised sequence modeling on biological data

Heterogeneity, complexity, dynamics and relationships among features can affect the performance of spatio-temporal data representation tasks [18], but deep learning techniques have proved their effectiveness in modeling different types of time-series, such as event data, trajectory data, audio data or video data [19]. Specifically, Recurrent Neural Networks (RNNs), like LSTMs [20] and Transformers [21], are nowadays widely used for modeling sequences of audio, video and text data.

Regarding unsupervised deep sequence modeling methods, the literature mostly tackles text-related problems, in the context of: translation [22], spelling [23], or text-to-speech/handwriting synthesis [24] tasks. Speech-related applications are also popular, autoencoders are used to summarize the temporal dependencies between observations in automatic speech recognition [25].

On the other hand, with respect to biological data, the literature is mainly focused on the modeling of DNA sequences and protein chains. Moreover, their sequence modeling is usually based on Natural Language Processing (NLP) proposals such as word2vec [26–28]. However, some methods propose their own modeling approaches, such as that of Hill et al. [29], which proposes a gated recurrent neural network

that can learn complex and long-range patterns in full-length human transcripts; or that of Kou et al. [30], which proposes a variational auto-encoder with a 4-dimensional latent space that produces patterns consistent with clinical impressions of esophageal manometry.

Finally, regarding cell behavior profiling, a 2D approach [31] characterizes two cell behaviors in live microglia using a Vector Quantized variational autoencoder (VQ-VAE), which also enforces temporal continuity in a 16-dimensional latent space. However, in this work the behaviors are not shared across groups. A very recent approach [32] proposes a 3D imaging-transcriptomics platform, based on dynamic time warping of cell temporal sequences to obtain a 2D representation with UMAP. The resulting 2D map can be used to identify different behaviors within the data. Neither of these approaches allows for explicit behavior transitions. Hence, they cannot model highly dynamic scenarios of immunology.

2.2. Self-supervised learning

Self-supervised learning is an emerging field in machine learning, that enables to use a supervised loss in tasks lacking of manually annotated data. Instead, algorithms learn from information intrinsically available within the data, but not directly related to the task at hand. These techniques have proven to be particularly effective in the fields of NLP and computer vision. In the field of NLP, many self-supervised tasks have been proposed, mainly involving predicting missing words from a sequence [33]. In image- and video-related problems, domain-specific Convolutional Neural Networks (CNN) are trained with unlabeled data by designing tasks in which supervisory variables can be automatically computed from the data, learning the spatial and temporal structures of the visual content. For example, authors in [34] extract multiple patches from a single image and ask the model to predict the spatial relationship between these patches, whereas the work in [35] validates frame order in video (useful for action recognition).

However, the application of self-supervised learning to the field of life sciences is very recent. Some approaches focus on cell-type identification [36,37] using prior biological information, such as the signaling pathways in genes. Another example combines unsupervised deep sequence modeling and self-supervised learning, using a Transformer to predict a missing fraction of amino acids sequences and protein sequences [38]. The resulting embedding is then analyzed using dimensionality reduction algorithms as t-SNE and PCA [39] to gain insight into the properties of the sequences. In another approach [40], an unsupervised adversarial autoencoder is combined with a self-supervised Latent Dirichlet Allocation (LDA), with the cell type acting as a regularizer. Then, the time-averaged cell descriptors in the 56-dimensional latent space are used as features to distinguish highly metastatic melanoma cells.

In comparison to all these approaches, in our scenario, behaviors are complex and migration mechanisms are related to a heterogeneous set of morpho-kinetic features, such as shape variation, relative position of the cell with respect to the vessel, cell trajectory, etc. Based on the expert knowledge of biologists, we have designed an initial set of spatio-temporal hand-crafted features that describes the dynamics of migration processes. Our proposal, relying on a LSTM-based architecture and a self-supervised method, accomplishes two key objectives: (1) it is capable of generating an embedding which models the short-term behaviors of the cells, improves the behavior separability even with dimensionality reduction, and detects potential irrelevant or redundant features in the initial set; and (2) it considers the temporal consistency in behavior assignment for each cell at the same time that it allows meaningful behavior transitions.

3. Materials and methods

The overview of our approach is shown in Fig. 1. Each 4D capture (3D space +time) containing neutrophils migrating within blood vessels is processed using the ACME software [41], available through [42]. This software performs individual cell segmentation and tracking, yielding a sequence of morpho-kinetic features per neutrophil. These features serve as input to our deep sequence modeling system, which generates the embedding representation. During the training phase, the embedding is optimized using our self-supervised method.

3.1. Materials

In this subsection we first describe our application scenario. While specific to our target task, it has attributes that make it easily extensible to other domains. Our scenario consists of four populations or groups, namely:

1. **wild-type/control**: mice without treatment.
2. **anti-Plt/control**: mice with platelet depletion.
3. **FGR-KO/therapy**: mice transplanted with bone marrow of a knockout mutant for FGR gene.
4. **FGR-INH/therapy**: mice treated with an inhibitor of the FGR protein.

Neutrophils in every group share the same set of behaviors, but in different and unknown proportions. Our objective during the training phase is to learn the set of behaviors over the control groups, assuming that they are applicable to the therapies as well. From a biological perspective, our final goal is to anticipate the effectiveness of a therapy by comparing its proportions of behaviors with those of the control groups.

Biological hypotheses suggest that the behavior proportions in the control groups must be maximally different. In the wild-type group, there must be a large proportion of larger neutrophils, which change their shapes and migrate over the blood vessel surface (pathogenic migration). In contrast, in the anti-Plt group, smaller, spherical and non-migratory neutrophils predominate (non-migratory). Successful therapies must have an intermediate composition between those of wild-type and anti-Plt groups, reducing pathogenic migration in favor of non-pathogenic ways of migration. For a deeper biological analysis, the reader is referred to [3].

Each capture consists of a 4D (3D spatial +time) video containing a set of moving neutrophils within blood vessels. Each cell is first segmented, tracked and parametrized over time using the technical solution described in [41] (which will be briefly described in Section 3.2). This process leads to a final dataset of *in vivo* microscopy that contains 147 4D captures composed of 2334 3D volumes of neutrophils migrating within venules in the cremaster muscles of mice, denoted as $\mathbf{X}$, belonging to the four different groups: wild-type/control, anti-Plt/control, FGR-KO/therapy and FGR-INH/therapy.

3.2. Baseline cell parameterization for behavior characterization

As mentioned, the ACME software segments, tracks, and describes each cell using a hand-crafted set of features (a more detailed description can be found in [41]):

1. First, from an input 3D+time volume, a 3D joint segmentation module generates three outcomes: (1) a mask defining the 3D volume corresponding to the blood vessel; (2) a set of binary masks containing the 3D regions susceptible to be cells; and (3) for each candidate region, its probability of being a well-segmented cell. All these outcomes are provided for every temporal index of the 3D+time volume.
2. Next, the cell masks are fed to the three-pass 3D tracking module. This system analyzes the time sequence of segmented regions and generates individual trajectories.

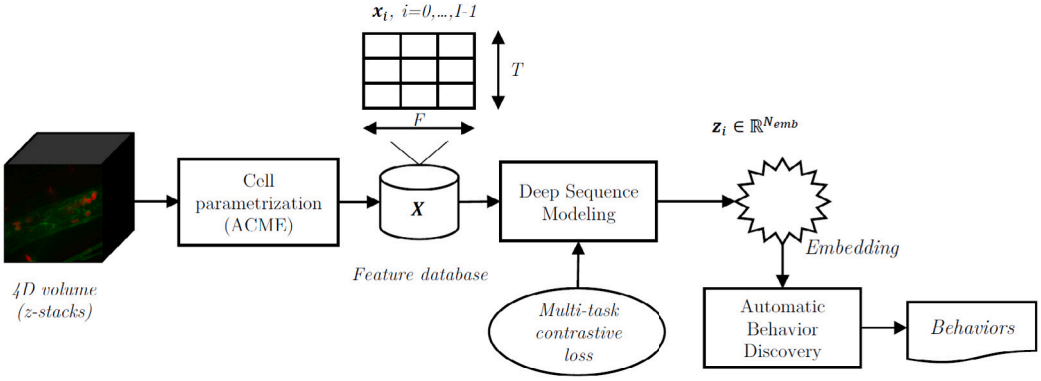


Fig. 1. Overview of the proposed approach. The 4D (3D space +time) captures are segmented, tracked and parametrized, yielding the feature 3D matrix $\mathbf{X} \in \mathbb{R}^{I\times T\times F}$. The deep sequence modeling system, trained with the multi-task contrastive loss, produces the latent space representation $\mathbf{z}_i \in \mathbb{R}^{N_{emb}}$. The automatic behavior discovery system subsequently generates the set of behaviors.

Table 1
Cell morpho-kinetic features used for cell behavioral description.

Index	Feature
1	Volume
2	Superficial area
3–5	Height, maximum width and height/width rate
6–8	Sphericity, prolate and oblate ellipticity
9–11	Principal axes length
12–13	Extent and solidity
14	Equivalent diameter
15–17	Cell X/Y/Z axis orientation respect to the bloodstream
18–19	Cell polar position (radius/angle)
20	Distance from cell center to blood vessel surface
21	Minimum distance from cell to blood vessel surface

3. Then, the feature extraction module characterizes each temporal instance of a cell through 21 instantaneous features, related to the position and shape of the cells at each time instant. The result for each trajectory is a sequence of morpho-kinetic features for each cell. Our set of features is more comprehensive than those from other state-of-the-art tools for dynamic cell behavior description [43,44].
4. Finally, the cell selection module chooses the final set of valid trajectories discarding:
   - incomplete tracks with less than 21 temporal instants, attributed to cells that are not suitable for behavior profiling (considering that recent literature establishes a minimum time span for considering a stable behavior [2–4]);
   - regions for which a valid position with respect to the blood vessel cannot be obtained (those located at the edges of the volume);
   - regions with a size far from the typical size of neutrophils;

   and classifies the remaining cells depending on their sequences of features and their previously computed probabilities of being correctly segmented. The precision at the end of this pipeline reaches a value of 95%.

Hence, the short-term spatio-temporal behavior of a cell i is described using a set of $F = 21$ morpho-kinetic features, computed at each instant t over $T = 21$ instants. It should be noted that every $\mathbf{x}_i$ is the time evolution of an individual cell from the set of I cells, yielding a 3D matrix $\mathbf{X} \in \mathbb{R}^{I\times T\times F}$ where $i = 0, \dots, I-1$ represents the cell track identifiers.

The sequence $\mathbf{x}_i \in \mathbb{R}^{T\times F}$ then becomes an input sample to our system. Table 1 enumerates the features used to parametrize the cell dynamics. It is worth noticing that each tracked cell in a capture may lead to several samples i, each one associated with a short temporal window centered at a different instant. We consider temporal windows with strong overlapping, by shifting the center by one temporal instant and consider cells that change their behavior over time [3].

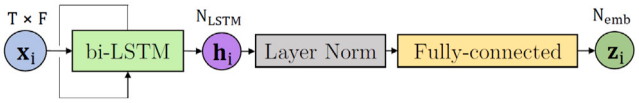


Fig. 2. Deep Sequence Modeling system (see Fig. 1). The input $\mathbf{x}_i$ consists of F features over T time instants. The LSTM produces a representation of size N_{LSTM}, which is compressed to the final N_{emb} dimension.

3.3. Deep sequence modeling

The Deep Sequence Modeling block generates a latent representation, the embedding, from the original features. Its architecture is depicted in Fig. 2. Each temporal sequence $\mathbf{x}_i = [\mathbf{x}_{i,t}], t = 0, \dots, T-1$, representing the short-term behavior of a cell, is fed to a bidirectional LSTM, which produces a unique representation $\mathbf{h}_i \in \mathbb{R}^{N_{LSTM}}$ by taking the output computed at the central instant of the sequence. Next, a *Layer Norm* and a *fully-connected* layer transform the output of the LSTM into the final sequence embedding $\mathbf{z}_i \in \mathbb{R}^{N_{emb}}$.

The goal of our Deep Sequence Modeling system is to obtain a latent representation from the original features that removes redundancies and irrelevant features, exploits inter-feature relationships, and enables behavior discovery in a scenario with two requisites: (1) it has to allow for discriminating different groups at a population level through their behavior proportions; and (2) although allowing for behavior transitions, it must strengthen temporal consistency of the embedding. Both conditions are integrated into a novel multi-task loss function $\mathcal{L} = \mathcal{L}_{BD} + \mathcal{L}_{TC}$, where BD in the first term stands for Behavior Discovery (described in Section 3.3.2), and TC in the second stands for Temporal Consistency (described in Section 3.3.3). Furthermore, since we aim at discriminating among groups at population level, we have designed a loss which involves bags of samples, instead of individual samples, as described in Section 3.3.1.

3.3.1. From samples to bags

Our proposal for behavior discovery relies on the hypothesis that, although any behavior may appear on any capture, the group to which each capture belong (control group or different therapies) strongly determines the proportions of behaviors. In consequence, our training process should be aware of these proportions. To that end, inspired by the approach in [16], we arrange cell samples into bags, and use them

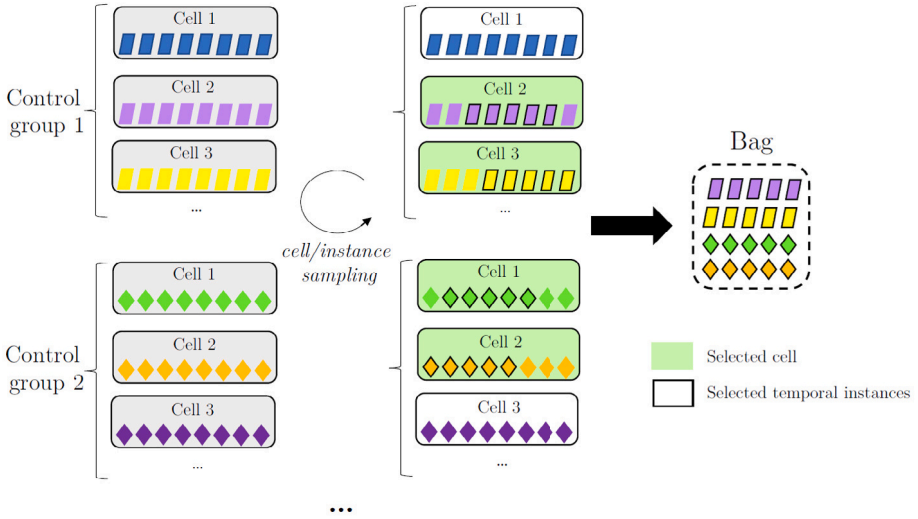


Fig. 3. Overview of the bag composition (example for two control groups, number of samples $N_B = 4$, and number of cell temporal instances $N_C = 5$). First, N_B cell temporal instances are selected through a balanced sampling (the same number for each group). Then, a random temporal instance is selected for every one of the cells. Finally, for each selected cell and temporal instance, N_C neighboring temporal instances are included in the bag. Thus, the number of elements in the bag is $N_B \cdot N_C$.

as batches in our training process. Each bag is composed of N_B samples, and should contain enough cells from each considered group G in an attempt to provide a statistically suitable representation of the behavior proportions.

Moreover, even though two consecutive inputs $\mathbf{x}_{i,t}$ and $\mathbf{x}_{i,t+1}$ from the same cell i differ by just one time instant, nothing explicitly ensures that our model will provide similar representations in the latent space for the both samples (i.e. temporal consistency). In order to ensure temporal consistency in the embedding, our bag must contain several consecutive samples of the cells to calculate the similarities among them.

Fig. 3 shows the way the bags are built. The process consists of two steps:

1. First, we perform a random selection of some cells from each of the groups. It is noteworthy that our approach will work as long as the batch is statistically representative of the real population of cell behaviors. For that end, we need to ensure that the batch size N_B is large enough and that the samples are evenly balanced among the groups. We will discuss this point in the experimental section.
2. Second, from the sequence representation of a given cell, we randomly choose a set of $\{C_i\}$ consecutive instants. Hence, each bag of N_B samples includes $\{C_i\}$ neighboring temporal instances for each selected cell i. In our case $|\{C_i\}| = N_C \ \forall i$. It should be noticed that using only N_C neighboring temporal instances (with $N_C < T$) allows for behavior transitions in our scenario, since the temporal consistency is guaranteed only in an specific range around the selected temporal instance for each cell. The optimal value of N_C will be also discussed in the experimental section.

3.3.2. A bag-based contrastive loss for behavior discovery

In our scenario, any behavior might appear in any of the groups, but the behavior proportions define the nature of the group. Given this assumption, instead of enforcing that a sample i is more similar to samples within its group, which will not be always true (especially when a cell exhibits a behavior underrepresented in its group), we define a contrastive loss that measures this similarity at the population level. Therefore, we define a novel bag-based Behavior-Discovery (BD) loss that, within a bag of size N_B, computes the ratio between S^-, the aggregated distances between negative cell pairs (i.e. cells belonging to different groups), and S^+, the aggregated distance between positive cell pairs within the bag (i.e. cells that belong to the same group):

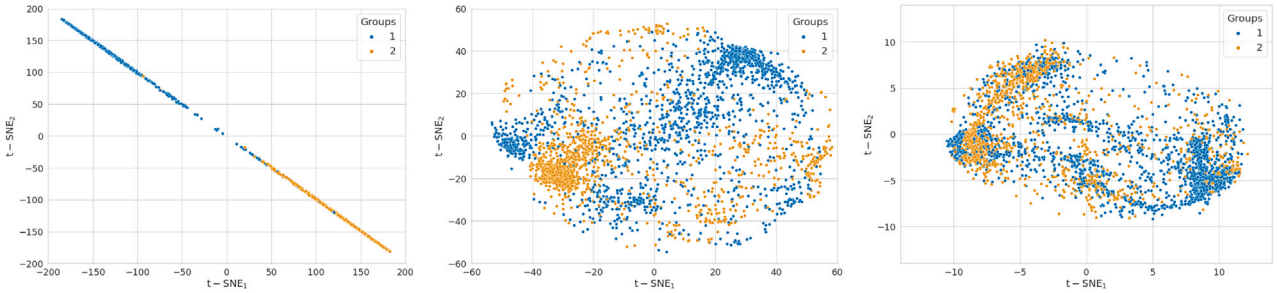
$$\mathcal{L}_{BD} = \frac{S^-}{S^+} = \frac{\sum_{i=1}^{N_B} \sum_{j \notin G(i)} e^{-\gamma_{BD} \|\mathbf{z}_i - \mathbf{z}_j\|}}{\sum_{i=1}^{N_B} \sum_{j \in G(i)} e^{-\gamma_{BD} \|\mathbf{z}_i - \mathbf{z}_j\|}} \tag{1}$$

where $G(i)$ represents the group associated with the capture that contains the sample i (i.e. an instant of a cell within the capture). The parameter γ_{BD} plays a fundamental role in our framework and is inversely proportional to the temperature parameter τ in the original supervised contrastive loss [15], In particular, it allows us to control whether the aggregated distances for each anchor sample i rely on only a few samples (large value of γ_{BD}) or many (low value of γ_{BD}). This means that γ_{BD} sets the relative importance of the most similar elements in the population (i.e. those cells exhibiting behaviors similar to that of i) with respect to that of the entire population (including cells with similar and different behaviors). In our hierarchical scenario in which samples of different groups may exhibit similar behaviors and differences are expected at the population level, we are more inclined to use a small value of γ_{BD}.

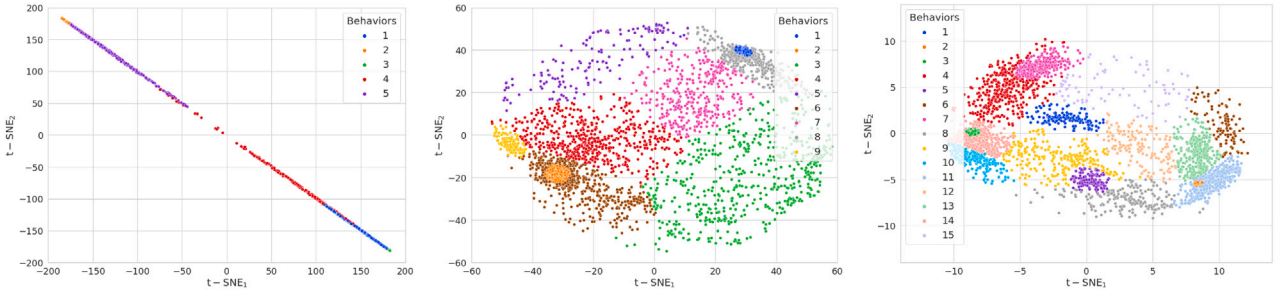
Furthermore, the value of γ_{BD} will have an impact on the separability of the learned embedded representations and the final number of discovered behaviors. In practice, we expect that large values of γ_{BD} will lead to data more sparsely distributed in the embedding space and result in a larger set of identifiable behaviors. Conversely, smaller values will organize data in a more compact way, generating a smaller set of behaviors. A qualitative analysis of the results under different values of γ_{BD} can be observed in Fig. 4. We will thoroughly analyze this point in the experimental section.

3.3.3. A bag-based contrastive loss for temporal consistency

Traditionally, to impose temporal coherence in recurrent embeddings, other authors have resorted to the prediction of the sample as a target task in the learning process [29], forcing the network to learn and predict the data dynamics. Instead of combining a supervised task (prediction) with an unsupervised one (behavior discovery), we rather propose to use the same sort of bag-based contrastive loss defined in Section 3.3.2, adapted to deal with temporal consistency. With our aggregation mechanism, which takes into account the different contributions to the aggregated similarity, the resulting embedding is robust to outliers (where outliers are the instant samples that are different to their temporal neighbors, for example, due to segmentation or tracking errors). From our point of view, this approach reduces the



(a) Labels identify groups: blue and 'o' markers for group 1 (wild-type) and yellow and 'x' markers for group 2 (anti-Plt).



(b) Colours identify behaviors. Note that as γ_{BD} increases, the embedding is sparser and the optimal number of behaviors also increases from $K_{opt} = 5$ (left), $K_{opt} = 9$ (center) and $K_{opt} = 15$ (right).

Fig. 4. 2-dimensional embeddings for $\gamma_{BD} = 0.01$ (left), $\gamma_{BD} = 0.6$ (center) and $\gamma_{BD} = 10$ (right).

influence of the outliers compared to a standard approach, where they would have a great impact on the learning process.

Therefore, we propose the use of a bag-based loss so that, within a bag of size N_B, we can enforce that the similarity between temporal instances of the same cell (taken in a temporal window of $T = 21$ instants), S_{intra}, is larger than the similarity between temporal instances of other cells and the cell under study, S_{inter}:

$$\mathcal{L}_{TC} = \frac{S_{inter}}{S_{intra}} = \frac{\sum_{i=1}^{N_B} \frac{1}{|\{C_{j|i}\}|} \sum_{j \in \{C_{j|i}\}} e^{-\gamma_{TC}\|\mathbf{z}_i - \mathbf{z}_j\|}}{\sum_{i=1}^{N_B} \frac{1}{|\{C_i\}|-1} \sum_{\substack{j \in \{C_i\} \\ j \neq i}} e^{-\gamma_{TC}\|\mathbf{z}_j - \mathbf{z}_i\|}} \tag{2}$$

where $\{C_i\}$ are the neighboring temporal instances of each cell i that are considered, and $\{C_{j|i}\}$ represents the set of the bag elements not belonging to the trajectory of the cell instance i. In our scenario, $|\{C_i\}| = C \ \forall i$, and, correspondingly, $|\{C_{j|i}\}| = N_B \cdot N_C - N_C$. The parameter γ_{TC} plays a slightly different role in ensuring temporal consistency than in behavior discovery. In this case, γ_{TC} can be considered as an indicator of the percentage of outliers in the dataset: focusing on the aggregated intra-cell similarity S_{intra}, small values give a similar importance to any time instance of a cell (appropriate for a scenario with low number of outliers), whereas large values make the loss focus on the most similar instances to the cell under consideration (suitable for scenarios with a high percentage of outliers). We will analyze its influence on the results in the experimental section.

This loss $\mathcal{L}_{TC}$ can also be interpreted as a regularization term for $\mathcal{L}_{BD}$, which is the primary loss.

3.4. Automatic behavior discovery

This section describes the process that, starting from either the original features or the embeddings, leads to the discovered behaviors. We have followed the next experimental protocol:

1. We have divided our dataset into training and test sets. The training set contains samples that belong to the control groups (wild-type and anti-Plt), whereas the test set contains samples from the therapy groups (FGR-KO and FGR-INH).
2. We have trained our Deep Sequence model using the training dataset, producing embeddings that separate wild-type and anti-Plt groups.
3. We have fitted a probabilistic clustering approach, a *Gaussian Mixture Model* (GMM) [45], over the training set after standard normalization. Each component in the mixture represents a data cluster, associated with a prominent behavior found in the control data. As the GMM requires to set the number of elements (behaviors), K, a priori, we chose K_{opt} as the value that maximizes the separability between the control groups (wild-type and anti-Plt) in terms of their distributions of behaviors. This is done as follows: first, for different values of $K \in [2, K_{max}]$ (i.e., for different numbers of potential clusters/behaviors), a normalized histogram $\mathbf{q}^K_{g_m} = \left[q^1_{g_m} \ q^2_{g_m} \ \dots \ q^K_{g_m}\right]$ is computed for each group $g_m = \{g_1, g_2\}$, measuring the proportion of cells $q^k_{g_m}$ belonging to each discovered behavior k. Then, the optimal number of behaviors K_{opt} is the one which minimizes the histogram intersection [46] between the two control groups:
$$K_{opt} = \underset{K \in [2, K_{max}]}{\arg\min} \ HI\left(\mathbf{q}^K_{g_1}, \mathbf{q}^K_{g_2}\right) \tag{3}$$
4. Once we have learned the Deep Sequence Model and the GMM parameters, we pass the test samples (FGR-KO and FGR-INH) through the model to generate their embeddings, and subsequently test how well the GMM fits the new data. To assess the quality of the fit, we will use several performance metrics described in the next section.

4. Results

Our approach has been implemented using Python and Pytorch. Z-score normalization by each feature is applied to the original data. We use a random train-validation split ratio of 90/10. The network is trained during 60 epochs using gradient descent with an Adam optimizer and takes around 12 h in an Intel(R) Core(TM) i7-7700 CPU equipped with a NVIDIA Geforce GTX 1080 Ti GPU. The inference process takes only a few minutes.

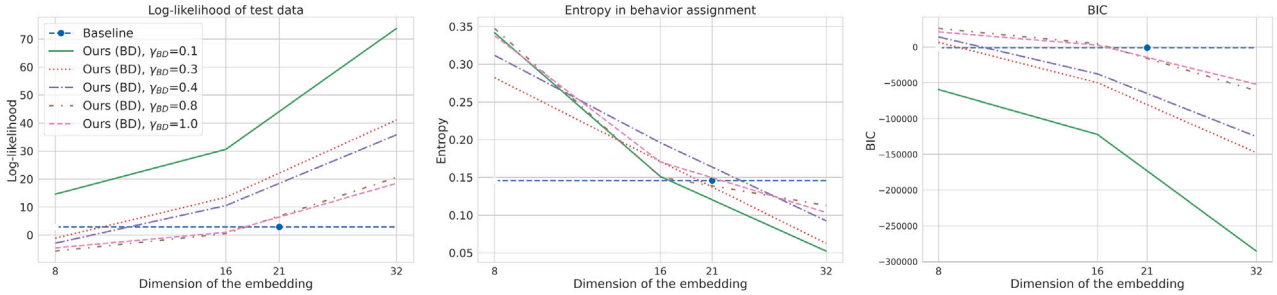


Fig. 5. Exploration of the influence of the γ_{BD} parameter. Test Log-likelihood (left), entropy in behavior assignment (center) and BIC (right).

During this Section, we will refer to the preliminary version of our approach using only the behavior discovery loss as Ours (BD), and our complete approach with the multi-task loss as Ours (BD+TC).

4.1. Performance metrics

In order to assess how our model trained on control groups fits the therapy groups (test data), we have used the following metrics:

- **Test Log-likelihood**. It evaluates how well the model trained in the control groups fits the new data (therapy groups):

$$\mathcal{L}_{test}\left(\theta \mid \mathbf{z}_i\right) = \sum_{i \in G(i)_{test}} \log\left(\sum_{k=1}^{K_{opt}} \alpha_k \phi\left(\mathbf{z}_i \mid \mu_k, \Sigma_k\right)\right) \tag{4}$$

  where $\theta = \left\{\alpha_1, \ldots, \alpha_{K_{opt}}, \mu_1, \ldots, \mu_{K_{opt}}, \Sigma_1, \ldots, \Sigma_{K_{opt}}\right\}$ are the parameters of the GMM (obtained with data from control groups) and ϕ is the Gaussian kernel. Higher values mean better fitting. In biological terms, it indicates if the behaviors discovered in the control groups are representative and fit well the cells in the therapy groups, i.e., if the behaviors are the same and shared between control and therapy groups.
- **Bayesian Information Criterion (BIC)**. It simultaneously considers model complexity and statistical fitting and evaluates how a model fits the new set of samples for a given complexity value. It considers the samples' likelihood along with the model's complexity (the number of learned parameters) and the total number of samples used to build the model [47]. Lower BIC values mean better model's performance and a less complex model.
- **Entropy in behavior assignment**. The GMM is a generative model that provides probabilistic assignments between samples and behaviors. We measure model's confidence in these assignments using the entropy of the assignments in the test set. Here, lower values mean better performance. From the biological point of view, it measures how well a model fitted in the control groups discriminates behaviors in the therapy groups i.e., how separable they are.

To carry out a fair comparison between the different alternatives and baselines, unless otherwise stated, we set the number of behaviors at $K_{opt} = 6$, which was determined as the optimal value using the original feature space.

4.2. Selection of hyperparameters

In the hyperparameter selection process for our dataset we consider a range of interest for the target embedding of $N_{emb} \in [12, 24]$ (surrounding the dimension of the original space, 21 features). We use 60-sample balanced bags ($N_B = 60$, with 30 samples per group), ensuring that each batch is statistically representative of the populations while keeping the computational complexity bounded. The number of temporal instants per cell varies across experiments. Finally, the dimension of the output representation of the bilateral LSTM is set to $N_{LSTM} = 256$, which matches the maximum N_{emb} value considered in our experiments, avoiding a bottleneck in the architecture). However, significant changes in this value, e.g. $N_{LSTM} = 128$, $N_{LSTM} = 512$, led to similar performance. Experiments in other datasets have yielded similar hyperparameters similar to the ones obtained in this paper. Furthermore, we propose these optimal search ranges for the hyperparameter validation process: $\gamma_{BD} \in [0.1, 1.0]$, γ_{TC} around 0.1 and $N_C \in [5, 11]$, and we provide the validation curves as Supplementary Material.

4.2.1. γ_{BD}: Tuning behavior discovery

First, we have focused on the behavior discovery loss, without the regularization term (temporal consistency) and using one sample per cell, i.e. $N_C = 1$. Fig. 4, which illustrates 2-dimensional embeddings ($N_{emb} = 2$) for three contrasting values of γ_{BD}, provides a deeper insight into the role of γ_{BD} in the sample distribution within the latent space. For $\gamma_{BD} = 0.1$, the embedding space shows a trivial solution, with samples perfectly separated in groups. While solution provides the best log-likelihood and BIC (as show on the left and right sides of Fig. 5), it is not suitable for our scenario as it challenges the assumption of shared behaviors across groups. Back to Fig. 4, for $\gamma_{BD} = 0.6$, the embedding space shows a more compact organization, where samples from each group are organized into a few big clusters (behaviors). In contrast, $\gamma_{BD} = 10$ yields a sparser data distribution, forming smaller clusters for each group. The characteristics of the data distribution in the embedding space have a direct impact on the number of cell behaviors (K_{opt} in our system). As γ_{BD} increases, the optimal number of behaviors also increases due to the sparser data distribution in the latent space. Specifically, for $\gamma_{BD} = 0.1$ the optimal number of clusters is $K_{opt} = 5$, for $\gamma_{BD} = 0.6$ it increases to $K_{opt} = 9$, and for $\gamma_{BD} = 10.0$ it further raises to $K_{opt} = 15$. It should be noticed that the experiment generating Fig. 4 is for visualization purposes, with a very reduced embedding dimensionality ($N_{emb} = 2$), and these values may change for different dimensionalities.

Fig. 5 shows the performance of our approach for different values of γ_{BD}, comparing it with the baseline working on the original feature space (blue point). The results reveal that our approach outperforms the baseline and allows us to select a value of $\gamma_{BD} = 0.8$ (the one with the best entropy in behavior assignment in the range of interest).

4.2.2. γ_{TC}: Tuning temporal consistency

Once the hyperparameter of the behavior discovery loss has been fixed, we can search for the optimal hyperparameter of the temporal consistency loss. First, we have validated γ_{TC} in an interval around the dimension of the original space. Furthermore, we have considered values for N_C in the range $[3, 13]$. Fig. 6 collects the performance measurements of the complete approach for different values of γ_{TC}, compared to the baseline. In this case, the value for $\gamma_{TC} = 0.01$ also yields a trivial solution, so we select as the optimal value $\gamma_{TC} = 0.1$, smaller than the optimal value of $\gamma_{BD} = 0.8$ for behavior discovery. This difference, as mentioned in Section 3.3.3, can be attributed to the

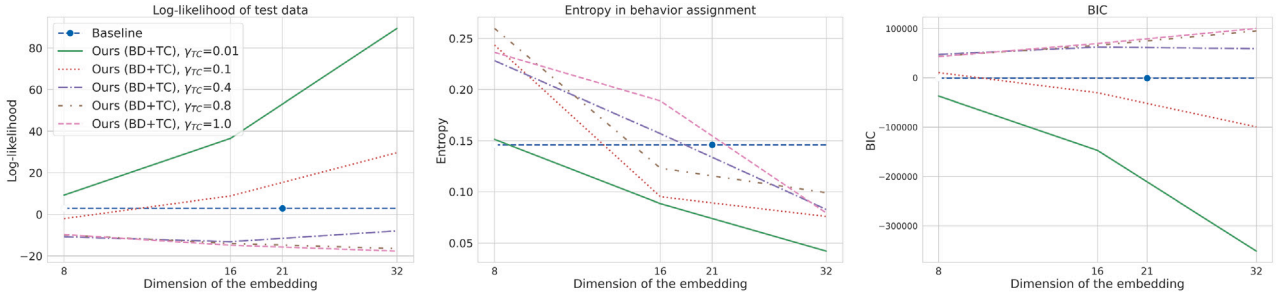


Fig. 6. Exploration of the influence of the γ_{TC} parameter. Test Log-likelihood (left), Entropy in behavior assignment (center) and BIC (right).

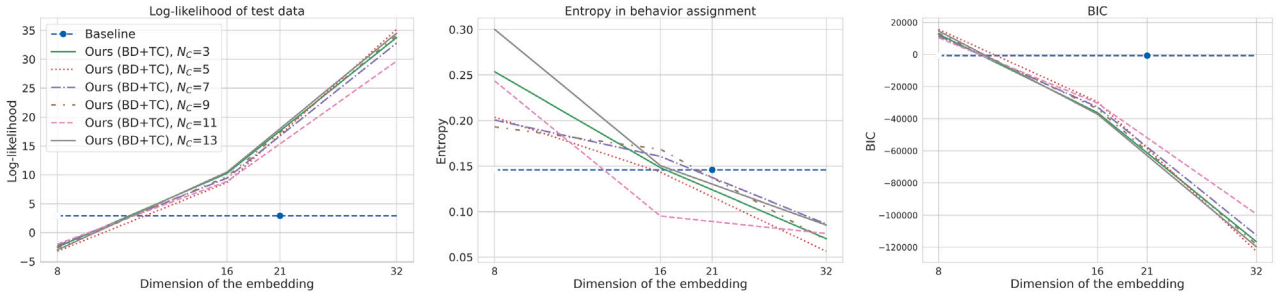


Fig. 7. Exploration of the influence of the N_C parameter. Test Log-likelihood (left), Entropy in behavior assignment (center) and BIC (right).

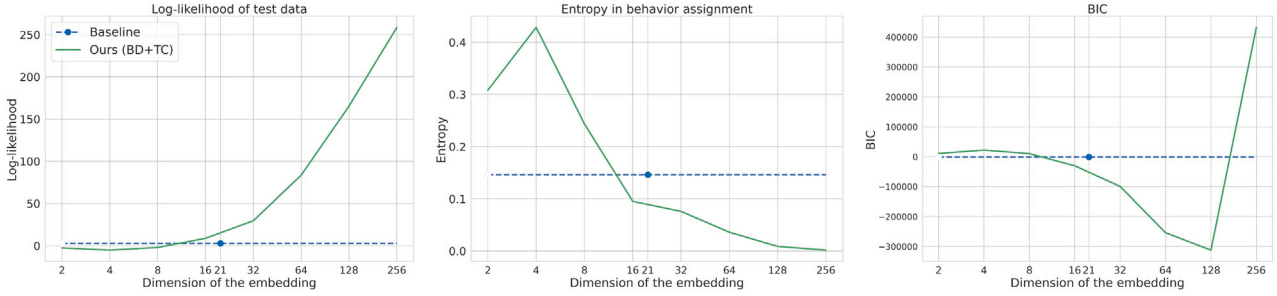


Fig. 8. Test Log-likelihood (left), Entropy in behavior assignment (center) and BIC (right) for $N_B = 60$, $\gamma_{BD} = 0.8$, $\gamma_{TC} = 0.1$, $N_C = 11$.

fact that the optimal hyperparameter is closely related to the number of outliers in the dataset. In our case, where the segmentation and tracking methods are robust, and outliers are few, lower values of γ_{TC} are favored, resulting in a more uniform weighting of the samples within.

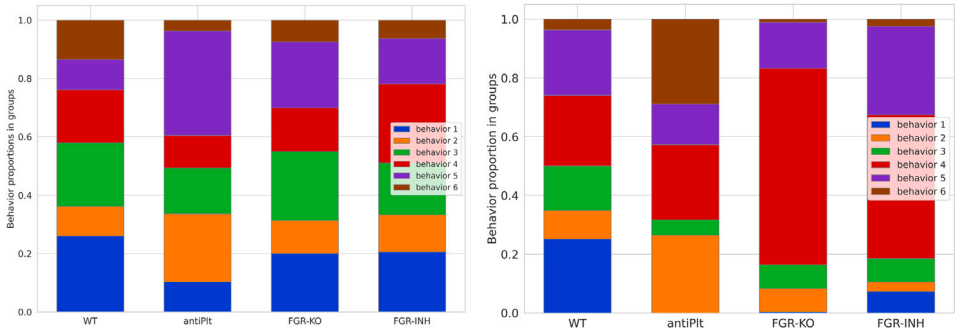
4.2.3. Number of temporal instances per bag

We have also analyzed the influence of N_C, the number of temporal instances included in the bag for ensuring temporal consistency. Fig. 7 illustrates the performance of our approach for different values of N_C in comparison with the baseline. The results show that the optimal performance is obtained when considering the $N_C = 11$ temporal instances of a cell. This configuration minimizes the entropy in behavior assignment, and results in negligible differences in test log-likelihood and BIC within the range of interest ($N_{emb} \in [12, 24]$). This observation reinforces our hypothesis in two senses: (1) the limited number of outliers justifies the use of a relatively large value of N_C for our purpose; and (2) since behavior transitions occur in our scenario, N_C values approaching the size of the considered temporal window ($N_C = 21$) are suboptimal. An increase in the time lapse (N_C) elevates the probability of a sample changing its behavior. Consequently, we cannot constrain the different time instances of a cell to remain within the same region of the latent space.

4.2.4. Size of the embedding

Once the values of $\gamma_{BD} = 0.8$, $\gamma_{TC} = 0.1$ and $N_C = 11$ have been selected through the previous experiments, Fig. 8 shows how our performance metrics evolve with the dimension of the embedding (N_{emb}). Several conclusions can be drawn from these results:

- First, the proposed embeddings outperforms the baseline, not only for the same dimensionality of the original data ($N_{emb} = 21$) but even for smaller sizes (e.g., $N_{emb} = 16$). Hence, our proposed system serves as a dual purpose: it re-arranges data in a latent space where behaviors can be better discriminated and accomplishes dimensionality reduction. Our intuition is that the initial set contains redundant features and, in some cases, irrelevant features, which are effectively removed in the transformed embedding space. We will present some examples of these features in Section 5.2.
- Second, the dimension of the embedding cannot be increased steadily. Whereas the test log-likelihood and entropy consistently improve with the dimension of the embedding, the BIC, which also considers the complexity of the system, worsens for very large sizes of the embedding (e.g. $N_{emb} = 256$). This indicates that the improvement in log-likelihood does not compensate for the increase in complexity.



(a) Histogram in the original space (instantaneous behaviors)

(b) Histogram in the proposed latent space (dynamic behaviors)

Fig. 9. Behavior proportion: instantaneous behaviors in the original feature space (left, with $K_{opt} = 6$) and dynamic behaviors in the proposed latent space (right, with $K_{opt} = 6$), for $N_B = 60$, $N_{emb} = 16$, $\gamma_{BD} = 0.8$, $\gamma_{TC} = 0.1$ and $N_C = 11$. *Please note that behaviors with the same numbering are not equivalent, and how the dynamic description of the scenario helps to strengthen the differences among the groups.

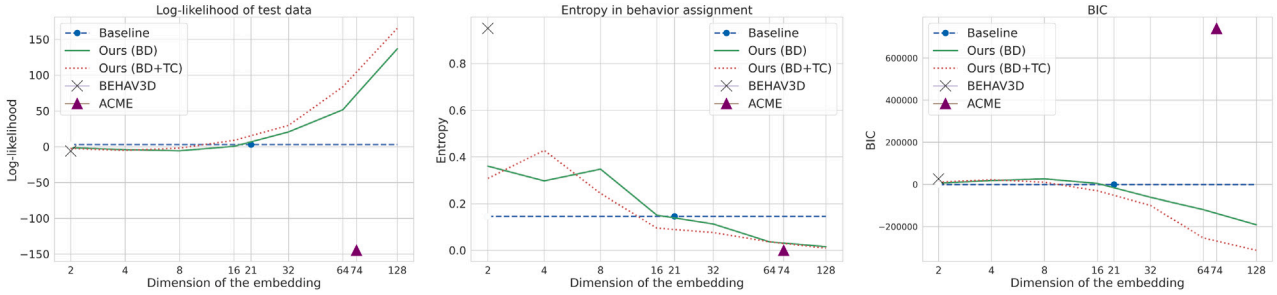


Fig. 10. Comparison of the performance measurements of the baseline and two state-of-the-art methods with two versions of our approach. Test Log-likelihood (left), Entropy in behavior assignment (center) and BIC (right).

Table 2
Number of detected cell behavior transitions, meaningful percentage and average duration of behaviors, for all the considered approaches.

Approach	# features	Transitions (%)	Meaningful transitions (%)	Average duration (time steps)
Baseline	21	36.00	48.12	7.27
ACME [41]	74	**7.03**	4.70	**13.06**
Ours (BD)	16	10.60	79.22	9.67
Ours (BD+TC)	16	**7.06**	**83.30**	**10.82**

Finally, in Fig. 9 the behavior proportions ($K_{opt} = 6$) are compared between the different groups in the original feature space and in the latent space with reduced dimensionality ($N_{emb} = 16$). As shown in the Figure, if cells' dynamics is not considered (left histogram), the behavior proportions in the group are more similar (differences among the behaviors lie in instantaneous cells' size, shape and positions, which can be stable across groups). However, when the cells' dynamics are considered, differences between groups are maximized (without breaking the requirement for shared behaviors between groups) and the resulting behaviors are biologically meaningful [3]. Moreover, our embedding holds the assumption that behaviors discovered in the control groups are also present in the therapy groups. However, in the embedding space, a non-shared behavior appears between the control groups (behaviors 1 and 2), which was not observed in the original feature space. The explainability procedure, described in Section 5.2, will provide some insight into the plausible biological hypotheses behind that.

4.3. Comparison with the state-of-the-art

Fig. 10 compares the two versions of our proposed approach with a baseline and two recent state-of-the-art method. Specifically:

- **Ours (BD)**: a preliminary version of our approach using only the behavior discovery loss.
- **Ours (BD+TC)**: our approach with the multi-task loss, which includes both behavior discovery and temporal consistency losses.
- **Baseline**: the baseline approach, describing each temporal instance of the cells through their original features (see Table 1).
- **ACME** [41]: a simple dynamic approach from the state of the art, where the instantaneous features are further processed to generate additional features aimed at modeling the temporal dynamics of the original ones, including basically their means, variances and some statistics over the trajectory of the cell. This results in a set of 74 instantaneous and dynamic features (for more information, refer to [41]).
- **BEHAV3D**: a very recent method in the literature [32]. The standard approach to perform sequence modeling over cell time series involves a dimensionality reduction algorithm prior to behavior identification [36,38]. In the case of BEHAV3D, a distance matrix between the cell temporal sequences $\mathbf{x}_i$ is computed using a dynamic time warping algorithm, which feeds UMAP to create 2D representations. Henceforth, BEHAV3D relies on a 2-dimensional representation of the data for behavior discovery. The behaviors, then, are discovered using clustering algorithm

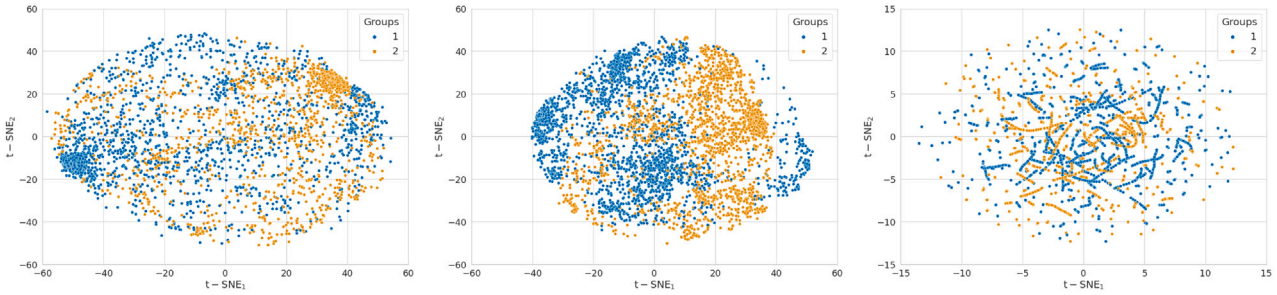


Fig. 11. 2-dimensional embeddings for our non-regularized approach (BD), our regularized version (BD+TC) and BEHAV3D [32]. Samples of group 1 are drawn with 'o' and samples of group 2 with 'x'.

(the authors suggest K-Means with outlier detection, but indicate that other methods could be used) over this 2D representation.

For our experiments, we have used the implementation of BEHAV3D available in [48]. To perform a fair comparison, we have used the same number of behaviors $K_{opt} = 6$ in the GMM algorithm for all the approaches. As observed in Fig. 10, our regularized approach outperforms the baseline approach, ACME, and BEHAV3D, in terms of BIC, entropy in behavior assignment and test log-likelihood. Hence, our sequence modeling approach with temporal consistency demonstrates superior performance in terms of separability and dimensionality reduction when compared to more standard approaches, including: (1) simple dynamic modeling through statistics extracted from the original features, such as the ACME approach; and (2) dimensionality reduction (2D-3D) and clustering, such as BEHAV3D.

Finally, Fig. 11 illustrates the 2-dimensional embeddings for our proposed approach, the regularized one (with temporal consistency) and BEHAV3D. As observed, samples are better organized in clusters by the proposed approach. It is also worth noting that in the case of the BEHAV3D representation, samples from the same cell are always in the same region of the feature space (aligned), not allowing behavior transitions.

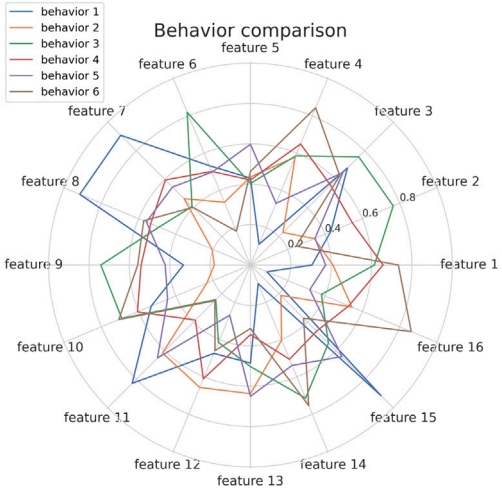


Fig. 12. Explainability of the discovered behaviors: medium values of the latent features for each behavior.

5. Discussion

In order to provide more insight into the capabilities and limitations of the proposed system, we first analyze the stability of behavior transitions and provide explainability to the results in Sections 5.1 and 5.2, respectively. Lastly, we examine the coherence of behavior transitions in Section 5.3.

5.1. Behavior transitions

This subsection is devoted to analyze the behavior transitions in our approach in comparison with: (1) the baseline approach, working with the initial set of 21 instantaneous features, and (2) the ACME approach, working with 74 instantaneous and dynamic features. Note that BEHAV3D is not included in the comparison, due to the fact that it does not allow behavior transitions explicitly (it analyzes the complete sequence of temporal features and transform it in a single 2D point representing its behavior).

Table 2 shows the number of detected behavior transitions for each of the described alternatives in comparison with our behavior discovery (BD) approach, with and without temporal consistency (TC). Here, the number of transitions are expressed as a percentage, indicating the proportion of cell behavior changes relative to the total number of time steps. According to the biologists' hypotheses, transitions are possible in our scenario, but they are considered *meaningful* only if cells remain in the new behavior for at least three temporal instants [3]. Some conclusions can be drawn from these results:

- First, the instantaneous modeling detects a share of 36% of behavior transitions. If we do not consider the temporal evolution of the features, the obtained behaviors are not descriptive enough of the conduct of the cells, and the resulting model lacks temporal consistency.
- Second, the dynamic modeling detects the smallest number of transitions (7%), but most of them are not meaningful. Even when we are considering a 21 time-step window to compute them and the features from adjacent time instances of the same cell should be very similar, the changes in cell behaviors are not temporally consistent.
- Our approach reduces considerably the number of transitions compared to the instantaneous modeling, and most of the transitions are meaningful. Even only considering the behavior discovery loss, the percentage of meaningful transitions reaches 79%.
- When considering temporal consistency as a regularizer for our behavior discovery approach, the number of transitions is reduced to 7%, approximately, and more than 83% are meaningful. The error analysis in Section 5.3 will examine the nature of these remaining non-meaningful behavior transitions.
- Finally, the mean duration of the behavior assignment is very similar with our approach (11 time steps) and the dynamic modeling (13 time steps).

As a conclusion, our approach provides meaningful and stable transitions.

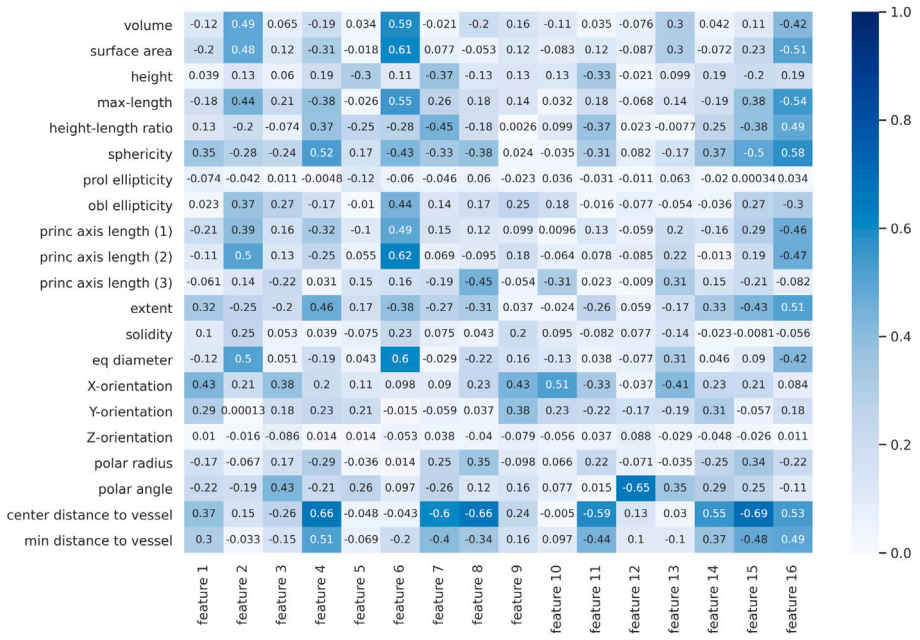


Fig. 13. Explainability of the discovered behaviors: correlation between the hand-crafted features and the latent features (color indicates absolute value of correlation).

Table 3
Textual description of the discovered behaviors.

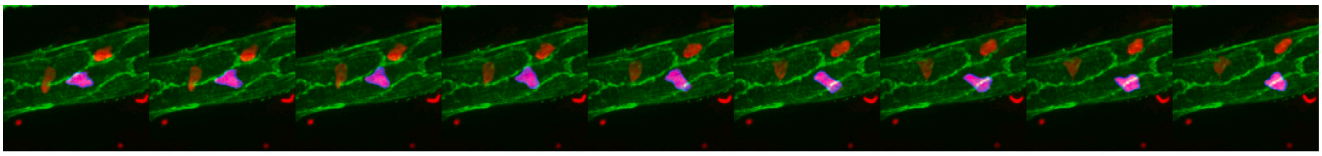
Behavior	Relevant latent features	Description
1	4^-, 14^-, 16^-, 7^+, 8^+, 15^+	This behavior contains big neutrophils, slightly spherical and they move erratically close to the blood vessel wall. Hence, they are polarized and this behavior is a pathogenic migration.
2	12^+, 8^-, 9^-, 10^-, 15^-	The neutrophils that present this behavior are small, spherical and stay practically still far from the blood vessel wall. Hence, it is a non-migratory behavior.
3	2^+, 3^+, 6^+, 9^+	Cells in this behavior are big, have an erratic movement and their positions are distant from the blood vessel wall. This behavior is very similar to the first behavior but less migratory, composed of cells starting to migrate pathogenically.
4	12^+, 13^-	The fourth behavior is composed of neutrophils of medium size, moving with linear trajectories and high velocity along the blood flow. Thus, it is a non-pathogenic migration.
5	5^+, 13^+, 12^-	Neutrophils in this behavior are of medium size, close to the blood vessel wall but moving along the blood flow with low velocity. Hence, this behavior is a non-pathogenic migration.
6	1^+, 4^+, 16^+, 6^-	Cells in this behavior are of medium size, spherical and their positions are distant from the blood vessel wall. Therefore, this is a non-migratory behavior.

5.2. Explainability and agreement with biological hypotheses

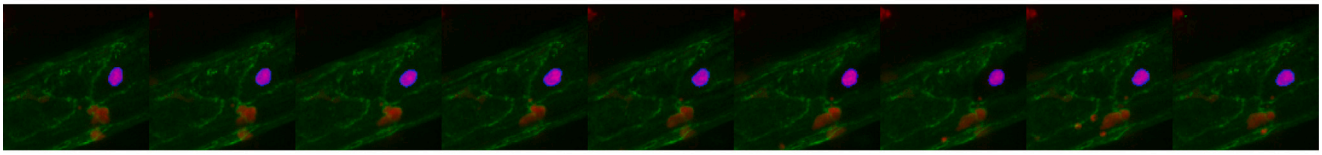
Explainability is a fundamental requisite in cell behavior analysis: it is necessary to understand the mechanisms behind cell migration to obtain more effective therapies. Although our system is based on self-supervised learning, explainability is preserved through the proportion of behaviors in each group, the specific values of each latent feature for each behavior, and the correlation between the latent features and the original ones.

Let us focus now on Figs. 12, 13 and 9(a), all of them obtained using our complete approach with the selected hyperparameters ($\gamma_{BD} = 0.8$, $\gamma_{TC} = 0.1$, $N_C = 11$ and $N_{emb} = 16$). Fig. 9(a) shows the histogram of discovered behaviors for each group, Fig. 12 shows a spider graph with the mean values of the features for each behavior and Fig. 13 presents the correlation matrix between the original features and the latent features. The combination of the last two allows us to obtain textual descriptions that are meaningful for the biologists. These textual descriptions are gathered in Table 3. The analysis of this information enables us to draw several conclusions of interest:

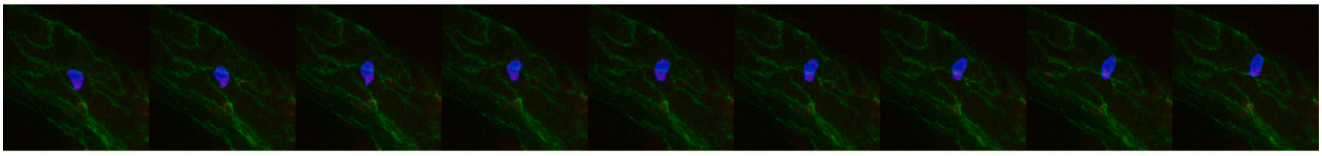
1. The histogram of behaviors demonstrates that our method successfully fulfills our hypothesis of shared behaviors across groups, even for those therapy groups that remained unseen during the learning of embeddings and the clustering steps. This graph is crucial for studying the effect of the therapies, by comparing their proportions with those of control groups (WT and anti-Plt).



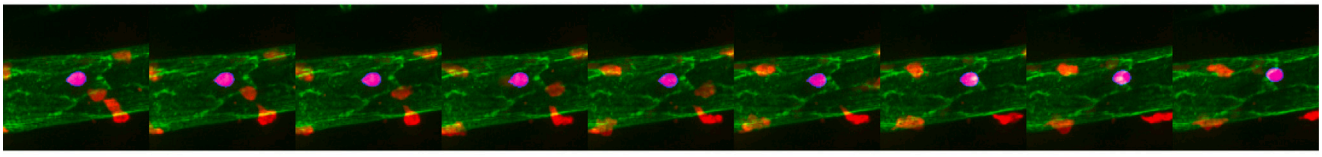
(a) Behavior 1. Pathogenic migration. Big cells with significant shape changes.



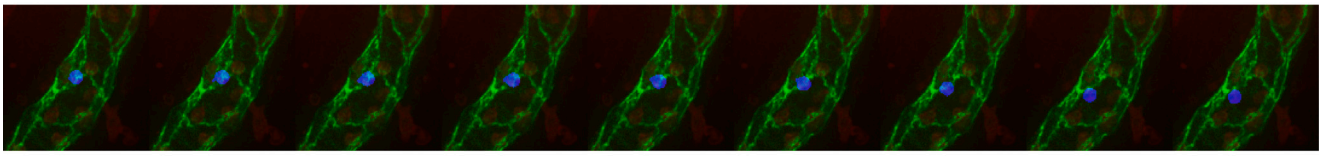
(b) Behavior 2. Non-migratory, smaller cells.



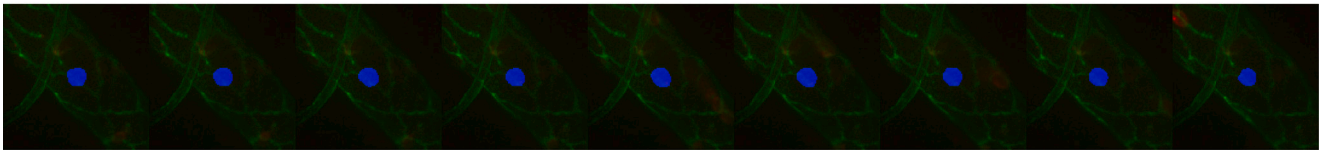
(c) Behavior 3. Cells starting to migrate, in a pathogenic manner going towards the blood vessel wall.



(d) Behavior 4. Non-pathogenic migration, notable displacement with the blood flow.



(e) Behavior 5. Non-pathogenic migration, low displacement.



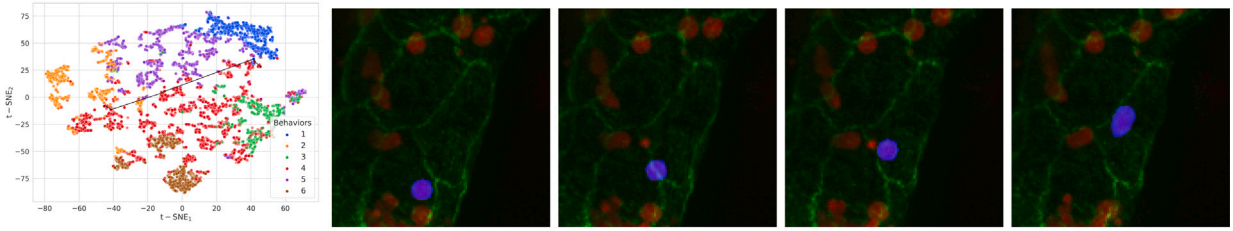
(f) Behavior 6. Non-migratory, bigger cells.

Fig. 14. Illustrative cases of discovered behaviors (Z-stack) in consecutive time instants (from left to right). Cells marked red, blood vessel in green and cell of interest in blue or magenta, depending on the red level of the capture. As an example, observe the different dynamics between a pathogenic migration (behavior 1) and a non-migratory behavior (behavior 2). Changes in shape and motion of the cells are quite noticeable in this case.

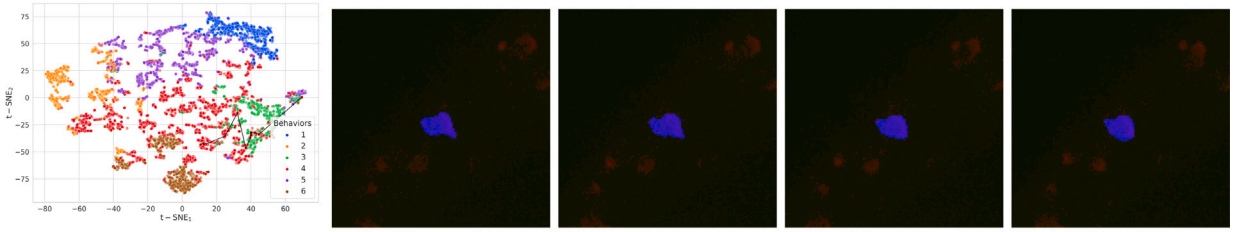
2. The spider graph allows biologists to quickly compare behaviors and identify which dimensions of the latent embedding are more relevant to identify a cell behavior.
3. Furthermore, by combining the previous information with the correlation between the latent embedding and the original features, we can obtain a textual description of each behavior, expressed in terms that are meaningful for the biologists. In particular, Table 3 enumerates the most relevant features for each behavior, with an indicator of the mean value of the features: "+" for a high value and "−" for a low value (second column). Additionally, the table includes a description for each behavior derived from the relevant features and the correlation (third column). Indeed, this allows to identify the features from the original set that are redundant (contributing similarly to every embedding feature) or irrelevant (having low correlation with all the embedding features).

The obtained conclusions are totally consistent with the biologists's hypotheses [3]. All the groups can be modeled using the same set of behaviors, but in different proportions. The wild-type group contains more neutrophils exhibiting migratory behaviors, 1 and 3, and the anti-Plt group has more cells of behaviors 2 and 6, which are non-migratory. The therapy groups show intermediate results in comparison with the control groups, which is what is intended with the therapies. Apart from that, the most migratory behavior (the first one) is only present in the wild-type group and FGR-INH group, which can be seen as an indicator of the effectiveness of the therapies (as this behavior is the most harmful). Finally, it is worth noting that, in our scenario, cell size and shape are correlated with migration (platelet recruitment for neutrophil migration implies a grow in the volume of cells [49]). Hence, differences in these features can constitute a behavior transition (changes in cell phenotypes in this scenario imply changes in their dynamics).

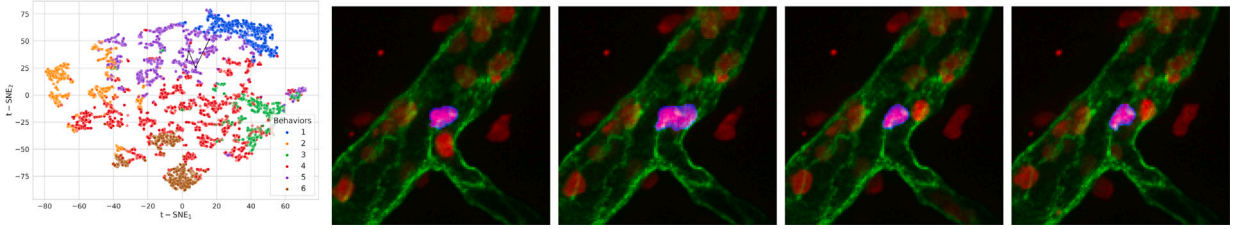
Regarding the identification of irrelevant or redundant features, the correlation between the original features and embedding features shown in Fig. 13 reveals that features like *Z-orientation* or *solidity* are practically irrelevant. In addition, as observed, all the features closely related to the cell size (*volume*, *surface area*, *equivalent diameter*, etc.) show similar contributions to the embedding features; thus, they are



(a) Significant change of position in one time instant.



(b) Shape changes due to migration.



(c) Unusual segmentation and tracking error, cell collision not separated.

Fig. 15. Illustrative examples of the most abrupt behavior transitions with a t-SNE plot of their trajectory (left) and consecutive time instants (Z-stacks, right). Cells marked in red, blood vessel in green and cell of interest in blue or magenta, depending on the red level of the capture. The transition occurs between the central time instants. Note that the disposition in the t-SNE is coherent: the fastest migration (blue, behavior 1) is top-right and the non-migratory behaviors are left (behavior 2) and bottom (behavior 6).

redundant features, and some of them could be removed from the analysis.

Finally, Fig. 14 shows one illustrative example of each discovered behavior in our scenario. The differences between the pathogenic migrations (involving shape changes and erratic trajectories along the blood vessel wall) and the non-migratory ones (involving practically immobile spherical cells) are easily noticeable. The non-pathogenic migrations show an intermediate conduct.

5.3. Error analysis and discussion

In this subsection, we aim to gain more insight into our latent space representation by analyzing cases that break the temporal consistency and may be attributed to errors in some steps of the processing pipeline. First, we will identify the most abrupt changes in the latent representations of consecutive temporal instants to evaluate if they are caused by segmentation or tracking errors. Second, we will discuss some non-meaningful behavior transitions, in which the cell remains less than three temporal instants in the destination behavior.

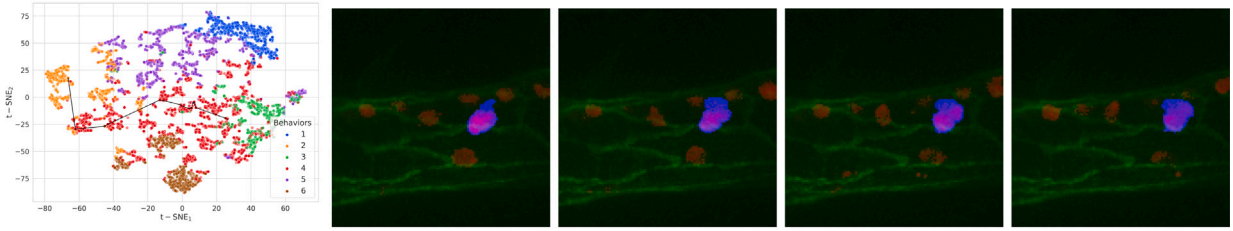
Fig. 15 shows the most abrupt displacements in our final embedding through the t-SNE trajectories and the Z-stacks of the cells in the time instants when the transition occurs. As observed, the abrupt transitions are related to: (a) abrupt changes of position in one time instant due to a notable acceleration of the migrating cell; (b) significant changes in cell shape due to migration dynamics, and (c) some segmentation and tracking errors, particularly, unusual difficulties of the collision management system to split the agglomerations of neutrophils. However, we found that, in general, the segmentation and tracking errors neither lead to abrupt transitions nor produce changes in behavior assignment. This means that they are not a significant proportion of the database, something expected as the segmentation and tracking systems include some filtering steps that remove low-confidence cells and provide a final 95% precision [41].

Fig. 16 shows some examples of the non-meaningful behavior transitions in our database. In these cases, some of the behavior transitions are related to exceptional changes in cell behaviors due to migration dynamics or slight segmentation inaccuracies in cell boundaries. Additionally, some transitions are linked to more challenging errors in cell segmentation and tracking. Hence, our latent space can be used as an outlier detection module.

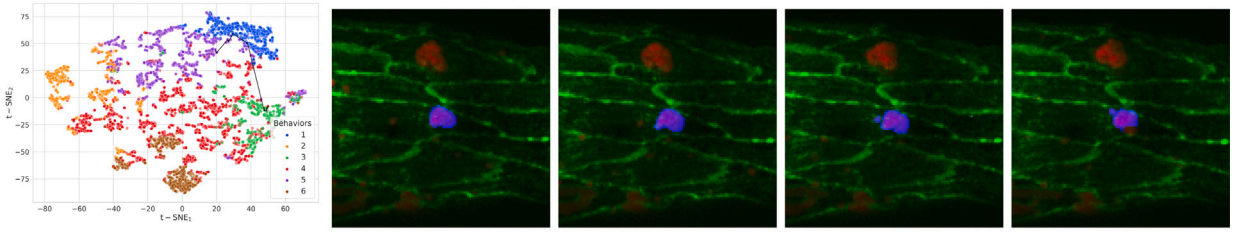
6. Conclusions

In this paper, we have proposed a self-supervised method that, starting from a set of hand-crafted cell migration features, achieves three main goals: (1) it creates embedded spaces where undesired effects, such as feature redundancy or irrelevance, are mitigated, and cell behaviors are better discriminated, facilitating their discovery; (2) it provides temporal consistency to the cell dynamics representation, while still allowing for behavior transitions; and (3) it aligns with our hierarchical scenario in which cells adopt behaviors that are shared across different groups, although in different proportions. To accomplish these objectives, we have proposed a multi-task bag-based contrastive loss that not only aims to separate populations of cells in the embedded space but also enforces temporal consistency in the obtained representations.

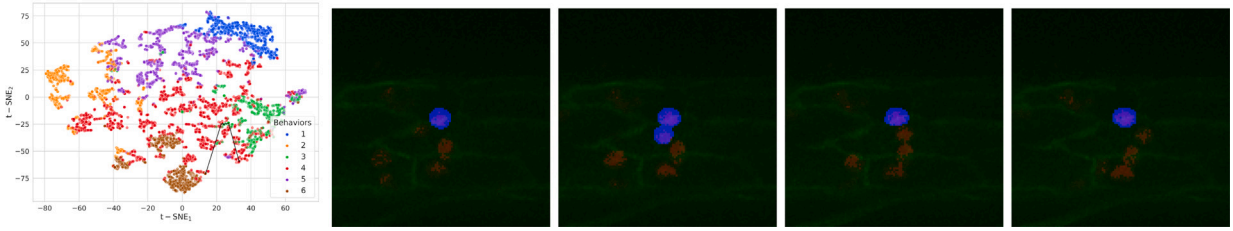
Our experiments have shown that our embeddings provide notable improvements with respect to the original features, even with reduced dimensionality. Furthermore, the proposed loss function is parametric: including a parameter γ_{BD} to adapt the behavior discovery to scenarios where biologists have intuitions about the expected number of behaviors; and a parameter γ_{TC}, which allows for adaptation to different levels of outliers in the dataset.



(a) A challenging error in segmentation and tracking, touching cells not separated.



(b) Small inaccuracies in cell borders.



(c) Unusual segmentation and tracking error, cell collision not separated.

Fig. 16. Illustrative examples of the remaining non-meaningful behavior transitions with a t-SNE plot of their trajectory (left) and consecutive timestamps (Z-stacks, right). The transition occurs between the central time instants (third and four column). Cells marked in red, blood vessel in green and cell of interest in blue or magenta, depending on the red level of the capture.

Moreover, these improvements do not compromise interpretability. By analyzing relationships between the original feature set and our embeddings, we have shown that our results are explainable, offering textual descriptions of the discovered behaviors in meaningful terms to biologists.

Regarding the robustness of the system, the errors in the segmentation and tracking step did not have a critical influence on our latent embedding. They neither lead to changes in the behavior assignment nor cause of the most abrupt behavior transitions in the embedding. Furthermore, the non-meaningful transitions (around 15%) could be removed from the analysis.

The main limitation of the proposed algorithm relates to its generalization ability: new therapies can generate completely new cell behaviors that cannot be assigned to those existing in the control groups. They may be identified in the embedding, but they cannot be explained. The recent field of open vocabulary learning [50] provides a solution: by providing the original handcrafted features as auxiliary supervision in training, it could identify new behaviors during inference and relate them to these interpretable feature. Other directions for further research involve exploring the use of autoencoders and other types of generative models to ensure that the latent representation contains all the information about the original feature space, and the inclusion of clustering constraints in the same network to improve behavior discovery and separability.

CRediT authorship contribution statement

Miguel Molina-Moreno: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Iván González-Díaz:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Ralf Mikut:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition. **Fernando Díaz-de-María:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by the National Grant PID2020-118504GB-I00 of the Spanish Ministry of Science and Innovation (State Research Agency) and Madrid Regional Government Grant Y2020/NMT-6660 (synergistic R&D interdisciplinary project COMPANION-CM). Miguel Molina-Moreno is supported by the Spanish Ministry of Education, Culture and Sports FPU Grant FPU18/02825. Ralf Mikut is supported by the Helmholtz Association in the programs Natural, Artificial and Cognitive Information Processing. Funding for APC: Universidad Carlos III de Madrid (Agreement CRUE-Madroño 2024).

The authors would like to kindly thank to Cardiovascular Inflammation Imaging and the Immune Response Laboratory (LIICRI) from CNIC for supplying the 4D volumes used for this research and providing biological support to this research. In particular, the authors would like to acknowledge Georgiana Crainiciuc, Miguel Palomino, Jon Sicilia and Andrés Hidalgo. The authors would like to also thank the members of the ML4TIME and ML4HOME groups at KIT for their meaningful insights into unsupervised learning, in particular: Marcel Schilling, Simon Bäuerle, Oliver Neumann and Luca Rettenberger.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cmpb.2024.108337.

References

[1] P. Haftbaradaran Esfahani, R. Knöll, Cell shape: effects on gene expression and signaling, Biophys. Rev. 12 (4) (2020) 895–901, http://dx.doi.org/10.1007/s12551-020-00722-4.

[2] M. Mukhopadhyay, Unraveling immune cell behavior, Nat. Methods (2022) http://dx.doi.org/10.1038/s41592-022-01435-0.

[3] G. Crainiciuc, M. Palomino-Segura, M. Molina-Moreno, et al., Behavioral immune landscapes of inflammation, Nature 601 (7893) (2021) 415–421, http://dx.doi.org/10.1038/s41586-021-04263-y.

[4] M. Di Pilato, R. Kfuri-Rubens, J.N. Pruessmann, et al., CXCR6 positions cytotoxic T-cells to receive critical survival signals in the tumor microenvironment, Cell 184 (17) (2021) 4512–4530.e22, http://dx.doi.org/10.1016/j.cell.2021.07.015.

[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (null) (2003) 993–1022.

[6] S. Gidaris, A. Bursuc, G. Puy, et al., OBow: Online bag-of-visual-words generation for self-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 6830–6840.

[7] J. Schindelin, I. Arganda-Carreras, E. Frise, et al., Fiji: an open-source platform for biological-image analysis, Nat. Methods 9 (2012) 676–682, http://dx.doi.org/10.1038/nmeth.2019.

[8] L. van der Maaten, G. Hinton, Visualizing high-dimensional data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[9] L. McInnes, J. Healy, N. Saul, et al., UMAP: Uniform manifold approximation and projection, J. Open Source Softw. 3 (29) (2018) 861.

[10] M.S. Cooleym, T. Hamilton, S.D. Aragones, et al., A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-seq data, 2022, http://dx.doi.org/10.1101/689851, bioRxiv.

[11] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828, http://dx.doi.org/10.1109/TPAMI.2013.50.

[12] Y. An, G. Cai, X. Chen, L. Guo, PARSE: A personalized clinical time-series representation learning framework via abnormal offsets analysis, Comput. Methods Programs Biomed. 242 (2023) 107838, http://dx.doi.org/10.1016/j.cmpb.2023.107838.

[13] J. Ke, Y. Shen, Y. Lu, Y. Guo, D. Shen, Mine local homogeneous representation by interaction information clustering with unsupervised learning in histopathology images, Comput. Methods Programs Biomed. 235 (2023) 107520, http://dx.doi.org/10.1016/j.cmpb.2023.107520.

[14] R. Wang, Q. Zhou, G. Zheng, EDRL: Entropy-guided disentangled representation learning for unsupervised domain adaptation in semantic segmentation, Comput. Methods Programs Biomed. 240 (2023) 107729, http://dx.doi.org/jcmpb.2023.107729.

[15] P. Khosla, P. Teterwak, C. Wang, et al., Supervised contrastive learning, in: Adv. Neural Inf. Processing Syst., 33, 2020, pp. 18661–18673.

[16] T. Martínez-Cortés, I. González-Díaz, F. Díaz-de-María, Training deep retrieval models with noisy datasets: Bag exponential loss, Pattern Recognit. 112 (2021) 107811, http://dx.doi.org/10.1016/j.patcog.2020.107811.

[17] G. Uribarri, G.B. Mindlin, Dynamical time series embeddings in recurrent neural networks, Chaos Solitons Fractals 154 (2022) 111612, http://dx.doi.org/10.1016/j.chaos.2021.111612.

[18] A. Hamdi, K. Shaban, A. Erradi, et al., Spatiotemporal data mining: A survey on challenges and open problems, Artif. Intell. Rev. 55 (2022) 1441–1448, http://dx.doi.org/10.1007/s10462-021-09994-y.

[19] S. Wang, J. Cao, P. Yu, Deep learning for spatio-temporal data mining: A survey, IEEE Trans. Knowl. Data Eng. (01) (2020) 1, http://dx.doi.org/10.1109/TKDE.2020.3025580.

[20] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[21] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Adv. Neural Inf. Processing Syst. 30 (2017).

[22] P. Ramachandran, P. Liu, Q. Le, Unsupervised pretraining for sequence to sequence learning, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 383–391, http://dx.doi.org/10.18653/v1/D17-1039.

[23] Y. Liu, J. Chen, L. Deng, Unsupervised sequence classification using sequential output statistics, in: Advances in Neural Information Processing Systems, 2017, pp. 3550–3559.

[24] J.R. Chang, A. Shrivastava, H. Koppula, X. Zhang, O. Tuzel, Style equalization: Unsupervised learning of controllable generative sequence models, in: International Conference on Machine Learning, Vol. 162, ICML 2022, PMLR, 2022, pp. 2917–2937.

[25] Z. Chen, J. Droppo, Sequence modeling in unsupervised single-channel overlapped speech recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2018, pp. 4809–4813, http://dx.doi.org/10.1109/ICASSP.2018.8461939.

[26] G.S. Han, Q. Li, Y. Li, Nucleosome positioning based on DNA sequence embedding and deep learning, BMC Genom. 23 (301) (2022) http://dx.doi.org/10.1186/s12864-022-08508-6.

[27] C. Liu, V. Ta, N. Le, et al., Deep neural network framework based on word embedding for protein glutarylation sites prediction, Life 12 (8) (2022) http://dx.doi.org/10.3390/life12081213.

[28] Z. Shen, W. Bao, D.S. Huang, Recurrent neural network for predicting transcription factor binding sites, Sci. Rep. 8 (2018) 15270, http://dx.doi.org/10.1038/s41598-018-33321-1.

[29] S.T. Hill, R. Kuintzle, A. Teegarden, et al., A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential, Nucleic Acids Res. 46 (16) (2018) 8105–8113, http://dx.doi.org/10.1093/nar/gky567.

[30] W. Kou, D.A. Carlson, A.J. Baumann, E. Donnan, Y. Luo, J.E. Pandolfino, M. Etemadi, A deep-learning-based unsupervised model on esophageal manometry using variational autoencoder, Artif. Intell. Med. 112 (2021) 102006, http://dx.doi.org/10.1016/j.artmed.2020.102006.

[31] Z. Wu, B.B. Chhun, G. Popova, et al., DynaMorph: self-supervised learning of morphodynamic states of live cells, Molecular Biol. Cell 33 (6) (2022) ar59, http://dx.doi.org/10.1091/mbc.E21-11-0561.

[32] J.F. Deckers, M. Alieva, A. Clevens, et al., BEHAV3D: an imaging and transcriptomics platform that unravels T-cell antitumor activity, Nat. Biotechnol. 35 (2022) http://dx.doi.org/10.1038/s41587-022-01398-9.

[33] R. Venu, S.T. Nabi, M. Kumar, Self-supervised learning: A succinct review, Arch. Comput. Methods Eng. 30 (2023) 2761–2765, http://dx.doi.org/10.1007/s11831-023-09884-2.

[34] C. Doersch, A. Gupta, A.A. Efros, Unsupervised visual representation learning by context prediction, in: 2015 IEEE International Conference on Computer Vision, ICCV, IEEE Computer Society, 2015, pp. 1422–1430, http://dx.doi.org/10.1109/ICCV.2015.167.

[35] I. Misra, C.L. Zitnick, M. Hebert, Unsupervised learning using sequential verification for action recognition, in: Computer Vision – ECCV 2016, Springer International Publishing, 2016, pp. 527–544.

[36] P. Gundogdu, C. Loucera, I. Alamo-Alvarez, et al., Integrating pathway knowledge with deep neural networks to reduce the dimensionality in single-cell RNA-seq data, BioData Mining 7 (2022) http://dx.doi.org/10.1186/s13040-021-00285-4.

[37] S. Xiaobo, S. Xiaochu, L. Ziyi, et al., A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq, Brief. Bioinf. 23 (2022) http://dx.doi.org/10.1093/bib/bbab567.

[38] A. Rives, J. Meier, T. Sercu, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. 118 (15) (2021) e2016239118.

[39] I. Jolliffe, Principal Component Analysis, Springer Verlag, 1986.

[40] A. Zaritsky, A.R. Jamieson, E.S. Welf, et al., Interpretable deep learning uncovers cellular properties in label-free live cell images that are predictive of highly metastatic melanoma, Cell Syst. 12 (7) (2021) 733–747.e6, http://dx.doi.org/10.1016/j.cels.2021.05.003.

[41] M. Molina-Moreno, I. González-Díaz, F. Díaz-de María, et al., ACME: Automatic feature extraction for cell migration examination through intravital microscopy imaging, Med. Image Anal. 77 (2022) 102358, http://dx.doi.org/10.1016/j.media.2022.102358, URL https://www.sciencedirect.com/science/article/pii/S1361841522000111.

[42] M. Molina-Moreno, Miguel55/ACME: ACME-v1.0, 2021, http://dx.doi.org/10.5281/zenodo.5638537.

[43] M.R. Hasan, N. Hassan, R. Khan, Y.-T. Kim, S.M. Iqbal, Classification of cancer cells using computational analysis of dynamic morphology, Comput. Methods Programs Biomed. 156 (2018) 105–112, http://dx.doi.org/10.1016/j.cmpb.2017.12.003.

[44] J. Opila, G. Krzysiek-Maczka, Direct tool for quantitative analysis of cell/object dynamic behavior – metastasis and far beyond, Comput. Methods Programs Biomed. 229 (2023) 107245, http://dx.doi.org/10.1016/j.cmpb.2022.107245.

[45] D.A. Reynolds, Gaussian Mixture Models, Springer US, 2009, pp. 659–663, http://dx.doi.org/10.1007/978-1-4899-7488-4_196.

[46] M.J. Swain, D.H. Ballard, Color indexing, Int. J. Comput. Vis. 7 (1991) 11–32, http://dx.doi.org/10.1007/BF00130487.

[47] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.

[48] M. Alieva, BEHAV3D: an imaging and transcriptomics platform that unravels T-cell antitumor activity, 2022, https://github.com/alievakrash/BEHAV3D.

[49] S. Pitchford, D. Pan, H.C. Welch, Platelets in neutrophil recruitment to sites of inflammation, Curr. Opin. Hematol. 24 (1) (2017) 23–31, http://dx.doi.org/10.1097/MOH.0000000000000297.

[50] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang, B. Ghanem, D. Tao, Towards open vocabulary learning: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2024) 1–20, http://dx.doi.org/10.1109/TPAMI.2024.3361862.