ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Interpretable Parametric Neighbour Embedding

Edouard Couplet[1], Pierre Lambert[1],
Michel Verleysen[1], John A. Lee[2,1], Cyril de Bodt[3]

1- UCLouvain - ICTEAM/ELEN
Place du Levant 3 L5.03.02, 1348 Louvain-la-Neuve - Belgium
2- UCLouvain - IREC/MIRO
Avenue Hippocrate 55 B1.54.07, 1200 Brussels - Belgium
3- UNamur - Dept. of Mathematics & naXys institute
Rue Joseph Grafé 2, 5000 Namur - Belgium

Abstract. Neighbour embedding methods effectively preserve local structures in low-dimensional spaces but are difficult to interpret due to their nonlinear nature, limiting their full potential for data exploration. Post-hoc interpretability methods exist but require extra effort and only approximate the embedding. We propose an interpretable-by-design neighbour embedding approach, where each data point is projected via a linear combination of shared basis matrices, enabling exact and direct explanations in terms of local coefficients and global directions. We demonstrate the approach on a single-cell dataset using a t-SNE loss, showing that it can provide useful interpretations while maintaining embedding quality.

1 Introduction

High-dimensional (HD) data can be explored visually through dimensionality reduction (DR) techniques, with neighbour embedding methods such as t-SNE [1] and UMAP [2] widely used in practice. Although these methods preserve local structure, their nonlinear nature makes it difficult to relate low-dimensional (LD) coordinates to original HD features, limiting interpretability [3]. Linear methods such as Principal Component Analysis (PCA) [4] provide interpretable projections but fail to capture complex structures. This contrast reflects a practical trade-off between DR quality and interpretability. Post-hoc interpretability approaches address this trade-off by fitting surrogate models, e.g., linear regressions [5, 6] or small decision trees [7], to explain nonlinear embeddings. These approaches maintain DR quality but yield explanations that only approximate the underlying embedding. We pursue the complementary direction: exact feature–embedding relationships, accepting a modest decrease in DR performance. In [8], we introduce parametric locally linear projections that produce high quality embeddings, with regularisation to control the interpretation granularity; however, this regularisation only partially constrains the projection matrices and does not impose their global coherence, which affects interpretation across data scales [9]. Therefore, we present a structured parametrisation that constrains these local projections to a common low-rank subspace spanned by a small set of basis matrices, improving both coherence and interpretation. Section 2 describes the parametrisation; Section 3 presents the implementation and its evaluation; Section 4 concludes with limitations and future directions.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2 Structured parametrisation

Let $\Xi = \{\xi_i\}_{i=1}^n$ be a set of n points in m dimensions. The goal of DR is to embed them as $X = \{\mathbf{x}_i\}_{i=1}^n$ in p dimensions, with $p \leq m$. Standard neighbour embedding methods optimise the low-dimensional coordinates directly, while parametric variants learn a mapping $\Xi \to X$, typically via neural networks, providing out-of-sample sample generalisation but still little interpretability. In [8], we propose an alternative parametric model $\mathbf{x}_i = \mathbf{W}_i^\top \xi_i$, assigning a distinct projection matrix $\mathbf{W}_i \in \mathbb{R}^{m \times p}$ to each data point. These matrices are learned individually with a Laplacian regulariser encouraging nearby points to have similar projections. While this setup produces competitive embeddings, it exhibits several limitations. Mainly, the set of matrices $\mathbf{W}_i$ has high cardinality and lacks a global structure despite regularisation, making it difficult to compare and interpret its elements. Also, the parameter count scales as $O(nmp)$, restricting applicability to small datasets. We address these issues by constraining each local projection $\mathbf{W}_i$ to lie in a low-rank subspace spanned by q basis matrices $\mathbf{B}_j \in \mathbb{R}^{m \times p}$: $\mathbf{W}_i = \sum_{j=1}^q \mathbf{c}_{ij} \mathbf{B}_j$, with $q \ll n$. The coefficients $\mathbf{c}_i = [c_{i1}, \ldots, c_{iq}]^\top$ are obtained from the input ξ_i via a simple parametric function: $\mathbf{c}_i = f_\theta(\xi_i)$. A schematic illustration of the full mapping $\xi_i \to \mathbf{x}_i$ is shown in Fig. 1.

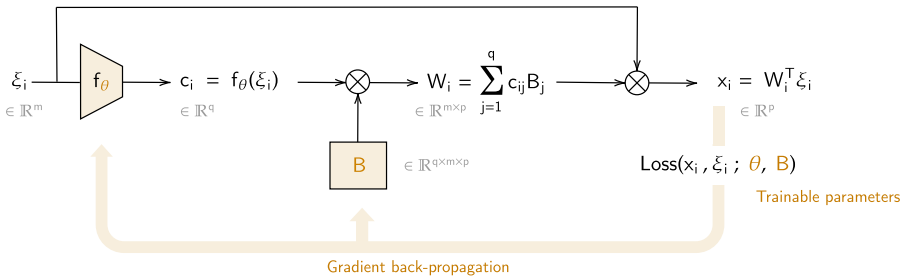

Fig. 1: Overview of the proposed parametrisation. LD coordinates are obtained via local linear projections: $\mathbf{x}_i = \mathbf{W}_i^\top \xi_i$. These projections are constrained to a low-rank subspace: $\mathbf{W}_i = \sum_{j=1}^q \mathbf{c}_{ij} \mathbf{B}_j$, with $\mathbf{c}_i = f_\theta(\xi_i)$.

In this approach, the learnable parameters are the tensor $\mathbf{B} \in \mathbb{R}^{q \times m \times p}$ collecting the basis matrices and the parameters θ of f_θ. These can be optimised jointly using a given neighbour embedding loss function. The parameter count drops from $O(nmp)$ to $O(qmp)$ (plus a small cost from f_θ), improving scalability. More importantly, given that f_θ can be chosen sufficiently smooth, similar points $\xi_\mathbf{i}$ will naturally map to similar coefficient vectors $\mathbf{c}_i$, and therefore similar projections $\mathbf{W}_\mathbf{i}$, without requiring explicit regularisation. Together with the shared global basis structure, this imposes coherence of local projections by design. The model becomes not only easier to optimise, but also to interpret. Indeed the decomposition $\mathbf{W}_i = \sum_{j=1}^q c_{ij} \mathbf{B}_j$ provides a clear interpretability mechanism: each basis matrix $\mathbf{B}_j$ represents a global projection direction, while each coefficient vector $\mathbf{c}_i$ specifies how data point i combines these directions.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

3 Experiments and Results

Dataset We test our approach on a single-cell transcriptomics dataset described in [10]. This dataset contains gene expressions for 23,822 cells from an adult mouse cortex. we preprocess it using a standard pipeline and select the 3000 most important genes, following the procedure suggested by [9]. As is common with single-cell data, we then perform a PCA to further reduce the dimensionality, keeping only the first 50 principal components. We will apply our interpretable neighbour embedding method in this space, with PCA loadings $\mathbf{V} \in \mathbb{R}^{3000\times 50}$ mapping genes to initial HD coordinates $\{\xi_i\}_{i=1}^{23,822}$, $\xi_i \in \mathbb{R}^{50}$.

Implementation We use the classical t-SNE loss, which measures the discrepancy between pairwise similarities in the LD space and in the HD space. LD similarities are computed with a Student-t kernel (one degree of freedom), and HD similarities are computed with Gaussian kernels whose bandwidth is set according to a user-defined perplexity. This parameter controls the scale of neighbourhoods which we want to preserve (local for low perplexities and global for high perplexities). We also allow averaging similarities across multiple preplexities to obtain a multiscale objective. The mapping f_θ is a standard neural network with two 128-unit layers, ReLU activations and softmax output. The number of basis matrices q is chosen according to the scale of interest in the interpretation: higher q emphasizes finer structure, while lower q targets broader structures. In our illustrative example, we focus on large scale structures and set $q = 3$. For visualisation, we project to two dimensions. The tensor $\mathbf{B} \in \mathbb{R}^{3\times 50\times 2}$ is initialised randomly.

Embedding quality We quantify embedding quality using three metrics: KNN, KNC, and CPD, capturing respectively local, regional and global structures. KNN is the average fraction of k-nearest neighbours in the HD space that remain among the k-nearest neighbours in the LD embedding, computed with $k = 10$. KNC is the average fraction of k-nearest class means in the HD space that remain among the k-nearest class means in the LD embedding, computed with $k = 10$. For our dataset, classes correspond to clusters extracted in [10]. CPD is the Spearman correlation between pairwise distances in the HD space and in the LD embedding, computed from 1000 sampled points. In Fig. 2, we compare the embedding obtained with our approach to the embedding obtained in [9], using their suggested t-SNE pipeline for single-cell transcriptomics data. We observe that both embeddings capture regional organisation equally well (KNC = 0.68). Moreover our model shows a reduced local neighbourhood preservation (KNN of 0.16 vs 0.41) while achieving a better global structure (CPD of 0.62 vs 0.51). This trade-off between local and global structure reflects the design choice of our method: by constraining projections to a shared low-rank subspace, we partly sacrifice fine-grained local neighbourhood fidelity in exchange for better preservation of mid- and large-scale structures, which are often more relevant in exploratory analysis.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

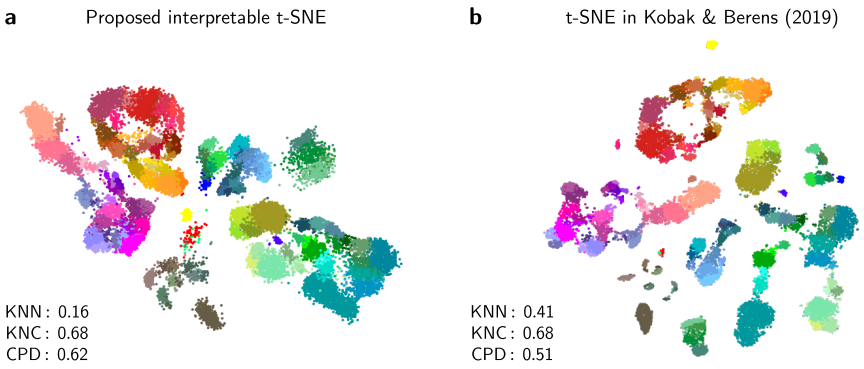

Fig. 2: Neighbour embedding applied to the Tasic et al. dataset. Cluster assignments and colours are from the original publication [10]. **a** Our proposed interpretable t-SNE, with a perplexity combination of 30 and $n/100 = 238$, random initialisation (of θ and $\mathbf{B}$), and adaptive learning rate (joint optimisation of θ and $\mathbf{B}$). **b** t-SNE from Kobak & Berens [9], with same perplexity combination, PCA initialisation (of LD coordinates), and learning rate $n/12 \approx 2000$ (direct optimisation of LD coordinates).

Embedding interpretability A key advantage of our structured parametrisation is that it provides several complementary layers of interpretability, each grounded directly in the model parameters rather than relying on post-hoc surrogate explanations. At the local level (Fig. 3a), the coefficient vectors $\mathbf{c}_i$ reveal how each point combines the global projection directions. In practice, we observe a gradual change of coefficients across the embedding which we attribute to the relative low comlpexity of our f_θ. This results in coherent local projections and prevents interpretations from being driven by isolated, noisy feature sets. At the regional/global levels (Fig. 3b), the basis matrices represent distinct projection modes that act consistently across all samples. Each mode can be interpreted by examining its dominant feature contributions. For the considered single-cell data, we recover gene contributions by projecting back to gene space via the PCA loadings $\mathbf{V}$. As an illustration (Fig. 3c), we identify top genes associated with $\mathbf{B}_1$ using $\text{argmax}_g ||(\mathbf{V}\mathbf{B}_1)_{g,:}||_2$. This basis projection primarily separates two broad clusters, driven by genes such as Gad1 and Gad2 (canonical GABAergic markers for inhibitory neurons) and Slc17a1 (a glutamatergic marker for excitatory neurons). This suggest that the two clusters correspond to inhibitory and excitatory neurons, which is consistent with the known annotations of the dataset [10]. In a real exploratory setting, we would not have these annotations and would need to verify any formulated hypotheses with targeted downstream analysis. Beyond descriptive analysis, the coefficient-basis decomposition also naturally supports interventional reasoning: amplifying or damping the effect of a basis matrix, perturbing specific feature contributions, or perturbing feature values, induces analytically predictable changes in the LD space, without the need to retrain the model. This enables direct "what-if" analyses that are inaccessible in standard nonparametric neighbour embedding methods.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964.
Available from http://www.i6doc.com/en/.

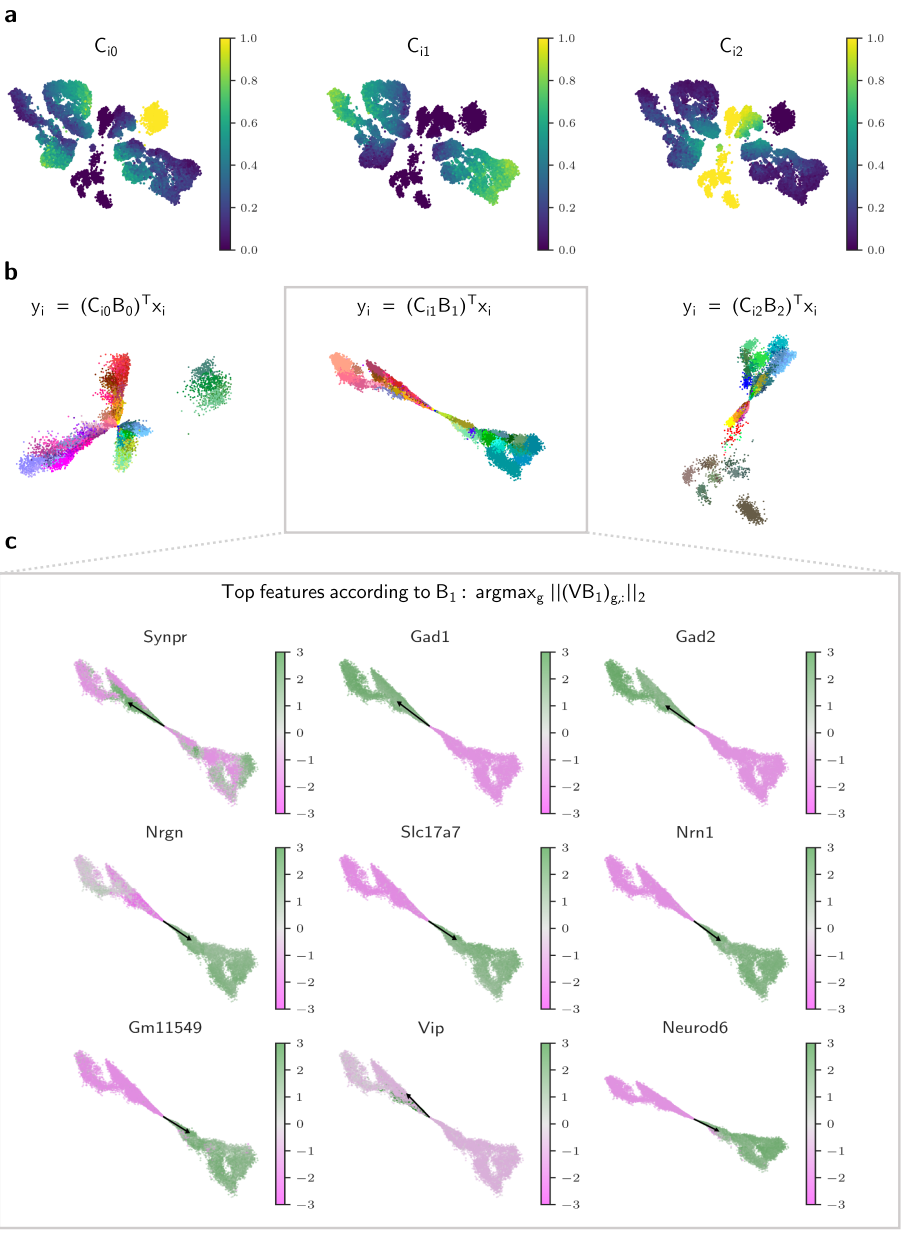

Fig. 3: Example interpretation of a t-SNE embedding produced with our approach on the Tasic et al. dataset [10]. Local projections are constrained to three basis projections. **a** Coefficient values across the embedding. **b** Samples projected according to each basis projection matrix. Cluster assignments and colours are taken from [10]. **c** Dominant gene contributions in $\mathbf{B}_1$ and corresponding (standardized) expression.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

4 Conclusion and Further Work

Our work introduces a structured, interpretable-by-design neighbour embedding method based on local linear projections, constrained to a shared low-rank subspace of basis matrices. This design provides exact feature–embedding relationships, enabling multi-level interpretability while maintaining strong embedding quality. More specifically, experiments on single-cell transcriptomics data show improved global structure preservation, though local neighbourhood fidelity is reduced. We argue that this local/global trade-off favors interpretability, as mid- and large-scale structures are often of primary interest. By analysing coefficients and dominant contributions in basis matrices, we reveal patterns aligned with biology. However, it is worth noting that the proposed parametrisation is domain-agnostic: identified features, directions, clusters, etc., are not guaranteed to carry biological meaning. One way to guide the model toward meaningful interpretations is to incorporate prior knowledge as an inductive bias in the coefficient structure. For example, single-cell datasets often exhibit a strong hierarchical organisation; the coefficients and basis matrices could be constrained to follow a tree structure reflecting this hierarchy. We plan to explore such approaches in future work. More broadly, our work contributes to a movement towards trustworthy data analysis, which is increasingly important as datasets grow in complexity and decisions rely ever more on data-driven insights.

References

[1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[2] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[3] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.

[4] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[5] Adrien Bibal, Viet Minh Vu, Géraldin Nanfack, and Benoît Frénay. Explaining t-SNE Embeddings Locally by Adapting LIME. In *ESANN*, pages 393–398, 2020.

[6] Pierre Lambert, Rebecca Marion, Julien Albert, Emmanuel Jean, Sacha Corbugy, and Cyril de Bodt. Globally local and fast explanations of t-SNE-like nonlinear embeddings. In *AIMLAI*, 2022.

[7] Adrien Bibal, Antoine Clarinval, Bruno Dumas, and Benoît Frénay. Ixvc: An interactive pipeline for explaining visual clusters in dimensionality reduction visualizations with decision trees. *Array*, 11:100080, 2021.

[8] Edouard Couplet, Pierre Lambert, Michel Verleysen, Dounia Mulders, John A Lee, and Cyril de Bodt. Natively Interpretable t-SNE. In *AIMLAI Workshop at ECML PKDD 2023*, pages 107–123. Springer, 2023.

[9] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):5416, 2019.

[10] Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.