LEARNING FMRI DICTIONARIES ACROSS INDIVIDUAL GEOMETRIES VIA OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Dictionary learning is a powerful tool for creating interpretable representations from complex data. When applied to functional magnetic resonance imaging (fMRI) data, the resulting patterns of brain activity can be used for various downstream tasks, such as brain activity decoding or population-level analysis. However, the variability in brain geometry across individuals poses a fundamental challenge. This is usually addressed by projecting each individual brain geometry onto a common template, which removes subject-specific information. In this work, we introduce a novel approach to dictionary learning on fMRI data that explicitly accounts for this variability. We use the optimal transport-based Fused Gromov-Wasserstein (FGW) distance to compare graphs with different geometries and features. To address the challenge of computing multiple FGW distances for large graphs such as those derived from fMRI data, we use amortized optimization to train a neural network that approximates the optimal transport plans. We also learn graph dictionary atoms that depend on the FGW trade-off parameter, which controls the balance between feature alignment and structural consistency. Numerical experiments on the HCP dataset demonstrate that the proposed approach captures meaningful information about both contrasts and subjects. It outperforms standard fMRI dictionary learning and is robust to transfer across graph scales.

1 INTRODUCTION

Dictionary learning for fMRI Functional magnetic resonance imaging (fMRI) measures local brain metabolism while subjects perform various cognitive tasks. The resulting 4D data are high-dimensional, noisy, and subject-specific, which makes learning representations particularly challenging. Dictionary learning (Mairal et al., 2009) provides an attractive framework for this setting because it represents data as an interpretable combination of atoms and produces compact and denoised representations. It has been successfully applied to fMRI data and outperforms classical decompositions such as Independent Component Analysis (ICA) (Hyvärinen & Oja, 2000) and Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999) on several downstream decoding tasks (Xie et al., 2017; Dadi et al., 2020). Prior work has applied dictionary learning to both task-based (Lv et al., 2015) and resting-state (Eavani et al., 2012) fMRI under the assumption that signals from multiple subjects are registered onto a common geometry. Multi-subject extensions have further considered learning shared dictionaries to estimate population-level functional atlases (Varoquaux et al., 2011).

The problem of inter subject variability A central challenge in multi-subject fMRI analysis is the substantial variability in brain anatomy and functional organization across individuals. These differences make it difficult to learn representations that can generalize across subjects while being sensitive to individual structure (Gordon et al., 2017; Thirion et al., 2021). Existing approaches often address this issue by projecting all subjects onto a common template geometry such as the FreeSurfer *fsaverage* template (Fischl, 2012), essentially discarding subject-specific information about brain structure. While practical, such registration procedures may rely on imperfect correspondences, distort subject-specific organization, and fail to fully account for inter-subject variability.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Graph dictionary learning When sampled on surfaces, individual fMRI data naturally admit a graph representation, whereby nodes correspond to brain locations and edges encode the neighboring structure between them. Dictionary learning was extended to graph-structured data in Vincent-Cuaz et al. (2021) through the use of the Optimal Transport based Fused Gromov Wasserstein (FGW) distance (Vayer et al., 2020), a powerful tool for comparing graphs with different geometries and node features. This framework represents graphs as linear combinations of learned graph atoms and has shown promising results on small graph datasets, where the learned representations capture both structural and feature information.
Applying graph dictionary learning (GDL) to fMRI data would overcome the need to register all subjects on a common geometry, as the FGW distance explicitly estimates alignments between graphs and therefore allows us to learn representations that account for geometrical variability. However, the computational cost of the FGW objective remains high, scaling cubically with the number of graph nodes. In practice, this restricts existing graph dictionary learning approaches to relatively small graphs containing only a few tens of nodes (Vincent-Cuaz et al., 2021). As a consequence, direct application of GDL to large-scale neuroimaging data such as fMRI graphs, which may involve hundreds or thousands of brain locations, remains challenging.

Contributions We propose Amortized Graph Dictionary Learning (AGDL), a new method for FGW graph dictionary learning that overcomes the two main limitations of FGW-based dictionary learning, which are the high computational cost of FGW and the need to choose the trade-off parameter α between feature and structure alignment. First, we use amortized optimization (Amos, 2023) to approximate the FGW distance, by introducing a neural network that learns solutions of the FGW problem in an unsupervised way from pairs of graphs sampled from the dataset. To that end we use the neural architecture from Unsupervised Learning of Optimal Transport (ULOT) (Mazelet et al., 2025) to approximate the FGW transport plans. This allows us to predict transport plans with a simple forward pass, which substantially reduces the computational cost of GDL and makes it applicable to large graphs. We also propose to learn dictionaries that explicitly depend on the α parameter to capture different levels of geometrical variability in the data and finely select the parameter with no computational overhead for downstream tasks. We apply AGDL to large scale fMRI data, on the HCP dataset (Van Essen et al., 2013), and show that it learns meaningful representations that account for both task-specific and subject-specific information. To the best of our knowledge, this is the first method for learning joint representations for brain activations and geometry on a large scale fMRI dataset.

2 FGW DISTANCE AND ITS APPROXIMATION WITH ULOT

In this section we introduce the Fused Gromov Wasserstein (FGW) distance, and present Unsupervised Learning of Optimal Transport (ULOT), a neural network architecture that learns to predict optimal transport plans between graphs.

2.1 FUSED GROMOV WASSERSTEIN DISTANCE BETWEEN GRAPHS

Definition of the FGW distance A graph is defined as $G = (\boldsymbol{F}, \boldsymbol{C})$ where $\boldsymbol{F} \in \mathbb{R}^{n\times d}$ denotes the node feature matrix, $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ encodes pairwise node distances or relations, typically through an adjacency matrix or a shortest-path matrix. Given two graphs G_1 and G_2 an optimal transport plan $\boldsymbol{P}(G_1, G_2)$ is a matrix $\boldsymbol{P} \in U(n_1, n_2)$ for $U(n_1, n_2) = \{\boldsymbol{P} \in \mathbb{R}_+^{n_1\times n_2} | \boldsymbol{P}\mathbb{1}_{n_2} = \frac{1}{n_1}\mathbb{1}_{n_1}, \boldsymbol{P}^T\mathbb{1}_{n_1} = \frac{1}{n_2}\mathbb{1}_{n_2}\}$, the set of couplings with uniform marginals, that minimizes the FGW loss $\arg\min_{\boldsymbol{P}\in U(n_1,n_2)} \mathrm{L}^\alpha(G_1, G_2, \boldsymbol{P})$ Vayer et al. (2020) for

$$\mathrm{L}^\alpha(G_1, G_2, \boldsymbol{P}) = (1-\alpha)\sum_{i,j=1}^{n_1,n_2} \left\|(\boldsymbol{F}_1)_i - (\boldsymbol{F}_2)_j\right\|_2^2 P_{i,j} + \alpha \sum_{i,j,k,l=1}^{n_1,n_2,n_1,n_2} |\,(\boldsymbol{C}_1)_{i,k} - (\boldsymbol{C}_2)_{j,l}\,|^2 P_{i,j}P_{k,l}. \tag{1}$$

The FGW distance is the combination of the Wasserstein term which measures discrepancies between node features in the left of equation 1, and the Gromov-Wasserstein term in the right of equation 1 that captures structural distortion between the two graphs. The parameter $\alpha \in [0, 1]$ controls the trade-off between feature alignment and structural alignment. The FGW distance is a

very general distance between graphs that can be used to compare graphs with different geometries, features and number of nodes. It has been successfully applied to a variety of tasks such as graph classification (Vayer et al., 2020), clustering and dictionary learning (Vincent-Cuaz et al., 2021).

Complexity of solving $\min_{\boldsymbol{P}\in U(n_1,n_2)} \mathbf{L}^\alpha(G_1, G_2, \boldsymbol{P})$ In order to solve the FGW problem one needs to minimize the FGW loss from equation 1 with respect to the transport plan $\boldsymbol{P} \in U(n_1, n_2)$. Because of the Gromov Wasserstein term, evaluating the FGW loss for a given plan $\boldsymbol{P}$ has a theoretical quartic complexity of $O(n_1^2 n_2^2)$. In the case of the square loss, Peyré et al. (2016) showed that this cost can be reduced to $O(n_1 n_2^2 + n_1^2 n_2)$, i.e., cubic complexity. However, this remains computationally prohibitive for large graphs, which in our setting typically contain around 1000 nodes. In the following section, we present a scalable solution to this problem based on Unsupervised Learning of Optimal Transport (ULOT) Mazelet et al. (2025).

2.2 Predicting FGW plans with amortized optimization

Unsupervised learning of optimal transport plans (ULOT) The FGW objective is non-convex and scales cubically with the number of nodes (Peyré et al., 2016), which makes direct minimization impractical for large graphs such as those arising in fMRI data. To compute transport plans efficiently, Mazelet et al. (2025) proposed to learn a neural network conditioned on α to learn approximation $\boldsymbol{P}^\alpha_\theta(G_1, G_2)$ of the optimal FGW transport plan for graph pairs $(G_1, G_2) \sim \mathcal{D}$, where $\mathcal{D}$ denotes a distribution of graphs. They introduced ULOT, a neural network architecture based on cross attention and Graph Convolutional Networks. The network is trained using amortized optimization, by directly minimizing the optimal transport loss. This enables efficient training without requiring ground truth correspondences. The training loss is the Fused Unbalanced Gromov-Wasserstein (FUGW) loss (Appendix, eq. 5), a relaxed version of the FGW loss that encourages the learned transport plan to satisfy the marginal constraints. A regularization parameter ρ controls the strength of this encouragement. Overall, ULOT is trained by minimizing the expected FUGW loss over graph pairs sampled from the dataset and values of α and ρ sampled from a distribution $\mathcal{P}$. At inference, predicting the transport plan between two graphs G_1 and G_2 is done by a simple forward pass through the trained model, which has a complexity of $O(n_1 n_2)$, thus enabling the application of FGW-based methods to large graphs. Note that the marginal relaxation of FUGW may discard parts of the graphs, resulting in a loss of information. This is undesirable in a dictionary learning setting, where we aim to preserve as much information from the input graphs as possible. So in this work we use ULOT with a high marginals penalization to obtain balanced plans and solve FGW.

3 Amortized dictionary learning across individual geometries

3.1 FGW graph dictionary learning

FGW graph dictionary learning The general formulation of FGW dictionary learning introduced in Vincent-Cuaz et al. (2021) is as follows. We represent a dataset of graphs $(G_i)_{i\in[N]}$ using a learned dictionary of *atoms* $(G^D_k)_{k\in[K]} = (\boldsymbol{F}^D_k, \boldsymbol{C}^D_k)_{k\in[K]}$. Each graph G_i is reconstructed as a linear combination of the dictionary atoms, where the reconstruction coefficients are referred to as *embeddings*. The FGW dictionary learning problem is defined as

$$\min_{\substack{(\boldsymbol{\omega}_i)_{i\in[N]}\in\Sigma_K \\ (\boldsymbol{F}^D_k, \boldsymbol{C}^D_k)_{k\in[K]}}} \frac{1}{N}\sum_{i=1}^{N} \min_{\boldsymbol{P}_i\in U(n_i,n)} \mathrm{L}^\alpha(G_i, (\boldsymbol{F}(\boldsymbol{\omega}_i), \boldsymbol{C}(\boldsymbol{\omega}_i)), \boldsymbol{P}_i) \tag{2}$$

where $\boldsymbol{F}(\boldsymbol{\omega}) = \sum_{k=1}^K \omega_k \boldsymbol{F}^D_k$ and $\boldsymbol{C}(\omega) = \sum_{k=1}^K \omega_k \boldsymbol{C}^D_k$, where $\boldsymbol{\omega} = (\omega_k)_{k\in[K]} \in \Sigma_K$ for $\Sigma_K = \left\{\boldsymbol{\omega} \in \mathbb{R}^+_K | \sum_k \omega_k = 1\right\}$. The dictionary is optimized using stochastic gradient updates computed on minibatches of graphs. At each stochastic step, the dictionary is first fixed and the unmixing step $\min_{(\boldsymbol{\omega}_i)_{i\in[N]}\in\Sigma_K} \frac{1}{N}\sum_{i=1}^N \min_{\boldsymbol{P_i}\in U(n_i,n)} \mathrm{L}^\alpha(G_i, (\boldsymbol{F}(\boldsymbol{\omega}_i), \boldsymbol{C}(\boldsymbol{\omega}_i)), \boldsymbol{P_i})$ is solved to obtain the embeddings for each graph in the minibatch, which are then used to compute the gradient of the dictionary atoms with the envelope theorem. Note that another approach for learning a graph dictionary with a GW barycenter model instead of a linear model has been proposed in Xu (2020), but using a GW barycenter is more computationally expensive.

Complexity of solving GDL The unmixing step consists in minimizing the FGW objective between each graph of the minibatch and its reconstruction, which is computed by minimizing the FGW loss with respect to the transport plan $\boldsymbol{P}_i \in U(n_i, n)$ for each graph G_i. Therefore, each unmixing step requires computing the optimal transport plan between each graph of the minibatch and its reconstruction q times, where q is the number of inner iterations for unmixing. This results in a complexity $O(pqn^3)$, where p is the number of outer iterations for dictionary updates, q is the number of unmixing inner iterations, and n is the number of graph nodes which we take to be the same for all graphs for simplicity. This complexity is prohibitive for large graphs, and therefore limits the application of GDL to small graphs with a few tens of nodes. This is several orders of magnitude smaller than the number of spatial units in fMRI data, which ranges from the hundreds to the thousands. In the next section, we present a scalable solution based on amortized optimization.

3.2 Amortized dictionary learning on fMRI data

Amortized dictionary learning We propose in this paper to replace the iterative FGW solver from GDL with a pretrained ULOT model (Mazelet et al., 2025). The ULOT model is augmented with positional information for the graph nodes. More precisely, ULOT receives augmented node features $\boldsymbol{F}_i^{pos} = [\boldsymbol{F}_i \mid \boldsymbol{Z}_i]$, where $\boldsymbol{F}_i$ contains the brain activation features and $\boldsymbol{Z}_i$ the corresponding 3D node positions. Importantly, $\boldsymbol{Z}_i$ is used only as positional information for ULOT and is not included in the feature term of the dictionary reconstruction loss. Thus, the feature term of the FGW reconstruction loss compares only the activation features $\boldsymbol{F}$, while its structural term compares the distance matrices $\boldsymbol{C}$. Given a graph G_i and its current reconstruction $\widehat{G}_i$, ULOT predicts the transport plan $\boldsymbol{P}_\theta^\alpha(G_i^{pos}, \widehat{G}_i^{pos})$, where G_i^{pos} and $\widehat{G}_i^{pos}$ denote the corresponding graphs augmented with 3D positional information. This predicted plan is then used to evaluate the FGW reconstruction loss between G_i and $\widehat{G}_i$.

Learning an α-dependent dictionary The FGW parameter α controls the trade-off between feature and structural alignment, and is given as input to the ULOT solver. Consequently, different values of α can lead to representations capturing different amounts of functional and geometric information. However, learning a dictionary from scratch for each value of α would be computationally expensive. Instead, we propose to amortize the dictionary itself by parameterizing its atoms as functions of α. We first encode α using a soft-binning representation $\hat{\alpha} = \mathrm{softbin}(\alpha)$. Two fully connected MLPs then map this representation to the activation features $\boldsymbol{F}_D(\alpha)$ and three-dimensional positions of the dictionary atoms $\boldsymbol{Z}_D(\alpha)$. The geodesic distance matrices $\boldsymbol{C}_D(\alpha)$ are computed from the 3D positions projected onto the unit sphere. The overall dictionary atoms are therefore given by:

$$\begin{cases} \boldsymbol{F}_D(\alpha) = \mathrm{MLP}_\psi(\hat{\alpha}), \\ \boldsymbol{Z}_D(\alpha) = \mathrm{MLP}_\phi(\hat{\alpha}), \end{cases} \qquad \begin{cases} \boldsymbol{F}_D^{pos}(\alpha) = [\boldsymbol{F}_D(\alpha) \mid \boldsymbol{Z}_D(\alpha)], \\ [\boldsymbol{C}_D(\alpha)]_{uv} = \arccos\big(\Pi_{\mathbb{S}^2}(\boldsymbol{Z}_D(\alpha)[u])^\top \Pi_{\mathbb{S}^2}(\boldsymbol{Z}_D(\alpha)[v])\big), \end{cases} \tag{3}$$

where $\Pi_{\mathbb{S}^2}$ denotes the projection onto the unit sphere; ϕ and ψ denote the parameters of the MLPs used to generate the dictionary atoms. For a graph G_i, the reconstruction associated with coefficients $\boldsymbol{\omega}_i^\alpha \in \Sigma_K$ is denoted $\widehat{G}_i(\alpha) = (\widehat{\boldsymbol{F}}_i(\alpha), \widehat{\boldsymbol{C}}_i(\alpha))$, where the reconstructed activation features and geometry are obtained by combining the α-dependent dictionary atoms according to $\boldsymbol{\omega}_i^\alpha$. We similarly denote by $\widehat{G}_i^{pos}(\alpha)$ the corresponding graph provided as input to ULOT, whose node features additionally contain the reconstructed 3D positions. During training, for each minibatch, we sample a value of α uniformly in $[0, 1]$, which is used to construct the dictionary atoms, for the ULOT model, and to evaluate the corresponding FGW reconstruction loss. The resulting optimization problem is given by

$$\min_{\psi,\phi} \ \mathbb{E}_{\alpha\sim\mathcal{U}(0,1)} \left[\frac{1}{N} \sum_{i=1}^{N} \min_{\boldsymbol{\omega}_i^\alpha \in \Sigma_K} L^\alpha \left(G_i, \widehat{G}_i(\alpha, \phi, \psi), \boldsymbol{P}_\theta^\alpha \left(G_i^{pos}, \widehat{G}_i^{pos}(\alpha, \phi, \psi) \right) \right) \right] \tag{4}$$

and can be solved by stochastic gradient descent, where the inner minimization over $\boldsymbol{\omega}_i^\alpha$ is solved using mirror descent for each graph in the minibatch.

3.3 Related works

Traditional dictionary learning (Tošić & Frossard, 2011; Mairal et al., 2009) has been successfully applied to vectorized fMRI data, where each brain location is treated as a feature and the dictionary

is learned on the resulting vectorized data (Lv et al., 2015; Shen et al., 2017; Iqbal et al., 2018; Khalid & Nauman, 2023). Large-scale versions using stochastic subsampling have been proposed to learn dictionaries on large fMRI datasets (Mensch et al., 2018; Dadi et al., 2020). Extensions have been introduced on graph structured data where the same structure is shared among all samples and the dictionary is learned on this shared structure (Thanou et al., 2014). However, these methods cannot account for the variability in brain geometry across subjects, which is an important aspect of fMRI data. In this work, we train on the individual geometry of the subject thanks to the use of OT that provides an alignment between structures. This is a crucial aspect of fMRI data as it encodes the spatial organization of brain activity.

AGDL builds on top of Graph Dictionary learning from Vincent-Cuaz et al. (2021), which introduced dictionary learning for graphs using the FGW distance. However this method relied on poorly scaling numerical solvers for the FGW distance, which makes the method inapplicable to large graphs. AGDL overcomes this limitation by using amortized optimization to learn a fast approximation of the FGW transport plans. This substantially reduces the computational cost and makes it applicable to large graphs.
Another important contribution of our method is that it learns a dictionary that explicitly depends on the α parameter. This makes it easy to compute dictionary atoms for different values of the trade-off parameter. This is particularly interesting because it can be seen as a multi-task learning problem for different values of the α parameter and is more efficient than the method in Vincent-Cuaz et al. (2021) that requires multiple training runs on a grid. Below, we demonstrate the application of AGDL to fMRI data, which is a particularly challenging setting due to the large size of the graphs and the high variability across subjects. In contrast, the method in Vincent-Cuaz et al. (2021) was only evaluated on small synthetic or molecule graph datasets.

4 Numerical experiments on fMRI data

4.1 Experimental setting

HCP dataset and DL baseline We train ULOT and perform dictionary learning on the HCP dataset (Van Essen et al., 2013) which consists of fMRI data of 1200 subjects performing a variety of tasks. We use the task fMRI data for 7 different tasks (emotion, gambling, motor, relational, language, social and working memory). The dataset contains two types of brain geometries for each subject, the common *fsaverage* geometry shared across individuals, and the individual geometry that is specific to each subject in terms of brain shape and folding. For each of these geometries, we build the dataset by running a General Linear Model (GLM) on both brain hemispheres using the Nilearn library (contributors) on the task fMRI data to obtain maps for the main contrasts in each subject. A contrast compares brain activity between two task conditions, highlighting the brain areas that are more or less active for the task of interest. For each subject we compute all its contrast maps, depending on which task signal is available. We detail the contrast computations in Appendix B.3.
Each contrast map is then represented as a graph, where nodes correspond to positions on the brain surface and edges encode the geodesic distance between these positions. Graphs are built by aggregating positions on the brain surface using the Ward algorithm (Thirion et al., 2014). This is done independently for each subject to keep their individual geometries. The node features are the average contrast value in each cluster concatenated with the 3d position of the node. We split the data per subjects into $70/15/15\%$ train/validation/test sets, and we use the same split for both the ULOT training and the dictionary learning.
As a baseline, we perform dictionary learning directly on the fMRI data projected onto the *fsaverage* common geometry, which is the standard practice in fMRI data analysis. In this setting, the reconstruction loss can be evaluated without solving an additional graph-alignment problem, and reduces to a standard least-squares objective. The proposed AGDL is done by training on native geometry graphs. We learn one dictionary for each brain hemisphere and choose the brain embedding as the concatenation of the left and right hemisphere embeddings. We detail the experimental setting in Appendix B.1.

FGW distance motivation Most dictionary-learning approaches for fMRI data rely on a least-squares loss function that projects brain activations onto a common geometry. The FGW distance considers not only brain activations, but also the brain's underlying geometry, which is essential for capturing subject-specific information. The α parameter controls the relative contribution of these two components, so that the distance can adapt to task-relevant information.

Distance used for kNN	Subject retrieval	Contrast retrieval
L2	0.23 ± 0.09	0.16 ± 0.05
FGW (best α)	$\mathbf{0.98 \pm 0.02}$	$\mathbf{0.22 \pm 0.06}$

Table 1: Accuracy of least-squares and FGW distances on a NN classification task. We report subject and contrast retrieval accuracy.

We experimentally compare those distances by evaluating their ability to discriminate contrasts and subjects by testing them on a simple nearest-neighbors classification task. Specifically, given a brain graph, we use a k-NN classifier to predict either the experimental contrast or the subject identity. As shown in Table 1 that reports classification accuracy, FGW consistently outperforms the least-squares loss, supporting its use as a discrepancy measure for our dictionary-learning framework.

Pre training for amortized optimization of the OT plans For amortized optimization of the FGW transport plans, we first pre-train a ULOT model on the HCP dataset by minimizing the FUGW loss equation 5 on pairs of graphs using either native-common or native-native pairs. We train the model stochastically by sampling $200\,000$ graph pairs per epoch and training for 105 epochs (see Appendix B.2), and jitter graph sizes between 900 and 1000 nodes to add some variability. We fix a batch size of 64 and a learning rate of 0.002. We sample α values from a Beta distribution with parameters $(0.5, 0.5)$ and ρ values from a log-uniform distribution on $[10^{-7}, 1]$. The model is then frozen for use in the subsequent dictionary learning step. For more details on the experimental setting and convergence analysis for ULOT pre training, see Appendix B.1.

ULOT transport plans validation Since AGDL relies on ULOT to predict the transport plan between each graph and its reconstruction, we need to assess the quality of these predicted plans. We perform this evaluation on smaller graphs with 100 nodes, for which transport plans can still be computed with a standard iterative solver at a reasonable computational cost. This evaluation is particularly important because previous experiments with ULOT on fMRI data were limited to a small dataset of 12 subjects (Mazelet et al., 2025), whereas here we consider a substantially larger-scale setting. Moreover, we evaluate ULOT on subjects that were not seen during training, allowing us to assess whether the pretrained model generalizes to new subjects without retraining.

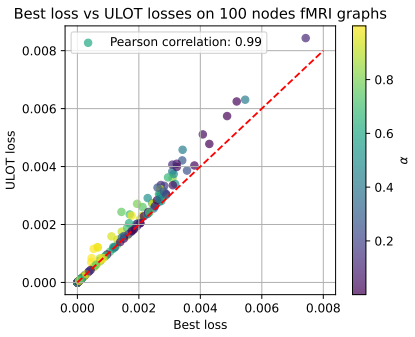


Figure 1: Comparison of the FUGW loss obtained with ULOT and the best loss for each graph in the test set.

For each pair of graph in the test set, we compute the FUGW loss using either the transport plan predicted by ULOT or a reference plan obtained with a standard iterative solver relying on batch coordinate descent and the IBPP solver (Xie et al., 2020; Thual et al., 2022). Since the FUGW objective is non-convex, the latter is not guaranteed to reach the global optimum. We therefore define the *best loss* as the minimum FUGW loss attained by either ULOT or the standard solver. Figure 1 compares the loss obtained with ULOT to this best attained loss. The two are highly correlated, with a Pearson correlation coefficient of 0.99, showing that ULOT consistently produces transport plans with objective values close to the optimal solution of the two methods.

Beyond objective values, we also compare the gradients of the FUGW loss with respect to the node features, since these gradients are used to update the dictionary atoms during training. The gradients computed using ULOT and the standard solver have an average cosine similarity of 0.93, indicating that ULOT provides a close approximation of the optimization signal used by AGDL. Finally, we evaluate how closely the ULOT transport plans predicted with the large value of $\rho = 0.9$ satisfy the marginal constraints during dictionary learning. The total variation distance between the predicted marginals and their corresponding uniform distributions is $6.58 \times 10^{-3} \pm 2.31 \times 10^{-6}$, indicating small deviations from the prescribed marginals. Further details are provided in Appendix B.2 figure

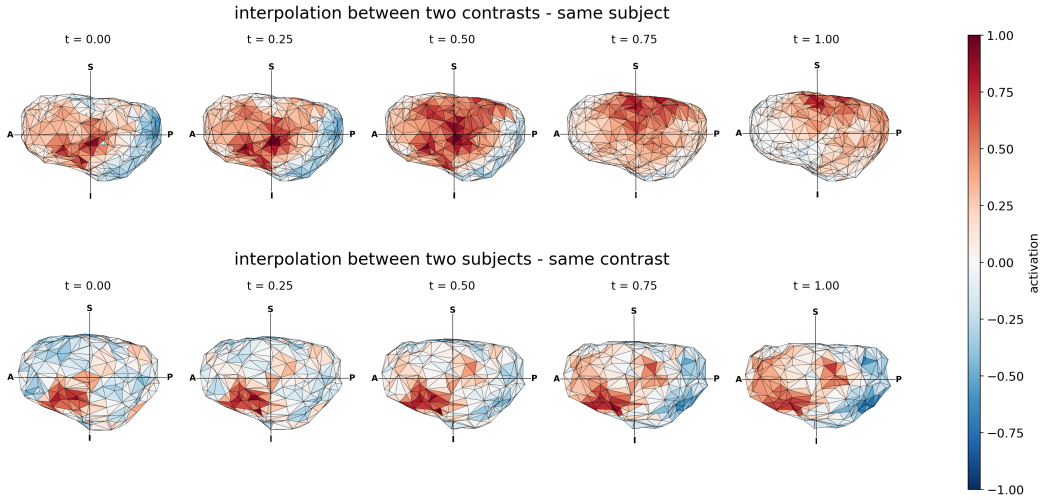


Figure 2: Interpolation between two reconstructed left hemispheres. The first row corresponds to the interpolation between contrasts *tongue - average* and *right foot - average* for a given subject, and the second row corresponds to the interpolation between two different subjects for the same contrast *tongue - average*. The axes are Anterior/Posterior and Inferior/Superior.

6. Together, these results support the use of ULOT as an efficient approximation of the transport optimization step required for dictionary learning.

AGDL We train AGDL stochastically on the HCP dataset for 1000 epochs with batch size 32 and learning rate 0.001, sampling 100 graphs per epoch. Atoms are initialized from HCP graphs on the common *fsaverage* geometry. Dictionary atoms and all training, validation, and test graphs contain 950 nodes. We consider $K = 60$ atoms, which is 4 times the number of contrasts, in order to get a model expressive enough to represent different levels of (subject) variability for each contrast. The α-conditioned dictionary MLPs have two layers with hidden dimension 16. The pairwise distance matrices are shortest path matrices on brain surfaces, and are normalized before evaluating the FGW loss so that their scale is comparable. For training, the values of α are sampled uniformly from $[0, 1]$. Further experimental details are provided in Appendix B.1.

4.2 Dictionary learning reconstruction and embeddings

Timing comparison between AGDL and GDL We compare the computational time of training AGDL and GDL on a single GPU for small graphs of size 100 nodes. For one epoch over 100 graphs with a batch size of 32, we find that GDL has an average runtime per epoch of 1265 seconds, while AGDL has an average runtime of 17.8 seconds. AGDL is about 70 times faster, which makes it applicable to large graphs, such as those arising in fMRI data, where GDL would be computationally prohibitive.

Interpolation between reconstructed brains Since AGDL reconstructs the geodesic distance matrices of the graphs, it is not designed to exactly reconstruct brain surfaces. However, we can visualize the reconstructed three-dimensional node positions and use them to obtain an approximate surface reconstruction. To illustrate the variability captured by AGDL, Figure 2 shows interpolations between pairs of reconstructed left hemispheres. The brain surfaces are reconstructed using the alpha shapes algorithm Edelsbrunner & Mücke (1994) Zhou et al. (2018).
The first row corresponds to an interpolation between the contrasts *tongue - average* and *right foot - average* for subject 115017. Since the subject is the same, we observe only limited geometric deformation, and the interpolation mainly reflects changes in brain activation between the two contrasts.
The second row corresponds to an interpolation between two different subjects for the same contrast *tongue - average*. In this case, we observe a much stronger geometric deformation, together with changes in the shape of the activation pattern. This illustrates that AGDL captures both subject-specific and contrast-specific information in the learned representations.

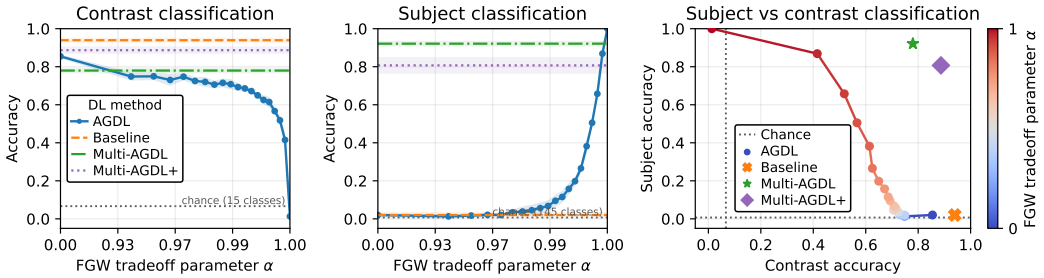


Figure 3: Classification accuracy for contrast (left) and subject (center) for different values of the FGW tradeoff parameter α from $\alpha = 0$ (Wasserstein) to $\alpha = 1$ (Gromov). We compare the baseline DL on common geometries. Joint visualisation of the two classification tasks (right). Multi-AGDL embeddings are obtained by concatenating the embeddings for 3 values of α, while Multi-AGDL+ includes the baseline embedding as well.

α-dependent embeddings for downstream classification tasks The fMRI graphs are characterized both by the identity of the subject from which they were obtained and by the experimental contrast that they represent. We therefore evaluate whether the learned representations capture both subject-specific and contrast-specific information.

We consider two downstream classification tasks: contrast classification and subject classification. For each value of α, we use the raw dictionary coefficients $\boldsymbol{\omega}$ as input to a logistic regression classifier. For contrast classification, the classifier is trained on embeddings from the training subjects and evaluated on embeddings from held-out test subjects, thereby measuring how well contrast information generalizes across subjects. For subject classification, we restrict the evaluation to the held-out subjects and split their contrast maps into training, validation, and test sets. We then train a logistic regression classifier to predict subject identity from the training embeddings and evaluate it on unseen contrasts from the same subjects.

As a baseline, we use linear dictionary learning on the common geometry and train the same logistic regression classifiers on the corresponding baseline embeddings. We repeat the complete evaluation procedure over five random seeds and report the average classification performance.

Classification performance of embeddings We show in Figure 3 (left) the accuracy of the AGDL contrast classifier for different values of α, and observe that the performance decreases with α with the highest value corresponding to $\alpha = 0$, which corresponds to dictionary learning with only the Wasserstein term which considers the node features only. The small gap in performance for $\alpha = 0$ likely comes from the fact that the baseline is designed to cancel-out subject variability with a fixed alignement on the common geometry whereas AGDL needs to recover an alignment that can introduce noise.

The subject identification task is more challenging and requires the embeddings to capture subject-specific information that is typically partially lost when projecting all subjects onto a common geometry. We show in Figure 3 (center) the performance of the classifier for different values of α, and observe that the best performance is obtained for $\alpha = 1$, which corresponds to a setting where the learned representations capture structural information only. This shows that subject specific information is mostly encoded in the geometrical structure of the graphs, and that discarding it results in a loss of subject specific information. These results highlight the strength of AGDL that can account for subject-specific information by performing dictionary learning directly in the native geometries as illustrated.

Learning a global representation across α values Different downstream tasks may benefit from different trade-offs between feature and structural information, as controlled by the FGW parameter α. Since embeddings obtained at different values of α emphasize different aspects of the data, we build a multi-view representation that aggregates information across multiple α values. For each graph, we concatenate the embeddings obtained for $\alpha \in \{0, 0.995, 1\}$ and train two logistic regression classifiers on the resulting representation, one for contrast classification and one for subject classification. As shown in Figure 3, this multi-view representation denoted Multi-AGDL outperforms the embeddings obtained from any single value of α on both tasks, achieving high accuracy for both contrast and subject classification. These results indicate that aggregating representations across

α values provides a strong representation capturing contrast-specific and subject-specific information. Importantly, this representation does not require selecting a task-specific value of α through validation, making it a convenient general-purpose representation for downstream classification. Additionally we define Multi-AGDL+ as the concatenation of the embeddings for $\alpha \in \{0, 0.995, 1\}$ and the baseline embeddings obtained from DL on common geometries, which further improves the classification performance on contrast classification. This shows that the baseline embeddings contain complementary information to the AGDL embeddings, and that combining them leads to a more discriminative representation.

Transfer learning across scales Contrary to dictionary learning on a shared geometry, AGDL does not require all graphs to be aligned and have the same number of nodes. This is because ULOT can be trained to predict transport plans between graphs of different sizes, without any modification to the architecture. This property allows us to perform transfer learning across scales, where we learn high resolution dictionaries and unmix graphs of different resolutions. This is particularly useful for fMRI data, where the number of nodes can vary significantly across subjects and datasets depending on the resolution. Also, this allows to perform super resolution of fMRI data by training a dictionary on high resolution graphs and unmixing low resolution graphs to obtain high resolution reconstruction. To do this, we trained a ULOT model on graphs of different sizes, from 80 nodes to 1020 nodes, and then trained a dictionary on the high resolution graphs of size 950 nodes. We then unmixed low resolution fMRI graphs of sizes $250, 500$ and 750 nodes on the high resolution dictionary. We find that the embeddings keep their discriminative power for both contrast and subject classification, as shown in Figure 4.

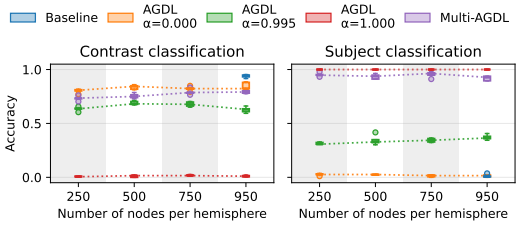


Figure 4: Classification accuracy for different sizes of low resolution graphs unmixed with a high resolution dictionary.

5 Conclusion

We propose a new method, AGDL, for doing optimal transport based graph dictionary learning on large graphs representing fMRI using amortized optimization with a pre-trained ULOT model to predict the transport plans used in the FGW loss and learning dictionaries conditioned on the optimal transport parameter α. We show that this method can be applied to large fMRI datasets and learns meaningful representations that capture both task-specific and subject-specific information.
This method goes beyond the common practice of projecting all subjects onto a common geometry, which biases the tight coupling between function and cortical geometry and thus degrades subject-specific information. AGDL is in line with Thual et al. (2022) that showed that projecting to a common geometry degrades subject-specific information, and that using FGW-based methods can help to capture this information. We go a step further by proposing a method for learning representations directly from the native geometries. To our knowledge, this is the first work to propose a scalable joint representation of graphs of different geometries while keeping information about their individual geometries. Moreover, the flexibility of FGW allows us to use the learned high resolution representations to unmix low resolution graphs. This also opens doors to transfer learning across datasets, to represent fMRI data from different datasets with a fixed dictionary.
This work is a first attempt to perform DL from fMRI data with heterogeneous geometries. As such, it has limits and calls for several potential extensions. First, even if using ULOT substantially reduces the computational cost of the method, it still requires evaluating the FGW loss at each iteration, which is cubic in the number of nodes. We could significantly reduce this cost by learning the atoms and the reconstruction coefficients in a fully amortized way (Amos, 2023) by training a neural network that takes as input a graph and outputs its embedding, which would remove the iterative optimization step for the reconstruction coefficients, and make the method applicable to even larger datasets and for near real time processing of fMRI data. Second, we could further study the robustness of our HCP dictionary by evaluating it on other fMRI datasets, and potentially learn a more general dictionary that can be used across datasets.

AI USE STATEMENT

In this work, we used generative AI tools for implementation, assisting with translation, and providing feedback on research experiments. We have not used generative AI tools to clean and reformat datasets, design research methodology or experiments, or interpret results. Generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the research paper to improve readability. We have reviewed all AI-assisted work: LLM-generated code was verified and tested for correctness by 1 author. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Brandon Amos. Tutorial on amortized optimization. *Foundations and Trends in Machine Learning*, 16(5):592–732, 2023.

Nilearn contributors. nilearn. URL `https://github.com/nilearn/nilearn`.

Kamalaker Dadi, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J. Gorgolewski, Demian Wassermann, Bertrand Thirion, and Arthur Mensch. Fine-grain atlases of functional modes for fmri analysis. *NeuroImage*, 221:117126, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.117126. URL `https://www.sciencedirect.com/science/article/pii/S1053811920306121`.

Harini Eavani, Roman Filipovych, Christos Davatzikos, Theodore D Satterthwaite, Raquel E Gur, and Ruben C Gur. Sparse dictionary learning of resting state fmri networks. In *2012 Second International Workshop on Pattern Recognition in NeuroImaging*, pp. 73–76. IEEE, 2012.

Herbert Edelsbrunner and Ernst P Mücke. Three-dimensional alpha shapes. *ACM Transactions On Graphics (TOG)*, 13(1):43–72, 1994.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Evan M Gordon, Timothy O Laumann, Babatunde Adeyemo, and Steven E Petersen. Individual variability of the system-level organization of the human brain. *Cereb. Cortex*, 27(1):386–399, January 2017.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Asif Iqbal, Abd-Krim Seghouane, and Tülay Adalı. Shared and subject-specific dictionary learning (shssdl) algorithm for multisubject fmri data analysis. *IEEE Transactions on Biomedical Engineering*, 65(11):2519–2528, 2018.

Muhammad Usman Khalid and Malik Muhammad Nauman. A novel subject-wise dictionary learning approach using multi-subject fmri spatial and temporal components. *Scientific Reports*, 13 (1):20201, 2023.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

Jinglei Lv, Xi Jiang, Xiang Li, Dajiang Zhu, Hanbo Chen, Tuo Zhang, Shu Zhang, Xintao Hu, Junwei Han, Heng Huang, et al. Sparse representation of whole-brain fmri signals for identification of functional networks. *Medical image analysis*, 20(1):112–134, 2015.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pp. 689–696, 2009.

Sonia Mazelet, Rémi Flamary, and Bertrand Thirion. Unsupervised learning for optimal transport plan prediction between unbalanced graphs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Jd8UcXJDCz.

Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Stochastic Subsampling for Factorizing Huge Matrices. *IEEE Transactions on Signal Processing*, 66(1):113–128, January 2018. doi: 10.1109/TSP.2017.2752697. URL https://hal.science/hal-01431618.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pp. 2664–2672. PMLR, 2016.

Hui Shen, Huaze Xu, Lubin Wang, Yu Lei, Liu Yang, Peng Zhang, Jian Qin, Ling-Li Zeng, Zongtan Zhou, Zheng Yang, et al. Making group inferences using sparse representation of resting-state functional mri data with application to sleep deprivation. *Human brain mapping*, 38(9):4671–4689, 2017.

Dorina Thanou, David I Shuman, and Pascal Frossard. Learning parametric dictionaries for signals on graphs. *IEEE Transactions on Signal Processing*, 62(15):3849–3862, 2014.

Bertrand Thirion, Gaël Varoquaux, Elvis Dohmatob, and Jean-Baptiste Poline. Which fmri clustering gives good brain parcellations? *Frontiers in neuroscience*, 8:167, 2014.

Bertrand Thirion, Alexis Thual, and Ana Luísa Pinho. From deep brain phenotyping to functional atlasing. *Current Opinion in Behavioral Sciences*, 40:201–212, August 2021. doi: 10.1016/j.cobeha.2021.05.004. URL https://inria.hal.science/hal-03289794.

Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in neural information processing systems*, 35:21792–21804, 2022.

Ivana Tošić and Pascal Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38, 2011.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Gaël Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, and Bertrand Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Biennial International Conference on information processing in medical imaging*, pp. 562–573. Springer, 2011.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty. Online graph dictionary learning. In *International conference on machine learning*, pp. 10564–10574. PMLR, 2021.

Jianwen Xie, Pamela K Douglas, Ying Nian Wu, Arthur L Brody, and Ariana E Anderson. Decoding the encoding of functional brain networks: An fmri classification comparison of non-negative matrix factorization (nmf), independent component analysis (ica), and sparse coding algorithms. *Journal of neuroscience methods*, 282:81–94, 2017.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pp. 433–453. PMLR, 2020.

Hongtengl Xu. Gromov-wasserstein factorization models for graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6478–6485, 2020.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.

A Appendix

A.1 AGDL algorithm

The algorithm used for AGDL is described in Algorithm 1, and the unmixing step is described in Algorithm 2. In practice, during unmixing we parameterize the simplex coefficients as $\omega_i = \text{Softmax}(z^{(i)})$ and optimize the logits $z^{(i)}$ by gradient descent.

A.2 Fused unbalanced Gromov-Wasserstein loss for amortized optimization of the transport plans

The ULOT model is trained by minimizing the Fused Unbalanced Gromov-Wasserstein (FUGW) loss, which is a generalization of the Fused Gromov-Wasserstein (FGW) distance that allows to compare graphs with different total mass and to learn transport plans that are not necessarily balanced. In this setting, graphs are defined with additional node weights γ representing their mass. We fix uniform weights $\gamma = 1/n$. It is defined as follows:

$$\begin{aligned}\text{FUGW}^{\alpha,\rho}(G_1, G_2, \boldsymbol{P}) = (1-\alpha) &\sum_{\substack{i=1,\dots,n_1\\ j=1,\dots,n_2}} \left\|(\boldsymbol{F}_1)_i - (\boldsymbol{F}_2)_j\right\|_2^2 P_{i,j} \\ &+ \alpha \sum_{\substack{i,k=1,\dots,n_1\\ j,l=1,\dots,n_2}} \left|(\boldsymbol{C}_1)_{i,k} - (\boldsymbol{C}_2)_{j,l}\right|^2 P_{i,j} P_{k,l} \\ &+ \rho\Big(\text{KL}(\boldsymbol{P}_{\#1} \otimes \boldsymbol{P}_{\#1} \,|\, \boldsymbol{\gamma}_1 \otimes \boldsymbol{\gamma}_1) \\ &\qquad + \text{KL}(\boldsymbol{P}_{\#2} \otimes \boldsymbol{P}_{\#2} \,|\, \boldsymbol{\gamma}_2 \otimes \boldsymbol{\gamma}_2)\Big).\end{aligned} \tag{5}$$

where $\boldsymbol{P}_{\#1} = \boldsymbol{P}\mathbb{1}_{n_2}$ and $\boldsymbol{P}_{\#2} = \boldsymbol{P}^T\mathbb{1}_{n_1}$ are the marginals of $\boldsymbol{P}$, and KL is the Kullback-Leibler divergence. The first two terms correspond to the FGW loss, while the last term is a penalty that encourages the transport plan to be close to a balanced plan. The ρ parameter controls the strength of this penalty, with larger values of ρ encouraging more balanced plans. In practice, we sample ρ from a log-uniform distribution on $[10^{-7}, 1]$ during training to learn transport plans with different levels of unbalancedness, which adds variability to the dataset and improves the generalization of the ULOT model.

B Experimental details

Training for ULOT and dictionary learning are performed on one H100 GPU and all the hyperparameters are validated using the Optuna library (Akiba et al., 2019). We detail the hyperparameters used for both trainings in the following sections.

B.1 Dictionary learning training

Hyperparameters for dictionary learning are given in Table 2. Dictionary learning training takes about 20 hours.

B.2 ULOT training

ULOT is trained for 105 epochs, and we show in Figure 5 the validation loss along epochs. Training takes approximately 200 hours. We use the hyperparameters in Table 3. We compute the mass of the

Table 2: Dictionary learning hyperparameters

Hyperparameter	
Learning rate outer loop	0.001
Learning rate inner loop	0.09
Batch size	32
Optimizer	AdamW
Weight decay outer loop	$1e^{-4}$
Weight decay inner loop	$1e^{-7}$
Number of bins for α encoding	7

Table 3: ULOT hyperparameters

Hyperparameter	
Learning rate	0.002
Batch size	64
Optimizer	AdamW
Number of node embedding layers	3
Embedding dimension for α	16
Node embedding layer final out dimension	64
MLP hidden dimension	1024
GCN hidden dimension	128
Temperature value a	10

ULOT transport plans for different values of ρ in Figure 6. We observe that for $\rho = 0.9$, the mass of the transport plans is close to 1, which corresponds to balanced plans. This is the value we use for AGDL, which produces balanced plans.

B.3 Experimental details on contrasts

We derive contrast maps from signals in both native geometry and fsaverage space from the HCP dataset (Van Essen et al., 2013). For each task, we select the main contrasts reported in Table 4. Geodesic distances on the cortical surface are approximated by shortest-path distances on the corresponding graphs. The medial wall is excluded from all contrast maps.

Table 4: HCP tasks used for alignment

Task	**Contrast**
GAMBLING	win - loss
EMOTION	fear - neutral
MOTOR	left hand - average
MOTOR	right hand - average
MOTOR	left foot - average
MOTOR	right foot - average
MOTOR	tongue - average
RELATIONAL	match - relation
WM	body - others
WM	faces - others
WM	tools - others
WM	places - others
WM	bk2 avg - bk0 avg
SOCIAL	mental - random
LANGUAGE	story - math

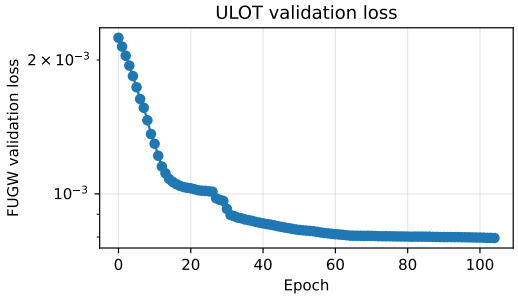


Figure 5: FUGW loss on validation dataset for ULOT training along epochs

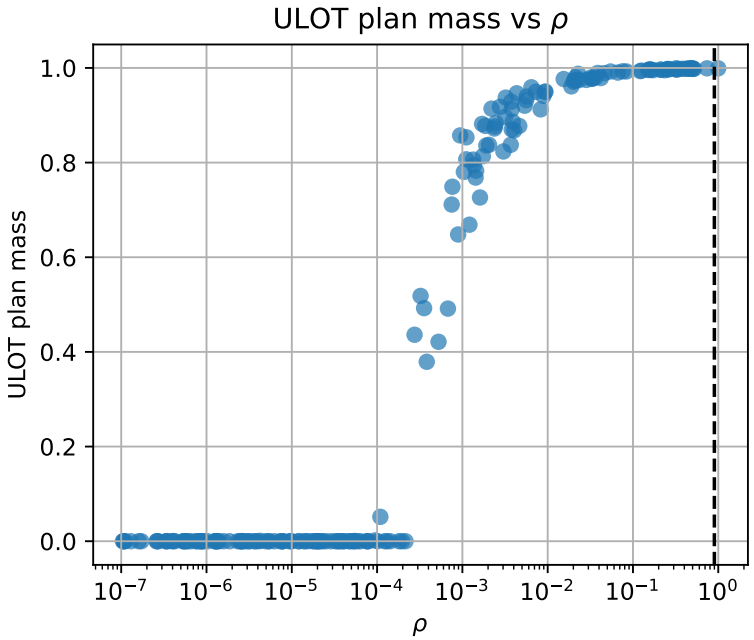


Figure 6: ULOT predicted plan mass for different values of ρ. The vertical line corresponds to $\rho = 0.9$ which is the value used for AGDL and produces balanced plans.

Algorithm 1 AGDL Dictionary Learning with ULOT-Based Unmixing

Require: Dataset $(G_i^{pos})_{i\in[N]}$ with $G_i^{pos} = (\boldsymbol{F}_i^{pos}, \boldsymbol{C}_i)$ and $\boldsymbol{F}_i^{pos} = [\boldsymbol{F}_i \mid \boldsymbol{Z}_i]$; feature projection $\pi_F([\boldsymbol{F} \mid \boldsymbol{Z}]) = \boldsymbol{F}$; initial models $\mathrm{MLP}_{\psi^{(0)}}$ and $\mathrm{MLP}_{\phi^{(0)}}$; batch size B; unmixing tolerance ε; unmixing step size η_ω; dictionary step size η_D; unmixing iterations T; outer iterations S

for $s = 0, \dots, S-1$ **do**

Sample minibatch indices $\mathcal{I}^{(s)} \subset [N]$ with $|\mathcal{I}^{(s)}| = B$

Sample $\alpha^{(s)} \in [0,1]$ uniformly

Compute the dictionary atoms at $\alpha^{(s)}$:

$$\widehat{\alpha}^{(s)} \leftarrow \mathrm{softbin}(\alpha^{(s)})$$

$$\left\{\boldsymbol{F}_k^{D,(s)}(\alpha^{(s)})\right\}_{k=1}^{K} \leftarrow \mathrm{MLP}_{\psi^{(s)}}(\widehat{\alpha}^{(s)}),$$
$$\left\{\boldsymbol{Z}_k^{D,(s)}(\alpha^{(s)})\right\}_{k=1}^{K} \leftarrow \mathrm{MLP}_{\phi^{(s)}}(\widehat{\alpha}^{(s)}).$$
$$\boldsymbol{F}_k^{D,pos,(s)}(\alpha^{(s)}) \leftarrow \left[\boldsymbol{F}_k^{D,(s)}(\alpha^{(s)}) \mid \boldsymbol{Z}_k^{D,(s)}(\alpha^{(s)})\right], \qquad k \in [K]$$
$$\left[\boldsymbol{C}_k^{D,(s)}(\alpha^{(s)})\right]_{uv} \leftarrow d_{\mathrm{geo}}\Big(\Pi_{\mathbb{S}^2}\left(\boldsymbol{Z}_k^{D,(s)}(\alpha^{(s)})[u]\right), \Pi_{\mathbb{S}^2}\left(\boldsymbol{Z}_k^{D,(s)}(\alpha^{(s)})[v]\right)\Big), \quad {\scriptstyle k\in[K],\ u,v\in[n]}.$$

Estimate the unmixing weights:

$$\left(\boldsymbol{\omega}_i^{(s)}\right)_{i\in\mathcal{I}^{(s)}} \leftarrow \textsc{Unmixing}\Big(\left(\boldsymbol{F}_k^{D,pos,(s)}(\alpha^{(s)}), \boldsymbol{C}_k^{D,(s)}(\alpha^{(s)})\right)_{k\in[K]}, (G_i^{pos})_{i\in\mathcal{I}^{(s)}}, \alpha^{(s)}, \varepsilon, \eta_\omega, T\Big).$$

for $i \in \mathcal{I}^{(s)}$ **do**

Reconstruct the concatenated node features:

$$\boldsymbol{F}^{pos,(s)}\left(\boldsymbol{\omega}_i^{(s)}; \alpha^{(s)}\right) \leftarrow \sum_{k=1}^{K} \omega_{i,k}^{(s)} \boldsymbol{F}_k^{D,pos,(s)}(\alpha^{(s)})$$

Reconstruct the pairwise distance matrix:

$$\boldsymbol{C}^{(s)}\left(\boldsymbol{\omega}_i^{(s)}; \alpha^{(s)}\right) \leftarrow \sum_{k=1}^{K} \omega_{i,k}^{(s)} \boldsymbol{C}_k^{D,(s)}(\alpha^{(s)})$$

Predict the transport plan:

$$\boldsymbol{P}_i^{(s)} \leftarrow \boldsymbol{P}_\theta^{\alpha^{(s)}}\Big((\boldsymbol{F}_i^{pos}, \boldsymbol{C}_i), \left(\boldsymbol{F}^{pos,(s)}(\boldsymbol{\omega}_i^{(s)}; \alpha^{(s)}), \boldsymbol{C}^{(s)}(\boldsymbol{\omega}_i^{(s)}; \alpha^{(s)})\right)\Big).$$

end for

Batch dictionary loss:

$$\mathcal{J}_{\mathcal{I}^{(s)}} \leftarrow \frac{1}{B} \sum_{i\in\mathcal{I}^{(s)}} \mathrm{L}^{\alpha^{(s)}}\Big((\pi_F(\boldsymbol{F}_i^{pos}), \boldsymbol{C}_i), \left(\pi_F\left(\boldsymbol{F}^{pos,(s)}(\boldsymbol{\omega}_i^{(s)}; \alpha^{(s)})\right), \boldsymbol{C}^{(s)}(\boldsymbol{\omega}_i^{(s)}; \alpha^{(s)})\right), \boldsymbol{P}_i^{(s)}\Big).$$

Update the MLP parameters:

$$\psi^{(s+1)} \leftarrow \psi^{(s)} - \eta_D \nabla_\psi \mathcal{J}_{\mathcal{I}^{(s)}},$$
$$\phi^{(s+1)} \leftarrow \phi^{(s)} - \eta_D \nabla_\phi \mathcal{J}_{\mathcal{I}^{(s)}}.$$

end for

return $\mathrm{MLP}_{\psi^{(S)}}, \mathrm{MLP}_{\phi^{(S)}}$

Algorithm 2 Unmixing

Require: Dictionary atoms $(\boldsymbol{F}_k^{D,pos}, \boldsymbol{C}_k^D)_{k\in[K]}$ with $\boldsymbol{F}_k^{D,pos} = [\boldsymbol{F}_k^D \mid \boldsymbol{Z}_k^D]$, minibatch $(G_i^{pos})_{i\in\mathcal{I}}$ with $G_i^{pos} = (\boldsymbol{F}_i^{pos}, \boldsymbol{C}_i)$ and $\boldsymbol{F}_i^{pos} = [\boldsymbol{F}_i \mid \boldsymbol{Z}_i]$, FGW parameter α, tolerance ε, step size η, and maximum number of iterations T

Initialize $\boldsymbol{z}^{(i,0)} \leftarrow 0_K$ for all $i \in \mathcal{I}$
$t \leftarrow 0$, $\Delta \leftarrow +\infty$
while $\Delta > \varepsilon$ **and** $t < T$ **do**
  for $i \in \mathcal{I}$ **do**
    $\boldsymbol{\omega}_i^{(t)} \leftarrow \mathrm{Softmax}(\boldsymbol{z}^{(i,t)})$
    Reconstruction:

$$\boldsymbol{F}^{pos}\left(\boldsymbol{\omega}_i^{(t)}\right) \leftarrow \sum_{k=1}^{K} \omega_{i,k}^{(t)} \boldsymbol{F}_k^{D,pos}$$

$$\boldsymbol{C}\left(\boldsymbol{\omega}_i^{(t)}\right) \leftarrow \sum_{k=1}^{K} \omega_{i,k}^{(t)} \boldsymbol{C}_k^{D}$$

    Predicted transport plan:

$$\boldsymbol{P}_i^{(t)} \leftarrow \boldsymbol{P}_\theta^{\alpha}\left((\boldsymbol{F}_i^{pos}, \boldsymbol{C}_i), \left(\boldsymbol{F}^{pos}(\boldsymbol{\omega}_i^{(t)}), \boldsymbol{C}(\boldsymbol{\omega}_i^{(t)})\right)\right)$$

  end for
  Batch loss:

$$\mathcal{L}_{\mathcal{I}}^{(t)} \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \mathrm{L}^{\alpha}\left((\pi_F(\boldsymbol{F}_i^{pos}), \boldsymbol{C}_i), \left(\pi_F\left(\boldsymbol{F}^{pos}(\boldsymbol{\omega}_i^{(t)})\right), \boldsymbol{C}(\boldsymbol{\omega}_i^{(t)})\right), \boldsymbol{P}_i^{(t)}\right)$$

  Gradient step:

$$\left(\boldsymbol{z}^{(i,t+1)}\right)_{i\in\mathcal{I}} \leftarrow \left(\boldsymbol{z}^{(i,t)}\right)_{i\in\mathcal{I}} - \eta \nabla_{(\boldsymbol{z}^{(i)})_{i\in\mathcal{I}}} \mathcal{L}_{\mathcal{I}}^{(t)}$$

$$\Delta \leftarrow \left(\frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \left\|\boldsymbol{z}^{(i,t+1)} - \boldsymbol{z}^{(i,t)}\right\|_2^2\right)^{1/2}$$

  $t \leftarrow t + 1$
end while
return $\left(\mathrm{Softmax}(\boldsymbol{z}^{(i,t)})\right)_{i\in\mathcal{I}}$