BunDLe-Net: Neuronal Manifold Learning Meets Behaviour

Akshey Kumar[1,4], Aditya Gilra[2,3], Mauricio Gonzalez-Soto[1], Anja Meunier[1,4], and Moritz Grosse-Wentrup[1,5,6]

[1] *Research Group Neuroinformatics, Faculty of Computer Science, University of Vienna, Austria*
[2] *Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam, Netherlands*
[3] *Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom*
[4] *UniVie Doctoral School Computer Science DoCS, University of Vienna, Austria*
[5] *Vienna Cognitive Science Hub*
[6] *Data Science @ Uni Vienna*

April 15, 2024

Abstract

Neuronal manifold learning techniques represent high-dimensional neuronal dynamics in low-dimensional embeddings to reveal the intrinsic structure of neuronal manifolds. A common goal of these techniques is to learn embeddings that allow a good reconstruction of the original data. We introduce a novel neuronal manifold learning technique, BunDLe-Net, that learns a low-dimensional Markovian embedding of the neuronal dynamics which preserves only those aspects of the neuronal dynamics that are relevant for a given behavioural context. In this way, BunDLe-Net eliminates neuronal dynamics that are irrelevant for decoding behaviour, effectively de-noising the data to reveal better the intricate relationships between neuronal dynamics and behaviour. We show that BunDLe-Net learns highly consistent manifolds across animals that reveal the building blocks of their neuronal manifolds on a variety of data sets, ranging from calcium imaging data recorded in the nematode *C. elegans* to spiking data from the rat hippocampus and primate somatosensory cortex.

1 Introduction

Advances in neuronal imaging techniques have increased the number of neurons that can be recorded simultaneously by several orders of magnitude [1, 2]. While these advances greatly expand our abilities to study and understand brain function, the complexities of the resulting high-dimensional data sets pose non-trivial challenges for data analysis and visualisation. Fortunately, individual neurons are embedded in brain networks that collectively organise their high-dimensional neuronal activity patterns into lower-dimensional neuronal manifolds [3, 4]. To understand the collective organisation of individual neurons into brain networks, we require algorithms that learn neuronal manifolds from empirical data. The goal of neuronal manifold

This work was supported under the CHIST-ERA grant (CHIST-ERA-19-XAI-002), by the Austrian Science Fund (FWF) (grant reference I 5211-N) and the Engineering and Physical Sciences Research Council United Kingdom (grant reference EP/V055720/1), as part of the Causal Explanations in Reinforcement Learning (CausalXRL) project.

learning is to find low-dimensional representations of the data that preserve particular properties. In neuroscience, a broad range of classical dimensionality reduction techniques is being employed, including but not limited to principal component analysis (PCA), multi-dimensional scaling (MDS), Isomap, locally linear embedding (LLE), Laplacian eigenmaps (LEM), t-SNE, and uniform manifold approximation and projection (UMAP) [5]. More recently, advances in artificial intelligence in general and deep learning methods, in particular, have given rise to a new class of (often non-linear) dimensionality reduction techniques, e.g., based on autoencoder architectures [6, 7, 8] or contrastive learning frameworks [9].

Common to all these techniques is their goal to reduce the data dimensionality while preserving particular properties of or information in the data. For instance, autoencoder-based frameworks typically focus on finding low-dimensional data representations that allow a good (or even perfect) reconstruction of the original, high-dimensional data. In contrast, we argue that reconstruction quality is only one of several desirable features for neuronal manifold learning. First, and in line with the argument by Krakauer et al. [10] that neuroscience needs behaviour, we argue that a neuronal manifold learning algorithm should not aspire to represent all but only those characteristics of high-dimensional neuronal activity patterns that are relevant in a given behavioural context. For instance, when studying an animal's ability to navigate a maze using visual cues, neuronal activity patterns that carry auditory or olfactory information are irrelevant in the behavioural context and should be abstracted away to better reveal the intricate relationships between neuronal representations of the visual cues and motor behaviour. Second, we argue that the reconstruction of the dynamics of the neuronal activity patterns should also take into account whether the low-dimensional embedding is causally sufficient in terms of the system's dynamics. To elaborate on this issue, consider the example of using a dimensionality reduction technique to learn the physical state description of a simple pendulum from a video stream showing the pendulum in action. Ideally, the dimensionality reduction technique should learn to represent the position and momentum of the pendulum for each video frame because these two variables constitute a full description of the system's physical state. In contrast, a dimensionality reduction technique that learns to represent the positions of the pendulum in the current and the past video frame only (without representing the pendulum's momentum) would also allow for a good reconstruction of the dynamics of the pendulum. This is the case because the pendulum's momentum, which is required to predict in which direction it will swing, can be approximately reconstructed from the difference in position across two video frames. However, this representation would not constitute a complete description of the actual physical state of the system. In analogy, a neuronal manifold learning technique should attempt to learn a complete physical state description of the underlying neuronal dynamics. Mathematically, this goal can be formulated as learning neuronal state trajectories that form a Markov chain because, in a Markov chain, the current state of the chain is causally sufficient for predicting the next state (in mathematical terms, the past and future states of the chain are statistically independent given the current state).

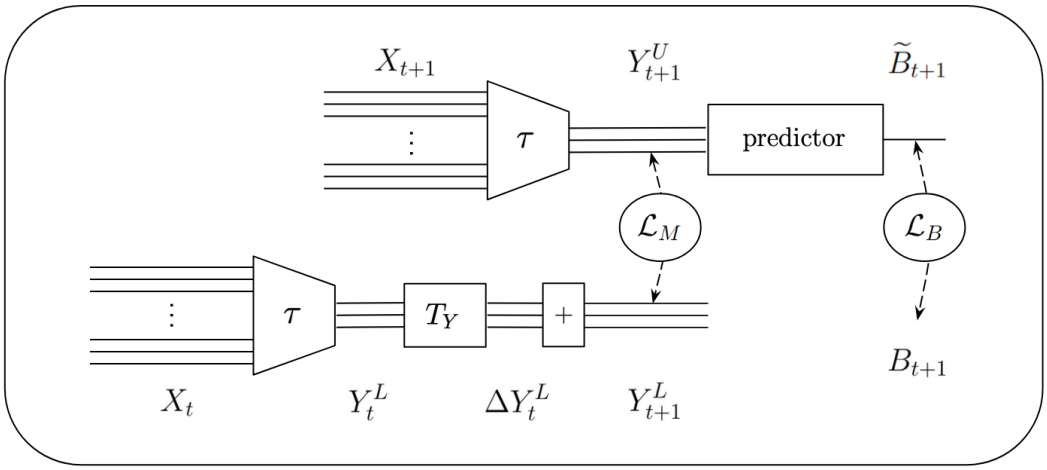

Figure 1: The BunDLe-Net architecture

Here, we introduce a novel framework for neuronal manifold learning, termed the Behaviour and Dynamics Learning Network (BunDLe-Net). BunDLe-Net learns a low-dimensional Markovian representation of the neuronal dynamics while retaining all information about a given behavioural context. It is based on the architecture shown in Fig. 1, which consists of two branches. In the lower branch, the high-dimensional neuronal trajectories X_t are first projected via a mapping τ to a lower-dimensional, latent trajectory Y^L_t. A first-order transition model T_Y then predicts the difference ΔY^L_t between the current and the next state to arrive at an estimate Y^L_{t+1} of the latent state at time $t+1$. This predicted latent state is compared to the true latent state at time $t+1$ in the upper branch, which is obtained by mapping the observed neuronal state X_{t+1} at time $t+1$ via the same τ as in the lower branch to the latent state Y^U_{t+1} via the loss function $\mathcal{L}_{\text{Markov}}$. By jointly learning the mapping τ and the first-order transition model T_Y that minimise the loss function $\mathcal{L}_{\text{Markov}}$ we obtain a latent, low-dimensional time-series Y_t that is Markovian by construction. This is the case because the transition model T_Y acts as a bottleneck that constrains the class of functions for τ for which the current state of the system is sufficient to predict the next state, in the sense that previous states do not provide any additional information. However, this architecture is not yet sufficient to learn a meaningful latent data representation because a mapping τ that projects the neuronal state trajectories to a constant ($Y_t = c$) would trivially fulfil the criterion of Markovianity. To obtain a meaningful latent representation, we also require that the behavioural context must be decodable from the latent representation Y_t by adding the loss function $\mathcal{L}_{\text{Behaviour}}$ that measures the reconstruction error between the true behavioural labels (B_{t+1}) and those predicted from the latent representation ($\tilde{B}_{t+1}$). By jointly learning that mapping τ and the first-order state transition model T_Y that minimise the two loss functions $\mathcal{L}_{\text{Markov}}$ and $\mathcal{L}_{\text{Behaviour}}$, the BunDLe-Net architecture learns low-dimensional Markovian representations of those aspects of the high-dimensional neuronal state trajectories that are relevant for a given behavioural context.

We remark that BunDLe-Net is a generic architecture in the sense that each of its modules (the mapping τ, the state transition model T_Y, and the prediction model for the behaviour)

can be realised by whatever models, e.g., linear or non-linear mappings which may be realised via (deep) neuronal networks or other modelling techniques, are most suitable for a certain type of neuronal data. The BunDLe-Net architecture is available as a *Python* toolbox at https://github.com/akshey-kumar/BunDLe-Net.

2 Results

We demonstrate the versatility of BunDLe-Net on three different data sets. We first focus on calcium imaging and behavioural data recorded in the nematode *C. elegans* [11] (see Section 4.7). We chose this dataset because probabilistic transitions between discrete behavioural states in this data set give rise to neuronal manifolds with a complex topology. In Section 2.1, we compare BunDLe-Net with existing state-of-the-art neuronal manifold learning techniques in terms of their ability to preserve dynamic and behavioural information in the data. We then compare the visual interpretability of the embeddings in terms of their topological structure in Section 2.2. In Section 2.3, we study the manifold's consistency across animals. Finally, we extend the analyses to spiking data with continuous-valued behavioural labels from rat hippocampus and primate somatosensory cortex in Section 2.4.

2.1 Quantitative evaluation of embeddings

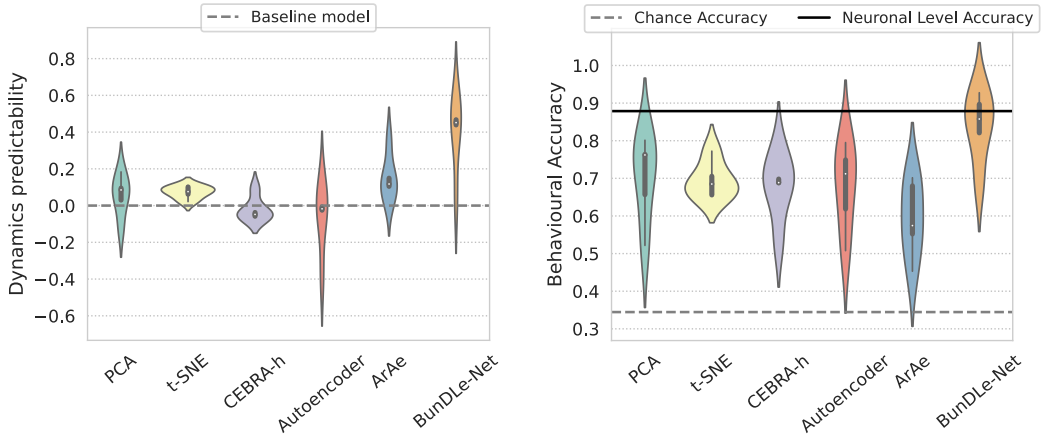

Figure 2: Quantitative comparison of BunDLe-Net with other commonly-used manifold learning techniques. (left) Evaluation of how well dynamical information is preserved in the embedding. The dashed line represents a baseline autoregressor which copies its input to the output. (right) Evaluation with respect to how well behavioural information is preserved. The dashed line represents the chance decoding accuracy, estimated by randomly shuffling the behavioural labels. The solid line represents the behaviour decoding accuracy from the raw neuronal traces.

We evaluate a latent space representation based on how well it preserves behavioural and dynamical information. As a metric for behavioural information, we use behavioural decoding accuracy. For dynamical information, we use a mean-squared-error-based predictability metric which estimates how well the latent state Y_t is able to predict Y_{t+1}, as elaborated in Section 4.4. We compare BunDLe-Net with other algorithms that are commonly used to learn high-level repre-

sentations in the field of neuroscience, such as PCA, t-SNE, autoencoder, an ANN autoregressor with an autoencoder architecture (ArAe)[1] and CEBRA-hybrid[2] (see Section 4.6). All embedding spaces were chosen to be three-dimensional for ease of comparison across algorithms and visualisation purposes. Implementations of the various models, training process, and evaluation procedures are available at https://github.com/akshey-kumar/comparison-algorithms.

Figure 2 presents the outcomes of our quantitative comparison, showcasing dynamical and behavioural prediction metrics in the left and right panels, respectively. Each panel depicts the predictability metric on the y-axis and the manifold learning technique on the x-axis, while the violin plots portray the metric's distribution across all five worm datasets. The variability across these plots underscores the diverse behavioural and dynamical attributes inherent in the dataset of each worm. For the dynamics evaluations, we compare all the models to a baseline model, which simply copies the input Y_t to the output as the predicted value for Y_{t+1}. For the behaviour evaluation, we compare with a chance-level behavioural decoding accuracy obtained by randomly shuffling the behavioural labels.

Turning to the results, we see that BunDLe-Net outperforms all other methods, including the state-of-the-art CEBRA, by a large margin. In the left panel, unsupervised methods like PCA, t-SNE, and the autoencoder show limited improvement over the baseline in predicting dynamics. Since they try to preserve maximum variance in the data in a low-dimensional space, they neglect to preserve minor details that may be crucial in determining future dynamics. CEBRA-hybrid, which also takes temporal information into consideration, does not perform better than the baseline model. The autoregressive-autoencoder, which seeks to reconstruct X_{t+1} from X_t, preserves some dynamical information and is seen to outperform PCA, t-SNE, and the autoencoder. Nonetheless, ArAe's reconstruction of the entire neuronal state at time $t+1$ can lead to irrelevant details persisting in latent space embedding. In contrast, BunDLe-Net's design focuses exclusively on retaining information pertinent to the latent space state at time $t+1$, which results in a markedly superior performance even compared to ArAe.

Shifting our attention to the right panel, all models surpass chance-level behaviour decoding accuracy. Notably, both CEBRA-h and the unsupervised methods (PCA, t-SNE, autoencoder) exhibit roughly the same performance on average. Despite this, their average decoding accuracy remains notably lower than neuronal-level decoding accuracy, indicating an inability to capture behavioural information at the neuronal level completely. This observation also applies to the ArAe despite its better performance in preserving dynamical information, which suggests that unsupervised preservation of dynamical attributes alone does not suffice for constructing behaviourally relevant models. In this regard, BunDLe-Net stands out by retaining all behavioural information as originally intended. On average, it even rivals the decoding performance achieved with raw neuronal data. For further evaluation of behavioural and dynamical performance of BunDLe-Net's embedding, please refer to Appendix A.

2.2 Visual interpretability of embeddings

In this section, we analyse the embeddings of BunDLe-Net and other competing neuronal manifold learning techniques. We visualise the embeddings of the Worm-1 in 3D and evaluate them qualitatively based on their structure and interpretability. We generalise the insights to all worms

[1] The ArAe would preserve dynamical information and embed it in a lower dimensional space due to the autoencoder architecture.

[2] Note that CEBRA as an algorithm was designed for continuous-valued behaviours. We cast our categorical behaviour (int) into a continuous behaviour (floating-point) and ran CEBRA on it.

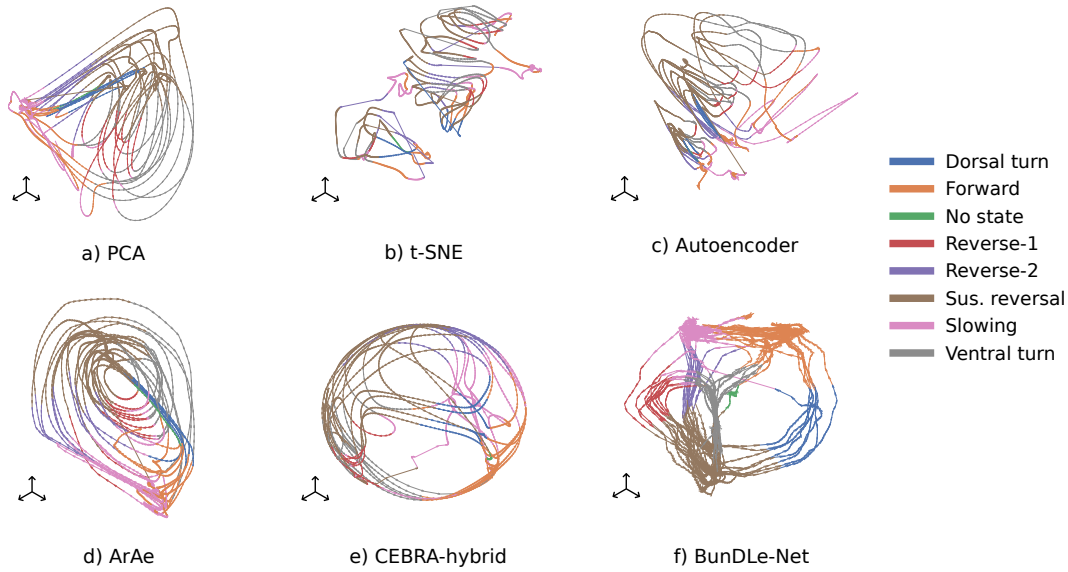

Figure 3: Neuronal manifolds learnt by various algorithms viz. a) PCA b) t-SNE c) Autoencoder d) Autoregressor-Autoencoder (ArAe) e) CEBRA-hybrid f) BunDLe-Net.

in the next section. Figure 3 shows the embeddings of Worm-1 by a) PCA, b) t-SNE, c) Autoencoder, d) Autoregressor-Autoencoder (ArAe), e) CEBRA-hybrid, and f) BunDLe-Net. In a), b) and c), we observe a noticeable drift in the PCA, t-SNE, and autoencoder embeddings. This drift drags out the dynamics in time, which is undesirable since we are searching for consistent mappings independent of time. The drift is also seen to obscure the recurrent nature of the dynamics to a large extent in b). The source of this drift could be a calcium imaging artefact or some neuronal dynamics irrelevant to our behaviour of interest. Since these models aim to preserve maximum variance for full-state reconstruction, they inadvertently embed the drift.

In contrast, in Figure 3 d), e), f), we see that this drift is largely absent, and the recurrent dynamics are more evident. These models have a common characteristic: they consider dynamics without attempting to reconstruct the entire neuronal state. Among the three methods shown, ArAe is unsupervised, while CEBRA and BunDLe-Net take behaviour into account. In both d) and e), we observe reasonably separated behaviours with minor trajectory overlaps. However, both embeddings demonstrate high variance *within* a trajectory of a given behaviour. In contrast, BunDLe-Net produces compact bundles that are well-separated from one another. The variance is low within each bundle, while a high variance is observed between different bundles. Consequently, BunDLe-Net's embedding exhibits distinct behavioural trajectories that are well-separated and along which the dynamics recur in an orbit-like fashion.

Additionally, in e), we observe that CEBRA-h embeds neuronal activity on the surface of a sphere. As a consequence, trajectories may be forced to intersect at certain points. Such intersection points are generally undesirable because they introduce ambiguity about the future trajectory. Ideally, intersection points should only occur when there is genuinely no information available about the subsequent behavioural trajectory. BunDLe-Net achieves this by producing highly compact embeddings with sparse intersections. These intersections can be interpreted as instances where BunDLe-Net encountered a lack of information about future trajectories, which

indicates the presence of true randomness at these points.

2.3 Consistency of neuronal manifolds across worms

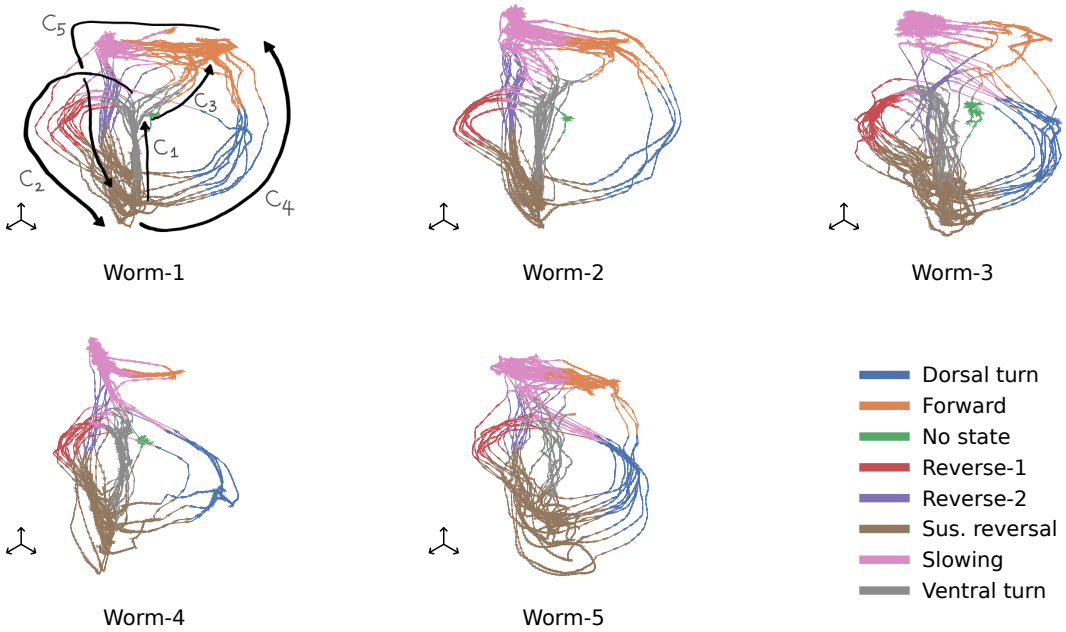

Figure 4: BunDLe-Net embeddings on five different *C. elegans* worm datasets which include neuronal recordings and behavioural labels. 3-D animations are available at https://github.com/akshey-kumar/BunDLe-Net/tree/main/figures/rotation_comparable_embeddings.

Here, we apply BunDLe-Net to all five worms in the dataset to visually compare the embeddings for consistency. This can be accomplished despite different subsets of neurons recorded from each worm by using the T_Y and the behaviour prediction layer from one worm (Worm-1 in this case) across all worms and only adapting the mapping τ to each individual worm, as detailed in Section 4.5.

Examining Figure 4, we observe a highly consistent branching structure in the trajectories of all five worms. The dynamics exhibit bundling of several segments, leading to recurring patterns along these bundles. Within each bundle, the dynamics are predominantly deterministic, while probabilistic *decisions* occur only at specific points in the trajectories. Focusing on the topology of Worm-1 and disregarding bundles consisting of fewer than two segments, we identify five prominent bundles,

$(\boldsymbol{C_1})$: sustained reversal ● → ventral turn ●

$(\boldsymbol{C_2})$: ventral turn ● → slowing ● → reversal-1 ● → sustained reversal ●

$(\boldsymbol{C_3})$: ventral turn ● → forward ●

$(\boldsymbol{C_4})$: sustained reversal ● → dorsal turn ● → forward ●

$(\boldsymbol{C_5})$: forward ● → slowing ● → reversal-2 ● → sustained reversal ●

These five motifs define the generic building blocks of the neuronal manifold in the sense that the neuronal trajectories are almost deterministic within each motif. As can be readily checked in Figure 4, these building blocks are highly consistent across worms, with similar behavioural motifs emerging across all worms. For example, motif $\boldsymbol{C_2}$ is consistently present in the embeddings of all worms, forming a loop (● → ● → ● → ●). The same holds true for motifs $\boldsymbol{C_1}$ and $\boldsymbol{C_5}$. However, motif $\boldsymbol{C_4}$ is not present in all worms and is notably absent in Worm-4. Instead, both Worm-4 and Worm-5 exhibit a slightly different motif (sustained reversal ● → dorsal turn ● → slowing ●). This variation in motifs may be due to the recording times, which may have been too short to capture all possible transitions for a given animal.

It is noteworthy that even though the individual worm recordings do not share an identical subset of neurons, the embeddings share a basic topological structure with only minor variations in transitions and branching points. These results demonstrate consistency in the embeddings across worms while preserving individuality in the behavioural dynamics in each worm and recording session.

2.4 BunDLe-Net on continuous behaviours

In this section, we demonstrate how the BunDLe-Net architecture can be readily adapted to different types of neuronal and behavioural data. In particular, we show its application to spike data with multidimensional, continuous-valued behavioural labels from rat and monkey brains. The Python implementation of these analyses is available at https://github.com/akshey-kumar/BunDLe-Net-continuous/.

Rat hippocampal data embeddings We used a publicly-available rat electrophysiology dataset consisting of neuronal and behavioural recordings as the rat navigates a 1.6-meter linear track [12] (see Section 4.7). The BunDLe-Net architecture required minimal modifications to accommodate new behavioural paradigms (see Section 4.2). Figure 5 shows the embeddings of four rats by BunDLe-Net and other commonly used algorithms. From the embeddings of all four rats, we see that the unsupervised methods such as PCA, t-SNE, autoencoder, and autoregressive-autoencoder (ArAe) fail at capturing any consistent structure in the neuronal data. From the embedding of rat-1 (Achilles), we see that embeddings of CEBRA and BunDLe-Net are *geometrically* different, with CEBRA embedding on the surface of a sphere and BunDLe-Net embedding on a plane. Despite these geometrical differences, BunDLe-Net and CEBRA's embedding manifest the same *topological structure* – a single closed loop. This suggests that both algorithms effectively capture the underlying topological structure of neuronal activity concerning the associated behaviour. In contrast to the embedding of *C. elegans* data, there are no discernible branching points in the trajectory of the rat's neuronal dynamics. Instead, a single deterministic loop is apparent, reflecting the rat's straightforward task of moving back and forth along the track to obtain rewards from either end, as indicated by human-annotated behaviour data. This also gives rise to the rectangular geometry since for each position, the rat can either be running to the right or left direction, corresponding to the two arms. Transitions only occur at the ends of the track, corresponding to the orthogonal arms.

Further, we see from BunDLe-Net's embeddings that both position and direction are well-preserved, as can be seen by the coloured labelling of the points. This implies that the recorded neurons encode information about the rat's position and direction in their spiking patterns. Our implicit behaviour decoding module within BunDLe-Net is clearly able to decode the positional and directional information from the neuronal activity. Notably, BunDLe-Net represents be-

havioural features orthogonally in the embedding, thus disentangling neuronal representations of position and direction found in the data. This finding is consistent with what is known about the hippocampus from decades of biological research [12].

Now, turning our attention to the embeddings of the other three rats, named Gatsby, Cicero, and Buddy, we see that BunDLe-Net's embedding remains quite consistent in terms of topological structure, i.e., a rectangle where one dimension encodes position and the orthogonal one encodes direction. In contrast, CEBRA's embedding becomes obscured by noise, resulting in a topological structure that is barely discernible. The data points in CEBRA's embedding are dispersed diffusely, while in BunDLe-Net they follow a more compact trajectory[3]. Additionally, we observe that BunDLe-Net's trajectories are far smoother in terms of transition from time step t to $t+1$. In contrast, CEBRA's embeddings show numerous discontinuous jumps in the trajectory that obscure any discernible dynamic structure. This contrast may be attributed to BunDLe-Net's dynamical loss function $\mathcal{L}_D$ and its very architecture, which is designed to preserve dynamical information.

These results demonstrate several strengths of BunDLe-Net: The ability to capture the underlying topological structure, robustness to noise in the data, smoother trajectories in terms of dynamics, orthogonal embedding of independent behaviours, and consistency of embeddings over animals.

[3] It should be noted that CEBRA embeddings over the surface of a sphere is an inherent feature of the algorithm itself and should not be interpreted as a specific characteristic of the neuronal manifold associated with this data.

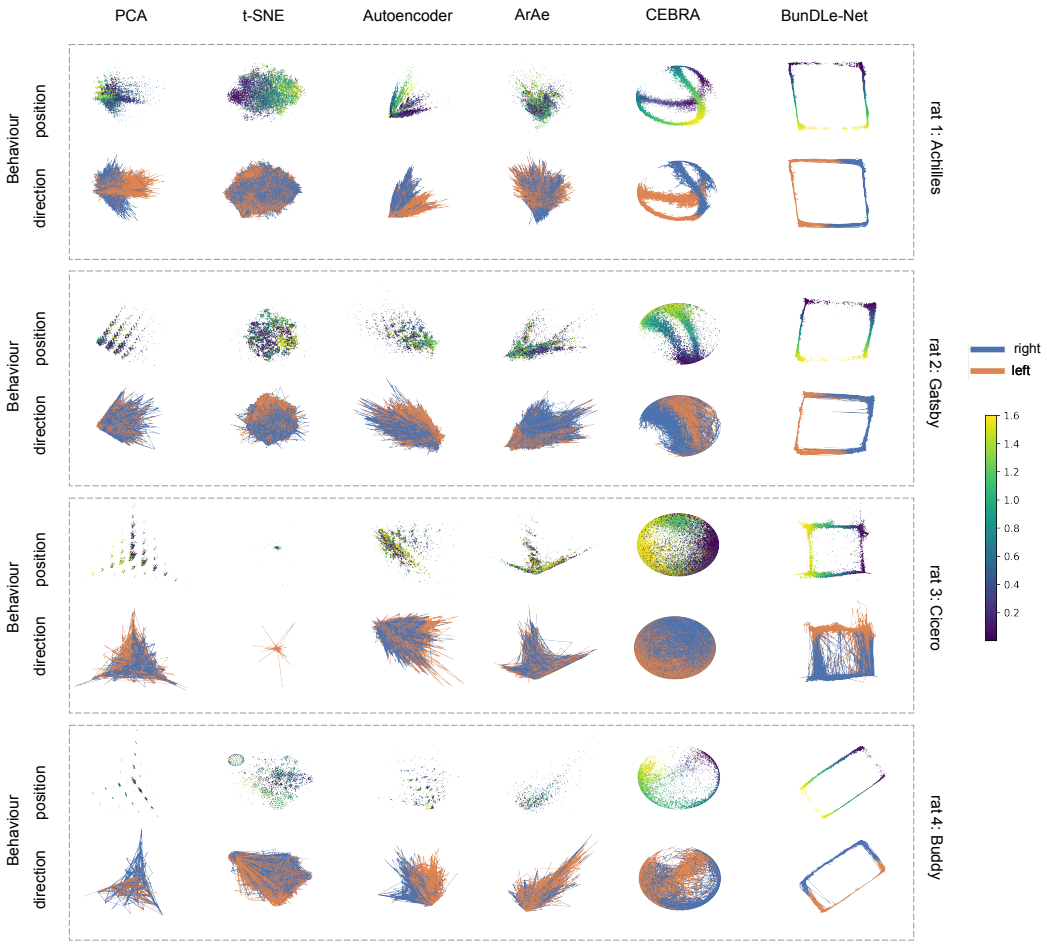

Figure 5: Comparison of rat hippocampus data embeddings in 3-D using various algorithms. Data from four rats are contained separately within the grey dashed boxes. Each column corresponds to the algorithm used for embedding the data. For each algorithm, the same embedding has been plotted twice, one labelled with a gradient colour map and the other with a binary colour map, which corresponds to the position and direction behaviour variable, respectively. This allows us to visualise how behaviours relate to the embedded neuronal activity. In the first plot, we plot the embedded data points, and in the second, we plot line segments between data points at time t and $t+1$ (to visualise how the trajectories evolve over time).

Primate dataset embeddings We now demonstrate BunDLe-Net using data collected from the primate somatosensory cortex during a center-out eight-direction reaching task [13] (see Section 4.7). Figure 6 shows the 3-D latent space embeddings of the neuronal activity by several algorithms. The colour labelling signifies the x and y coordinate of hand position. From the embedding by PCA and the autoencoder, we see that behavioural information is preserved to some extent since different behavioural states occupy distinct regions of 3-D space in the embedding. This would allow us to decode behaviour from these embeddings. Note, however,

that in the autoencoder embeddings dissimilar behaviours are mapped near each other while similar ones are separated. The violet and yellow regions are located near each other even though they correspond to positions that are the furthest apart. On the other hand, some green regions are fragmented and interspersed with dissimilar behaviours (lower plot). BunDLe-Net's embedding on the other hand demonstrates both a high behaviour decodability while also respecting the similarity structure of the behaviour. In fact, it is the only embedding where the task of reaching out in eight directions is evident from the topology[4]. The primate data embedding is different in structure from the worm and rat embeddings in that it does not have a recurrent structure to the dynamics, but instead a clear beginning and end of the reaching action. In BunDLe-Net's embedding, we see the beginning of the action depicted by the centre of the embedding, where there is no information of which direction the monkey is going to reach out to. Over time we see this information appearing as the embedding branches off into one of its eight arms, each of which corresponds to a direction in which the monkey reaches out to.

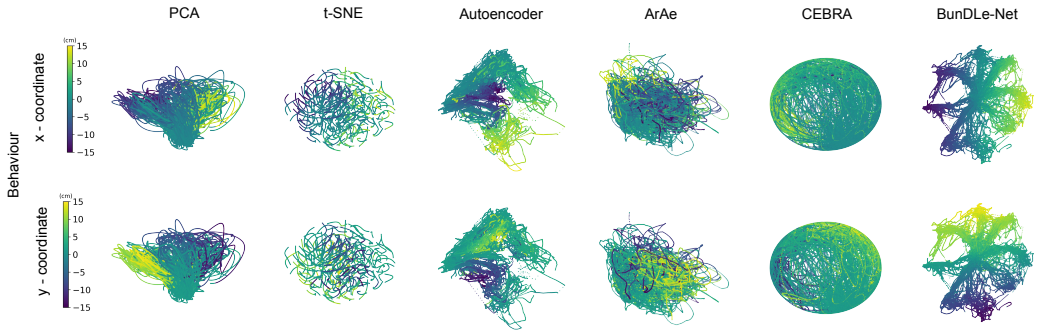

Figure 6: Embeddings in 3-D space of primate neuronal data.

3 Discussion

We have demonstrated the superiority of BunDLe-Net to other neuronal manifold learning techniques across various neuronal datasets from *C. elegans*, rats, and primates. We showed how BunDLe-Net can easily be extended to other imaging modalities and model organisms by adapting its learning modules while maintaining the overall structure shown in Figure 1. As such, BunDLe-Net is not one algorithm but a generic architecture for learning consistent state representations from neuronal data based on simple but vital principles. In the following, we further elaborate on the relevance of these principles for neuronal manifold learning.

On a fundamental level, the concept of a neuronal manifold can be interpreted as a scientific discovery that sheds new light on how large numbers of neurons coordinate their activities to represent information, implement computations, and generate behaviour. In this view, the goal of neuronal manifold learning techniques is to reveal the true, intrinsic structure of the neuronal manifold from empirical data. Alternatively, neuronal manifold learning algorithms can be interpreted as data compression and visualisation techniques. In this view, the particular

[4]No post-hoc averaging was required to reveal the behavioural patterns in the embedding unlike other methods as in [9]

shape of a neuronal manifold results from a model-based dimensionality-reduction technique that attempts to preserve certain data properties. Notably, these two viewpoints are not mutually exclusive, i.e., the observed shape of the neuronal manifold may be influenced by its intrinsic structure as well as by the particularities of the dimensionality reduction technique.

Indeed, our results in Figure 3 show substantial qualitative differences in the manifolds across various learning techniques, indicating that different model assumptions inherent to the various algorithms influence the shapes of the learned manifolds. On the other hand, the results obtained by BunDLe-Net shown in Figure 4 demonstrate that highly consistent manifolds can be learned across multiple animals, supporting the concept of an intrinsic structure of the neuronal manifold. Remarkably, BunDLe-Net achieves this consistency in *C. elegans* despite only 22 out of more than 100 neurons per animal being shared across the five data sets. We attribute this ability to reconstruct consistent manifolds to the time-delayed embedding of the neuronal dynamics for learning the latent dynamics (cf. Section 4.1), which due to Taken's theorem [14] allows the reconstruction of a Markovian representation of a dynamical system (i.e., the neuronal dynamics on the manifold) regardless of the specific observation function (i.e., the recorded neurons for each worm). We note that the number of time lags that need to be considered in this embedding is determined in BunDLe-Net by minimising the Markovian loss function $\mathcal{L}_{\text{Markov}}$, i.e., the number of time lags is increased until no further decrease in the loss function is observed.

Together with the constraint that the behavioural information must be preserved, BunDLe-Net's ability to learn a Markovian latent embedding results in almost deterministic trajectories that only exhibit a high degree of randomness at a discrete number of branching points (cf. Section 2.2 and Figure 4). This distinction in the neuronal dynamics between periods of high certainty with apparently random behaviour at a discrete number of branching points renders the neuronal manifold of *C. elegans* particularly interesting. Specifically, we interpret the almost deterministic trajectory bundles as the basic building blocks of the neuronal manifold that are fused together at the branching points to create the manifold's intrinsic structure.

The branching points act as decision points regarding the worm's future behaviour. However, it is presently unclear how *C. elegans* makes these decisions. In general, the randomness in the branching points could be due to intrinsic randomness in the neuronal activity or due to latent, unobserved neurons, i.e., observing these neurons could disentangle the branching points and again result in deterministic trajectories. However, BunDLe-Net's ability to learn Markovian representations would disentangle the branching points if such information were present in the time delay embeddings of the neuronal dynamics. Because this is not the case, our empirical results align with an interpretation in which the randomness in the branching points is intrinsic neuronal noise. However, we conjecture that such randomness might be overwritten by external stimuli, which were not part of the experimental design.

Regardless of the nature of the noise in the branching points, the learned neuronal manifolds reveal the behavioural flexibility of *C. elegans* in the context of its neuronal dynamics. In particular, they reveal when, i.e., at which points on the neuronal manifold, *C. elegans* makes decisions about its future behaviour. As such, we predict that external perturbations of the neuronal activity, e.g., by optogenetic stimulation, are most effective when applied at times when the neuronal state is in one of the branching points. Conversely, we hypothesise that the neuronal dynamics are more robust against external perturbations if these are applied when the neuronal dynamics follow one of the highly deterministic trajectory bundles. To generalise from this argument, we consider neuronal manifold learning algorithms in general and BunDLe-Net in particular, to be of extraordinary utility in neuroscience because these methods allow us to

make empirically testable predictions on how large-scale neuronal dynamics are coordinated to generate behavioural flexibility.

To conclude this article, we outline potential extensions of BunDLe-Net. We consider the extension of BunDLe-Net to multiple non-mutually exclusive behaviours, which could be used to study how large-scale neuronal activity coordinates multi-dimensional behaviours, straightforward. Naturally, this approach could be extended to include stimuli to study how external information is encoded in neuronal manifolds and translated into behaviour. Each of these changes would merely require adapting the behavioural prediction layer. Regarding the learning module for the latent embedding, we note the growing body of literature on the topic of (causal) representation learning. Representation learning addresses the problem of learning high-level (causal) variables from low-level observations [15, 16]; a topic with potentially rich synergies with neuronal manifold learning that are yet to be explored.

4 Methods

In this section, we first provide further information on the theoretical principles that motivate BunDLe-Net. Subsequently, we elaborate on the architectural framework that arises from these principles. We then proceed to provide a comprehensive overview of BunDLe-Net's implementation, encompassing the learning modules and the details of the training process. Finally, we present the competing methods that serve as benchmarks for evaluating the performance of BunDLe-Net.

4.1 Theoretical principle

BunDLe-Net employs a fundamental theoretical principle to embed neuronal data with respect to a given set of behaviours. The core idea is to ensure that the resulting embedding Y contains all information about the dynamics and behaviour that is present at the neuronal-level X. To elucidate this concept, consider the diagram in Figure 7, where T_X denotes a transition model at the X level. For illustrative purposes, we presently assume that the X level is Markov, but will later relax this assumption. The embedding Y is obtained by applying a function τ on the X level. Generally, the resulting transition model at the Y level may not be Markov, implying that Y_t might not fully capture the information about Y_{t+1} present in the system, either at the X level and/or in the past states Y_{t-n}, where $n \in \mathbb{Z}^+$. Such an embedding would be of limited use since one might need to refer back to the X level to answer certain questions about the Y level.

$$\begin{tikzcd} X_t \arrow[r, "\mathbf{T}_X"] \arrow[d, dashed, "\tau"] & X_{t+1} \arrow[d, dashed, "\tau"] \\ Y_t \arrow[r, "\mathbf{T}_Y"] & Y_{t+1} \end{tikzcd}$$

Figure 7: Commutativity diagram where the X and Y level signify the neuronal and latent space dynamics respectively.

To ensure a more comprehensive and self-contained embedding, we aim for Y to be Markov and independent of the X level. This requires the diagram (Figure 7) to commute, i..e. it

should not make a difference if we first time-evolve and then transform with τ, or the other way round. Put in terms of conditional independence, our requirement takes the form $Y_{t+1} \perp X_t | Y_t$, meaning that knowledge of X_t provides no additional information about Y_{t+1} beyond what is already known from Y_t. In this way, the dynamics at the Y level are self-contained and *sealed-off* from the details at X level. This is what makes our embedding so useful and interpretable: our embedding has all the relevant information from the X level, enabling it to be viewed as a distinct and meaningful dynamical process in its own right.

Non-Markovian neuronal dynamics To handle non-Markov neuronal dynamics at X, we consider time windows that include the previous n time steps, i.e., $(X_t, ..., X_{t-n})$ as input to our model. By choosing a large enough value for n, we can ensure that the resulting process becomes Markov [14], allowing us to model it in the same way as described above. Note that while earlier we were mapping a single time slice to a point in latent space, now we are mapping an entire time window of length n to a single point in latent space. Such a transformation does not merely coarse-grain over the neuronal or *spatial* level of granularity but also over the *temporal* domain of patterns.

Learning meaningful embeddings While the requirement of a Markov embedding may be very useful in terms of elegance and interpretability, it is not sufficient to ensure meaningful embeddings. For example, consider a transformation τ that uniformly maps every neuronal state to a constant. In this scenario, the resultant process would exhibit Markov dynamics as a single-state process. However, such an embedding fails to yield any meaningful insights regarding the underlying dynamics or behaviour. Remarkably, for BunDLe-Net, such a process would yield a perfect $\mathcal{L}_{\text{Markov}}$ loss, irrespective of the input data.

An additional requirement must be imposed to avoid such *trivial* embeddings. We demand that the behaviour B can be decoded from the embedding, thereby preventing the transformation from reducing everything to a mere constant. By upholding this crucial condition, we preserve the behavioural intricacies that render the embedding purposeful and informative, aligning with the ideals espoused by Krakauer et al. [10].

4.2 BunDLe-Net architecture

Here, we explain how the BunDLe-Net's architecture in Figure 1 arises from the commutativity diagram of Figure 7. The upper and lower arms in the architecture correspond to the possible paths from X_t to Y_{t+1} in the commutativity diagram. The lower arm in the architecture involves first coarse-graining X_t, followed by implementing a transition model on the Y-level. In practice, the transition model outputs ΔY_t from which Y_{t+1} is estimated as $Y_t + \Delta Y_t$. Since the transition model T_Y outputs Y_{t+1} with only Y_t as input, the Y-level is first-order Markov by construction. The upper arm of BunDLe-Net coarse-grains the time-evolved X_{t+1}[5]. Both arms result in estimates of Y_{t+1} which we distinguish by upper indices Y^L_{t+1} and Y^U_{t+1}. We add a mean-squared error term to our loss function $\mathcal{L}_{\text{Markov}}$ that forces Y^L_{t+1} and Y^U_{t+1} to be equal, thus ensuring that our requirement of commutativity in Figure 7 is satisfied,

$$\mathcal{L}_{\text{Markov}}(Y^U_{t+1}, Y^L_{t+1}) = \left\| Y^U_{t+1} - Y^L_{t+1} \right\|^2.$$

[5]Since we have time-series data, we need not learn T_X of the commutativity diagram, but simply feed X_{t+1} directly into the network.

The estimated Y_{t+1} is then passed through a predictor layer which learns to output the behaviour B_{t+1} given Y_{t+1}. Correspondingly, we add a term $\mathcal{L}_{\text{Behaviour}}$ to our loss function, which forces the predicted behaviour to match the true behaviour. This ensures that Y_t contains the same amount of information about B_t as X_t. Here, we use the cross-entropy loss where $B_{t+1}^{(j)}$ represents the j-th component of a one-hot encoded label vector of B_{t+1}, and $\tilde{B}_{t+1}^{j}$ is the softmax output of the predicted $\tilde{B}_{t+1}$.

$$\mathcal{L}_{\text{Behaviour}}(B_{t+1}, \tilde{B}_{t+1}) = -\sum_{j=1}^{8} B_{t+1}^{(j)} \log(\tilde{B}_{t+1}^{(j)}).$$

Both terms are weighted by a hyper-parameter γ and the loss function is given as,

$$\mathcal{L} = (1-\gamma)\mathcal{L}_{\text{Markov}} + \gamma\mathcal{L}_{\text{Behaviour}}.$$

All the layers in BunDLe-Net are learned simultaneously, and both loss terms ensure that the learned τ and T_Y preserve information about the behavioural dynamics. An open-source Python implementation of the BunDLe-Net architecture is available at https://github.com/akshey-kumar/BunDLe-Net.

Adaptation for continuous-valued multidimensional behaviour To deal with the new behavioural paradigms for the rat and monkey data, we modify the behaviour predictor module in Figure 1 to have as many outputs as there are behaviours. We use a mean squared loss function for $\mathcal{L}_{\text{Behaviour}}$ to deal with the continuous-valued position behaviour variable. Aside from this, the architecture and parameters are left unchanged.

4.3 Learning modules

Architecture parameters The τ layer (encoder) of our network consists of a series of ReLU layers [17], followed by a normalisation layer. An encoder of identical architecture is used later in the autoregressor-autoencoder (ArAe) model to facilitate comparison across models. For the *predictor* and T_Y layer, we use a single dense layer each. This sufficed to achieve good performance on the datasets used in this work. For T_Y, we also add a normalisation layer so that the output remains in the scaling of the latent space learned by τ. The details of the individual layers are provided in the Python code in Appendix B.

Gaussian noise against overfitting To safeguard against overfitting of the model, we introduce Gaussian white noise in the latent space by incorporating it in the τ layer. Injecting Gaussian white noise is a well-established regularisation technique that makes the model robust to overfitting [18, 19]. Since we are working with relatively limited data in the context of artificial neural networks, guarding against overfitting becomes particularly crucial.

Latent space dimensionality We choose the dimensionality of the Y-level to be three. This is because, in 3-D, we can connect any finite number of points without the edges crossing each other. This allows for embeddings of neuronal activity in the form of trajectories with nodes and edges that do not intersect. This might not always be possible in 2-D, where one can have a constellation of data points that cannot be connected without crossings. It is possible however, to embed any arbitrary graph in three dimensions without the edges having to intersect. [20].

Intersection points are undesirable for the embedding of a dynamical process due to the ambiguity they introduce. A meaningful embedding should exhibit smooth trajectories without self-intersections. An intersection point of two trajectories would mean that the past state at time $(t-1)$ contains additional information about the future state $(t+1)$ than the present state at (t), thus rendering the dynamics non-Markovian. Avoiding such intersections and non-Markovian dynamics enhances the interpretability of the embedded data and allows an enhanced prediction of future dynamics.

Model validation / parameter tuning To determine the optimal parameters for the model, including the number and types of layers, we use a held-out validation set on Worm-1. The neuronal and behavioural data of Worm-1 is partitioned into seven folds along the time axis, and one fold is randomly selected as the validation set from the time-ordered dataset. The remaining data forms the training set. By choosing an entire fold in the data as a validation set, we ensure that the model performs as well on unseen data. This would not be the case if we created our validation set by *iid* (independent and identically distributed) sampling due to high time correlations in the time series. After selecting the optimal model parameters through validation on Worm-1, we train models with the same parameters on the other worms. Since we only use Worm-1 for parameter tuning, if the model performs well on other worms, we can be confident that its success is not due to overfitting.

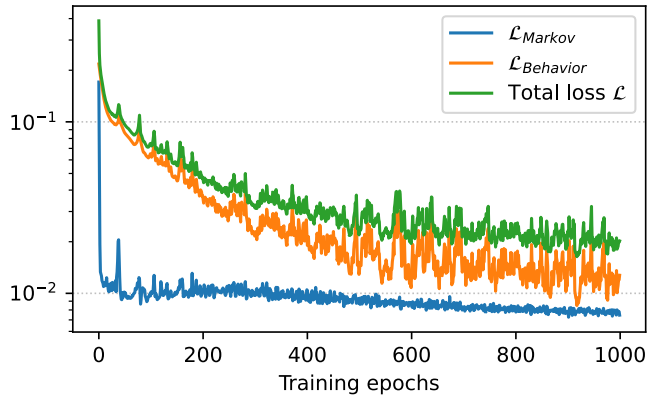

Figure 8: Markov and behavioural loss during training of BunDLe-Net on a log plot

Training details Since the neuronal data was found to be non-Markovian[6], we use time-windows of length 15 as input to BunDLe-Net. Reducing the window length decreased model performance while increasing it further had no significant effect. Training was performed with the ADAM optimiser [21] with a learning rate of 0.001 and batch size of 100. The γ parameter of BunDLe-Net was chosen to be 0.9 to ensure that $\mathcal{L}_{\text{Markov}}$ and $\mathcal{L}_{\text{Behaviour}}$ are roughly the same

[6]We tested for non-Markovianity using an autoregressor model and found that including multiple time steps from the past boosted the prediction performance of the model.

order of magnitude during training (see Figure 8). We trained BunDLe-Net until the losses converged.

4.4 Evaluation metrics

We evaluate a latent space representation based on how well it preserves behavioural and dynamical information. To estimate the *behavioural information* of an embedding, we train a simple feed-forward neural network in a supervised setting[7] to predict behaviour from the embedding. The decoding accuracy is then used as a metric for the information content about B in the embedding, with the decoding accuracy obtained on the raw, high-dimensional neuronal traces serving as the baseline. To evaluate the *dynamical information* in the embedding, we train an autoregressor network to predict Y_{t+1} from Y_t. The mean squared error between the predicted and true Y_{t+1} is estimated. From this, we compute a predictability metric for the dynamics, defined as $1 - \text{MSE}_m/\text{MSE}_{io}$, where MSE_m is the mean squared error of the model and MSE_{io} is the mean squared error of a trivial autoregressor that copies its input to the output (baseline performance). We trained all evaluation models on a training set of the embedded data and performed the evaluation on a held-out test set to prevent overfitting (see Model validation in Section 4.3).

4.5 Comparable embeddings across worms

To produce comparable embeddings[8], we first trained a model on each worm separately. We then extracted the T_Y layer and behaviour predictor layer from the model with the least loss (Worm-1, in this case). We then trained fresh models on each worm, with the chosen T_Y and behaviour predictor layers from Worm-1 frozen in, until the losses converged. Thus the new models would only have to learn the mapping τ for each worm while the other layers remained unchanged throughout the learning process. Notably, this approach was feasible despite recording different neurons from each worm. By adopting this strategy, we ensured consistent geometries across the worms, allowing us to effectively compare differences in topology, should they be present. The embeddings are illustrated in Figure 4. A latent dimension of three was again chosen for ease of visualisation, and can also be justified by a graph-theoretical argument detailed in Section 4.3.

4.6 Description of competing methods

Here, we describe the other commonly-used neuronal manifold learning algorithms used in the comparison. All models are used to project the neuronal data to a three-dimensional space for purposes of fair comparison. An implementation of the various models, training process, and evaluation procedures can be found at https://github.com/akshey-kumar/comparison-algorithms.

PCA Principal component analysis [22] has been applied to neuronal datasets to enable visualisation and interpretation of the data. It is a linear transformation that aligns the data along

[7] We use a simple architecture consisting of a single linear layer since it already demonstrated a high decoding accuracy (~ 0.94) on the raw neuronal traces. Hence, more complex models are not required to evaluate the embeddings.

[8] We could also simply fit separate models on each worm's data, as was done for the evaluation in Figure 2. Due to differing initialisations of BunDLe-Net this would result in visually different embeddings. These embeddings, however, share the same topology independent of the initialisation. For ease of visual comparison between embeddings, we adopt the above procedure to have latent spaces that can be mapped to one another.

the directions of maximum variance. Typically, the first three principal components are chosen and plotted in 3-D space [11]. The resulting trajectories can provide a rough perspective of the neuronal dynamics at a high level. Since this is a commonly-used method to coarse-grain data, we use PCA as our first baseline model.

t-SNE t-distributed stochastic neighbour embedding is a popular tool for visualising high-dimensional data, including neuronal data [23, 24]. It is essentially a non-linear dimensionality reduction method that tries to preserve distances between the data points.

Autoencoder Arguably, autoencoders (or some variant thereof) are currently one the most predominant methods for learning low-dimensional representations of data [6, 7]. Typically, an autoencoder learns a representation by attempting to reconstruct the training data using an ANN composed of an encoder and decoder [25]. Here, we consider the deterministic vanilla autoencoder with a deep encoder and decoder. The depth of the layers, number of neurons, and other training-related hyperparameters were tuned to obtain reasonably optimal performance.

Autoregressor-autoencoder (ArAe) An autoregressor is generally used on time-series data to predict the future state based on the past. Here we implemented an autoregressor with an ANN with an autoencoder-like architecture[9] and refer to it as ArAe. Such architectures have been used before to learn low-dimensional representations of time-series data [6, 26]. We implement our ArAe as an ANN with a deep encoder and decoder that tries to predict X_{t+1} given X_t as input with Y_t as the latent space as seen in Figure 10.

CEBRA CEBRA [9] is a state-of-the-art neuronal manifold technique. It uses contrastive learning to optimise the encoding of data by maximising the similarity between related samples and minimising the similarity between unrelated samples. The algorithm employs neural network encoders and a similarity measure to optimise the embeddings based on user-defined or time-only labels. In our experiments, we used CEBRA-hybrid, which takes both behaviour and time dynamics into account for the embedding.

4.7 Description of datasets

C. elegans data We apply BunDLe-Net to calcium-imaging whole brain data from the nematode *C. elegans* from the work by Kato et al. [11]. This dataset is ideal for demonstrating the capabilities of BunDLe-Net due to its high-dimensional neuronal recordings labelled with motor behaviour[10], multiple animal recordings, eight different behavioural states, behavioural flexibility and multiple repetitions of behavioural states over time. It includes time-series recordings of neuronal activation from five worms with human-annotated behaviours for each time frame. Each recording consists of approximately 2500-3500 time samples, spanning around 18 minutes (sampled at ~ 2.9 Hz) in which around 100 - 200 neurons are recorded. A low-pass filter with a cut-off frequency of 0.07 Hz is applied to mitigate high-frequency noise in the raw neuronal traces. Not all recorded neurons could be identified; hence, only a subset is labelled for each

[9] We use an autoencoder architecture since an autoregressor, in general, need not map the data to a low-dimensional space. Hence, we use an encoder to obtain an embedding that can be compared with the other methods

[10] The motor behavioural labels were inferred from the activity of the neurons AVAR, AVAL, SMDVR, SMDVL, SMDDR, SMDDL, RIBR, RIBL while the worms were immobilised. Hence, we removed these neurons from the dataset to ensure we are not inferring behaviours directly from these neurons.

worm, with different yet overlapping subsets identified across worms. The human-annotated behaviours B denote the motor state of the worm at a given instant of time and can take on one of eight states: forward, slowing, dorsal turn, ventral turn, no-state, sustained reversal, reversal-1, and reversal-2.

Rat hippocampal data We use a publicly available rat electrophysiology dataset from a study conducted by Grosmark and Buzsáki [12]. This dataset has been widely employed to assess different neural embedding techniques [9, 27]. It consists of simultaneously recorded electrophysiological and behavioural data as the rat navigates a 1.6-meter linear track. It is markedly different from the *C. elegans* dataset since we have spiking neuronal data (discrete-valued) instead of calcium image data (continuous-valued). Also, there are three behaviour variables that are simultaneously recorded and correspond to the rat's position and running direction (left or right). Notably, the position variable is continuous-valued while the direction is discrete-valued. Another significant difference is that the neuronal recordings are only from a small subset of the brain. Hence, this data set is more challenging with respect to finding a latent embedding that is causally sufficient for the neuronal dynamics. The BunDLe-Net architecture required minimal modifications to accommodate new behavioural paradigms (see Section 4.2).

Primate cortex data The data comprises neuronal recordings from Area 2 of the primary somatosensory cortex of a primate engaged in a centre-out eight-direction reaching task. This dataset, initially presented by Chowdhury et al. [13], is publicly available and has served as a benchmark for neuronal manifold learning [9, 28]. We used pre-processed data from [9], where time bins were employed and the estimated spikes of individual units were convolved using a Gaussian kernel to derive estimates of time-varying average spike rates. The behavioural variable comprises the x and y coordinates representing the position of a cursor controlled by the monkey during the reaching task. Simultaneously, extracellular neuronal recordings from the primary somatosensory cortex were captured.

5 Acknowledgements

We would like to thank Sebastian Tschiatschek and Simon Rittel for enriching discussions at the Causal Representation Workshop 2021 which was hosted at the Faculty of Computer Science, University of Vienna. We would also like to thank Manuel Zimmer and his lab, especially Kerem Uzel, for collaborating with us and providing the neuronal calcium imaging data from *C. elegans*. We also thank Verity Cook for discussions about the illustrations and figures.

References

[1] Jason ND Kerr and Winfried Denk. Imaging in vivo: watching the brain in action. *Nature Reviews Neuroscience*, 9(3):195–205, 2008.

[2] Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods*, 10(5):413–420, 2013.

[3] Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

[4] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571 (7765):361–365, 2019.

[5] Rufus Mitchell-Heggs, Seigfred Prado, Giuseppe P Gava, Mary Ann Go, and Simon R Schultz. Neural manifold analysis of brain circuit dynamics in health and disease. *Journal of Computational Neuroscience*, 51(1):1–21, 2023.

[6] N. Wahlström, Schön T. B., and M. P. Deisenroth. From pixels to torques: Policy learning with deep dynamical models. *Deep Learning Workshop at the 32nd International Conference on Machine Learning ICML 2015*, 2015.

[7] Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

[8] Chethan Pandarinath, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, L. F. Abbott, and David Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815, September 2018. doi: 10.1038/s41592-018-0109-9. URL https://doi.org/10.1038/s41592-018-0109-9.

[9] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, May 2023. doi: 10.1038/s41586-023-06031-6.

[10] John W Krakauer, Asif A Ghazanfar, Alex Gomez-Marin, Malcolm A MacIver, and David Poeppel. Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3): 480–490, 2017.

[11] Saul Kato, Harris S. Kaplan, Tina Schrödel, Susanne Skora, Theodore H. Lindsay, Eviatar Yemini, Shawn Lockery, and Manuel Zimmer. Global brain dynamics embed the motor command sequence of caenorhabditis elegans. *Cell*, 163(3):656–669, October 2015. doi: 10.1016/j.cell.2015.09.034. URL https://doi.org/10.1016/j.cell.2015.09.034.

[12] Andres D. Grosmark and György Buzsáki. Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. *Science*, 351(6280):1440–1443, 2016. doi: 10.1126/science.aad1935. URL https://www.science.org/doi/abs/10.1126/science.aad1935.

[13] Raeed H Chowdhury, Joshua I Glaser, and Lee E Miller. Area 2 of primary somatosensory cortex encodes kinematics of the whole arm. *eLife*, 9:e48198, jan 2020. ISSN 2050-084X. doi: 10.7554/eLife.48198. URL https://doi.org/10.7554/eLife.48198.

[14] Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer-Verlag, 1981. doi: 10.1007/BFb0091924.

[15] Thomas N. Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1gax6VtDB.

[16] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

[17] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[18] Alexander Camuto, Matthew Willetts, Umut Simsekli, Stephen J Roberts, and Chris C Holmes. Explicit regularisation in Gaussian noise injections. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16603–16614. Curran Associates, Inc., 2020.

[19] Richard M. Zur, Yulei Jiang, Lorenzo L. Pesce, and Karen Drukker. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical Physics*, 36(10):4810–4818, 2009. doi: https://doi.org/10.1118/1.3213517.

[20] Robert F. Cohen, Peter Eades, Tao Lin, and Frank Ruskey. Three-dimensional graph drawing. In Roberto Tamassia and Ioannis G. Tollis, editors, *Graph Drawing*, pages 1–11, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg. ISBN 978-3-540-49155-2.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

[22] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901. doi: 10.1080/14786440109462720.

[23] George Dimitriadis, Joana P. Neto, and Adam R. Kampff. t-SNE visualization of large-scale neural recordings. *Neural Computation*, 30(7):1750–1774, July 2018. doi: 10.1162/neco_a_01097. URL https://doi.org/10.1162/neco_a_01097.

[24] Jinke Liu and Martin Vinck. Improved visualization of high-dimensional data using the distance-of-distance transformation. *PLoS Computational Biology*, 18(12):e1010764, December 2022. doi: 10.1371/journal.pcbi.1010764. URL https://doi.org/10.1371/journal.pcbi.1010764.

[25] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. doi: https://doi.org/10.1002/aic.690370209. URL https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690370209.

[26] Boyuan Chen, Kuang Huang, Sunand Raghupathi, Ishaan Chandratreya, Qiang Du, and Hod Lipson. Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science*, 2(7):433–442, July 2022. doi: 10.1038/s43588-022-00281-6. URL https://doi.org/10.1038/s43588-022-00281-6.

[27] Ding Zhou and Xue-Xin Wei. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7234–7247. Curran Associates, Inc.,

2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/510f2318f324cf07fce24c3a4b89c771-Paper.pdf.

[28] Felix Pei, Joel Ye, David M. Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity. In *Advances in Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2109.04463.

A Further evaluation of BunDLe-Net's embedding

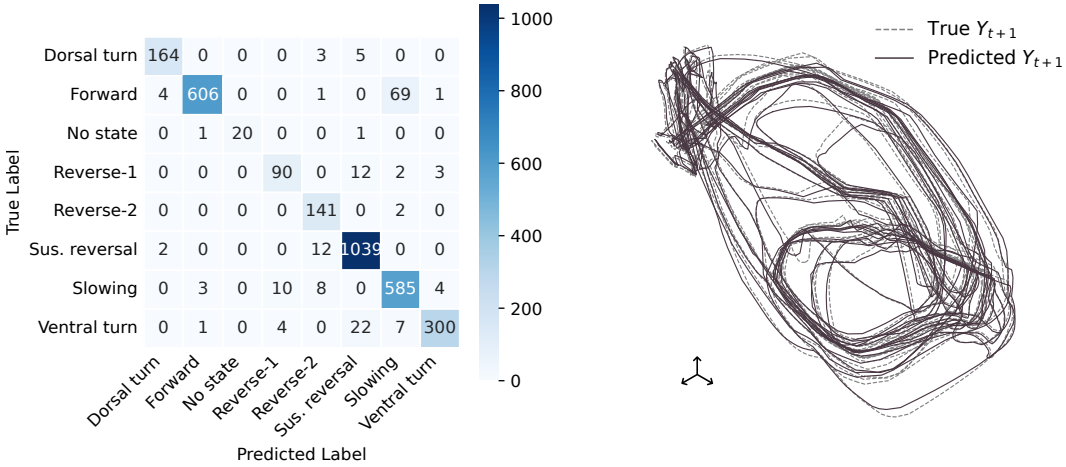

Figure 9: (left) Confusion matrix of behaviour predictions from BunDLe-Net's embedding for Worm-1 (right) True dynamics and dynamics predicted by BunDLe-Net for Worm-1.

In the following, we provide further information to build an intuition for the behavioural and dynamic prediction performance of BunDLe-Net. In Figure 9 a), we present the confusion matrix for BunDLe-Net's behavioural *prediction layer* from the ANN architecture. BunDLe-Net achieves a decoding accuracy of 94.3%, with the few decoding errors dominated by confusion of forward and slowing, two behaviours that are qualitatively similar and only quantitatively differ in the speed of the motion. To evaluate the dynamical performance of the model, we use the *transition model layer* T_Y to estimate Y_{t+1} from Y_t and compare it with the true Y_{t+1}, obtained as $\tau(X_{t+1})$. Figure 9 b) shows that the predicted dynamics indeed track the true dynamics rather well. These results indicate that the behaviour predictor and transition model within BunDLe-Net do well to preserve dynamical and behavioural information as intended.

B BunDLe-Net architecture

```python
class BunDLeNet(Model):
    """Behaviour and Dynamical Learning Network (BunDLeNet) model.

    This model represents BunDLe-Net's architecture for deep
    learning and is based on the commutativity diagrams. The
    resulting model preserves information relevant to the
    behavioural dynamics

    Args:
        latent_dim (int): Dimension of the latent space.
    """
    def __init__(self, latent_dim):
        super(BunDLeNet, self).__init__()
        self.latent_dim = latent_dim
        self.tau = tf.keras.Sequential([
            layers.Flatten(),
            layers.Dense(50, activation='relu'),
            layers.Dense(30, activation='relu'),
            layers.Dense(25, activation='relu'),
            layers.Dense(10, activation='relu'),
            layers.Dense(latent_dim, activation='linear'),
            layers.Normalization(axis=-1),
            layers.GaussianNoise(0.05)
        ])
        self.T_Y = tf.keras.Sequential([
            layers.Dense(latent_dim, activation='linear'),
            layers.Normalization(axis=-1),
        ])
        self.predictor = tf.keras.Sequential([
            layers.Dense(8, activation='linear')
        ])

    def call(self, X):
        # Upper arm of commutativity diagram
        Yt1_upper = self.tau(X[:,1])
        Bt1_upper = self.predictor(Yt1_upper)

        # Lower arm of commutativity diagram
        Yt_lower = self.tau(X[:,0])
        Yt1_lower = Yt_lower + self.T_Y(Yt_lower)

        return Yt1_upper, Yt1_lower, Bt1_upper
```

B.1 BunDLe-Net loss function

```python
def bccdcc_loss(yt1_upper, yt1_lower, bt1_upper, b_train_1, gamma):
    """Calculate the loss for the BunDLe Net

    Args:
        yt1_upper: Output from the upper arm of the BunDLe Net.
        yt1_lower: Output from the lower arm of the BunDLe Net.
        bt1_upper: Predicted output from the upper arm of the BunDLe Net.
        b_train_1: True output for training.
        gamma (float): Tunable weight for the DCC loss component.

    Returns:
        tuple: A tuple containing the DCC loss, behaviour loss, and total
                                                loss.
    """
    mse = tf.keras.losses.MeanSquaredError()
    scce = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True)

    DCC_loss = mse(yt1_upper, yt1_lower)
    behaviour_loss = scce(b_train_1, bt1_upper)
    total_loss = gamma*DCC_loss + (1-gamma)*behaviour_loss
    return gamma*DCC_loss, (1-gamma)*behaviour_loss, total_loss
```

C Other architectures of ANN models

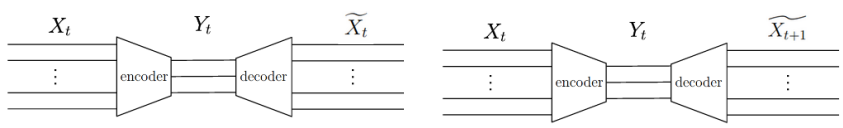

Figure 10: Architecture of Autoencoder and autoencoder-autoregressor (ArAe)

D Embedding of states in distinct behavioural motifs

Figure 11: (left) BunDLe-Net trajectory of Worm-1 highlighting the embedding of slowing behaviour (in pink) within two distinct bundles or behavioural motifs. (right) In contrast, the sustained reversal state (in brown) is represented by a single intersection point on the right side.

Behaviour can be modelled at different levels of granularity. In the present data set, the worms' behaviour is described in terms of high-level behavioural patterns such as forward and reversal movements. Alternatively, one could analyse the angular positions and velocities of the various segments of the worms' bodies, resulting in a more fine-grained representation. Both fine-grained and coarse-grained models hold value in specific contexts. However, it is crucial to maintain consistency within a model's state space to describe the dynamics accurately. If we utilise a model to understand fine-grained elements but only have access to coarse-grained information, the resulting model will be incomplete or inconsistent in the sense that it lacks the essential information required to predict features of the behavioural dynamics at the desired level of granularity. Here, we demonstrate how BunDLe-Net adeptly handles the coarse-graining of data while still preserving the crucial distinctions between states that are instrumental in explaining the overall dynamics.

We present the discovery of two distinct behavioural states with identical labels based on BunDLe-Net's neuronal embedding concerning the given set of behaviours. Consider branches $\boldsymbol{C_2}$ (● → ● → ● → ●) and $\boldsymbol{C_5}$ (● → ● → ● → ●) of Figure 4. The *slowing* behaviour (in pink) occurs in both these branches, i.e., they are represented distinctly in the latent space and are not fused together even though they have been assigned the same behavioural label. Branch $\boldsymbol{C_2}$ has a much shorter *slowing* segment than branch $\boldsymbol{C_5}$. We name the new behavioural states corresponding to $\boldsymbol{C_2}$ and $\boldsymbol{C_5}$ as *slowing 1* and *slowing 2*, respectively. These different types of slowing movements are embedded in distinct behavioural motifs since they differ in their neuronal realisation and their relevance for the model dynamics, i.e., one would predict different future trajectories depending on whether the state is *slowing 1* or *slowing 2*. We note that this is not the case for other behavioural states, e.g., the sustained reversal (in brown) for which all trajectories form one coherent bundle in the embedding. This implies that, in the behavioural state of a sustained reversal, BunDLe-Net found no information at the neuronal level to predict whether a dorsal or ventral turn is more likely to occur next. In summary, BunDLe-Net can maintain distinct representations or fuse trajectories depending on whether dynamical information about future behaviours is present. Accordingly, if provided with a set of behaviours that are not consistent or complete for the construction of a full dynamical model, BunDLe-Net can discover extra distinctions or *states* that complete this set of behaviours, provided this information is present in the neuronal level.

E Learning process

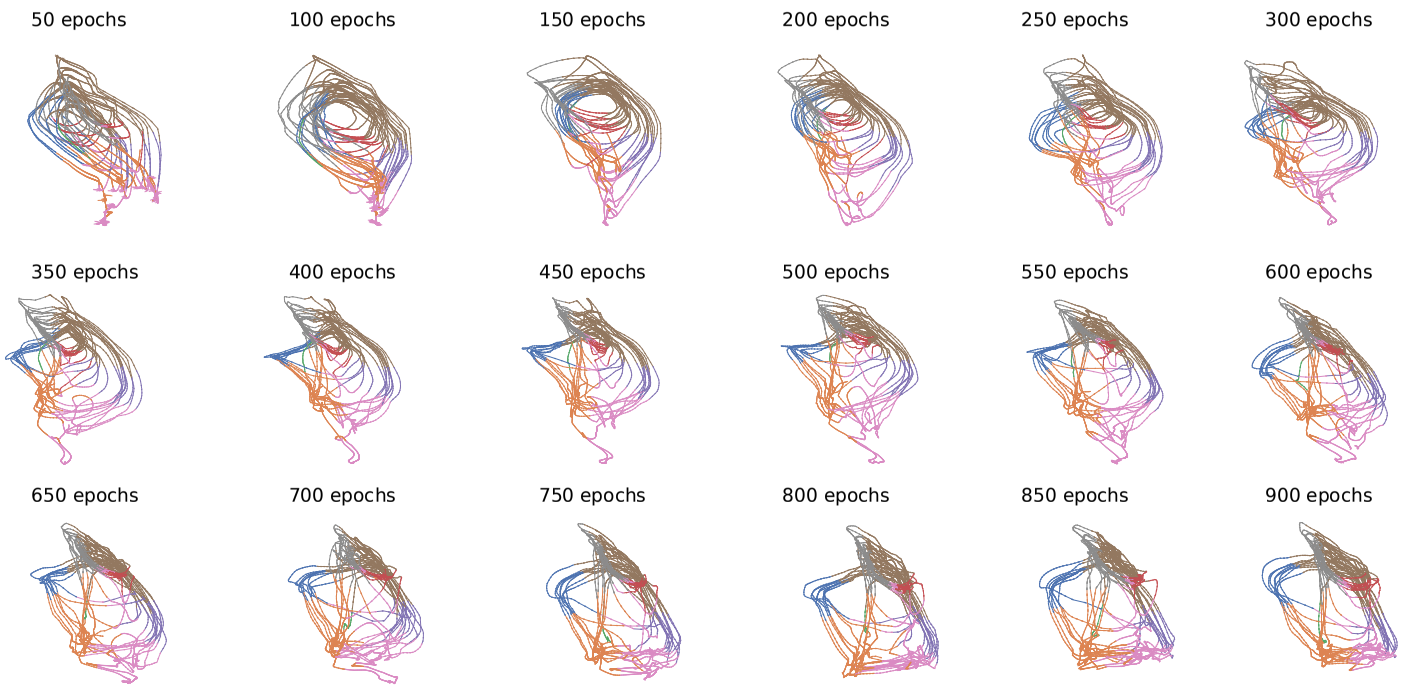

Figure 12: Visualisation of the learning process as a function of epochs