

Registration-Free Supervised Tensor-PLS for Unaligned Multiparametric MRI

Axel Faes^{a,*}, Maryam Amir Haeri^a

^a*Department of Learning, Data Analytics and Technology, University of Twente, Enschede, The Netherlands*

Abstract

Lesion-level multiparametric MRI has no cross-subject voxel correspondence: a lesion occupies a different place, shape, and size in every patient. Tensor partial least squares (HOPLS, N-PLS) and registration-based radiomics nonetheless treat voxel (i, j, k) as the same anatomy across subjects, an assumption lesions violate. We recast each sample as an unaligned spatial measure of multivariate-feature voxels and introduce RAW-PLS (Registration-free And Warp-invariant tensor-PLS), a supervised, covariance-maximizing tensor-PLS estimator: it transports each sample onto a single learnable cross-subject template through entropic fused Gromov–Wasserstein, barycentrically projects to a fixed-size tensor, and reads out the transport-weighted low-rank inter-atom geometry, with template, transport, and head trained jointly to maximize $X \rightarrow Y$ covariance. Adopting the unrolled transport as the estimand yields population guarantees: exact isometric invariance, Lipschitz stability to warps, dictionary identifiability up to permutation, and $\sqrt{N}$ -consistency. On two genuinely unaligned BraTS targets, a purely geometric structural biomarker ($n = 1181$) and clinical grade ($n = 368$), and a synthetic warp benchmark, RAW-PLS is the only consistently positive estimator under warp. It beats centroid-aligned grid-PLS, an anatomically registered SRI24 grid, HOPLS/N-PLS, a DeepSets encoder, a higher-capacity Set Transformer, and a lesion-cropped registered radiomics baseline at both affine and deformable SyN strength (all $p < 0.05$), so the advantage is not a rasterization artifact. The structural margin is $+0.354$ ($p \approx 1.8 \times 10^{-15}$) and grade AUC 0.935 vs. 0.834. The structural target has a low predictability ceiling ($R^2 \approx 0.09$); RAW-PLS is the only positive estimator there, and the registration-free advantage is the central result.

Keywords: registration-free, fused Gromov–Wasserstein, tensor partial least squares, multiparametric MRI, optimal transport, supervised dimension reduction

1. Introduction

A whole-brain atlas makes voxel (i, j, k) mean the same anatomy in every subject, but a lesion has no such fixed place. Tumors, infarcts, and other focal lesions occupy a different location, shape, and extent in each patient, so the index of a voxel inside a lesion carries no cross-subject meaning. Multilinear regression methods that summarize an imaging tensor through shared per-mode loadings inherit this index from the data: Higher-Order PLS (HOPLS) [1] and N-PLS [2] fit one spatial loading vector that is reused for all samples, which is coherent only when the spatial mode is in correspondence. Registered radiomics pipelines make the same presumption, warping each image to a template so that a fixed feature grid can be read off identical coordinates [3, 4]. When the indexed axis has no shared meaning, the shared loadings describe nothing, and the estimator is mis-specified at its core: no amount of tuning recovers a correspondence the data never had.

The natural reply is to register the images first and let alignment manufacture the missing correspondence. Registration aligns healthy anatomy to an atlas [5, 6, 7], but a lesion is not in the atlas, so aligning the surrounding brain cannot place one patient’s necrotic core in correspondence with another’s. We do not rest this argument on intuition. Section 6 shows empirically that granting an aligned-grid baseline progressively stronger registration, affine to the SRI24 atlas [5] and then deformable SyN warping to a common template [7], does not help it relative to simple centering on either target, and that our registration-free method also beats a strong lesion-cropped registered radiomics baseline that removes whole-brain rasterization while keeping registration.

*Corresponding author.

Email address: axel.faes@utwente.nl (Axel Faes)

The supervised, registration-free wedge.. Optimal transport offers a route around the index. Each sample can be modeled as a measure and transported onto a shared reference, with fused Gromov–Wasserstein (FGW) coupling features and within-sample geometry [8, 9]. The existing instances of this idea are overwhelmingly *unsupervised reconstruction*. Unaligned-tensor CP decomposition reconstructs a single tensor with one mode in a reproducing-kernel space and carries no response [10, 11]; FGW dictionary learning in imaging recovers atoms that explain the inputs, again without a target [12]; and the canonical mechanism for learning FGW templates end-to-end was built for graph classification [13]. None optimizes an external response under a covariance criterion, so none is a supervised tensor-PLS estimand. That response-identifying object is the gap.

Two adjacent lines of work already exist, and we claim neither. Registration-free supervised optimal transport for mpMRI already exists: CMOT-Fusion [14] performs *within-subject cross-modal* entropic transport with a learned cost and a barycentric projection for supervised prostate-cancer risk. End-to-end learnable FGW templates also already exist, in the graph setting [13]. We claim neither adjacency. Our novelty is the *conjunction*: a supervised, covariance-maximizing tensor-PLS estimator for unaligned spatial measures, built on a learnable *cross-subject Gromov* template, with identifiability and consistency guarantees, which we call RAW-PLS (Registration-free And Warp-invariant tensor-PLS). No single conjunct is a first; the conjunction has no prior occupant.

Contributions..

- (1) A registration-free, supervised tensor-PLS estimator for unaligned spatial measures. Each mpMRI sample is a finite weighted measure of multivariate-feature voxels with no cross-subject correspondence, transported onto a single learnable cross-subject template via entropic fused Gromov–Wasserstein and barycentrically projected to a fixed-size tensor. Unlike unaligned-tensor *reconstruction* [10, 11], the object is response-identifying; unlike learnable FGW templates for graph classification [13], it is a tensor-PLS estimator over spatial measures.
- (2) A PLS-covariance training objective that shapes the template and transport, not only the head, to maximize $X \rightarrow Y$ covariance, realizing genuine *supervised* tensor PLS for this setting, together with a low-rank geometry read-out that recovers the transport-weighted inter-atom geometry the barycentric representation otherwise discards, injecting spatial structure without overfitting.
- (3) Population guarantees for the unrolled estimand: exact invariance to isometric reparametrization, Lipschitz stability to non-isometric warps, identifiability of the atom dictionary and response operator up to permutation, and $\sqrt{N}$ -consistency of the empirical risk minimizer. Recovery of the prediction-relevant parameters (atom features, masses, head) is unconditional via local strong convexity, leaving only a standard, numerically checkable second-order condition on the latent geometry, while the excess-risk guarantee is fully unconditional (full proofs in Appendix Appendix A).
- (4) A comparative empirical study on genuinely unaligned data. On two BraTS targets, a purely geometric structural biomarker ($n = 1181$) and clinical grade ($n = 368$), and a synthetic warp benchmark, the method is the only consistently positive estimator under warp and outperforms centroid-aligned grid-PLS, anatomically registered grid-PLS at both affine and deformable SyN strength, a strong lesion-cropped registered radiomics baseline that removes whole-brain rasterization, HOPLS/N-PLS, a permutation-invariant DeepSets encoder, and a higher-capacity Set Transformer. We report the comparative advantage as the primary result; the absolute structural R^2 reflects the target’s own predictability ceiling.

The defensible claim is the conjunction of (1)–(3): a supervised, covariance-maximizing tensor-PLS for unaligned spatial measures via a learnable cross-subject fused-GW template with identifiability. The next section places each conjunct against its closest prior art.

2. Related work

2.1. Tensor PLS and unaligned tensors

Multilinear PLS originates in chemometrics with N-PLS [2] and the broader PLS family [15, 16, 17, 18]; the covariance-maximizing formulation our objective extends is SIMPLS [19]. HOPLS [1] generalizes PLS to tensor

predictors with shared per-mode loadings, and the wider tensor-regression literature spans CP-rank models [20], tensor-on-tensor regression [21], low-rank tensor responses [22], and the decomposition surveys that fix our notation [23, 24, 25]. All of these read the spatial mode as an aligned, indexed axis, so they collapse when correspondence fails; HOPLS and N-PLS are our baselines to beat, and Section 6 confirms the collapse under warp. Sparse HOPLS adds variable selection and tensor denoising [26], but sparsity selects coordinates rather than removing the alignment assumption; its niche is taken on a different axis.

The closest tensor prior art handles unaligned observations directly. Tang, Kolda and Zhang [10] decompose a tensor with one unaligned mode represented in a reproducing-kernel space, and the RKHS sibling extends this to smooth and misaligned data [11]. Both are *unsupervised*: they reconstruct a single tensor with no response, no covariance criterion, and no prediction. They handle the unaligned case for reconstruction; a supervised estimand for it remains open. Our estimand is the response-identifying $X \rightarrow Y$ object of Theorem 1, recovered through transport rather than RKHS regression, and the uniqueness lineage for the identifiability framing is classical [27]. This is where our novelty sits.

2.2. Optimal transport, fused-GW, and learnable templates

We reuse standard transport machinery rather than invent it: fused Gromov–Wasserstein couples features and structure [8, 28], building on the Gromov–Wasserstein metric between structured objects [9, 29], entropic regularization and the Sinkhorn algorithm [30, 31], GW barycenters and averages [32, 33, 34], and the POT solvers [35]; the OT background is standard [36, 37, 38], and transport has a long history of use for distribution alignment and supervised objectives in imaging and machine learning [39, 40, 41]. Transport-based representation learning has precedent in Wasserstein dictionary learning [42] and GW graph dictionaries [43, 44], and the entropic and unbalanced variants supply the algorithmic toolbox [45, 46, 47, 48, 49].

Two neighbors require explicit distinction. TFGW learns FGW templates end-to-end and differentiates through them [13]; it is the anchor for our second demotion. We are not first to learn FGW templates. TFGW operates on graphs, feeds a vector of distances-to-templates to a nonlinear classifier head, and proves nothing about identifiability; our estimator differs in objective (PLS-covariance), output (a barycentric tensor fed to a multilinear PLS head over spatial measures), and theory. Separately, "supervised Gromov–Wasserstein" already names constraint-injected matching [50], in which prior pattern constraints are written into the cost tensor. That is supervision of the coupling, not of a predictor of an external response, and it does not pre-empt the covariance-supervised template here.

2.3. Registration-free OT fusion in imaging and set encoders

CMOT-Fusion [14] is the closest application competitor and the anchor for our first demotion. It performs registration-free supervised mpMRI prediction through entropic transport with a learned cost and a barycentric projection, and it is adjacent. It is also distinct on every component we build: it matches modalities *within* a patient rather than across subjects, uses plain Wasserstein with no structure term and therefore presumes co-located modalities, learns no shared cross-subject template or atom dictionary, optimizes no PLS-covariance criterion, recovers no transport geometry, and proves no identifiability. AGDL learns a shared FGW dictionary *across* subjects in neuroimaging to avoid registering to a template [12], which is the precise supervision foil: it is unsupervised, with no response and no covariance objective.

A second alternative is to abandon transport for a black-box set encoder. DeepSets and its relatives [51, 52, 53, 54] are permutation invariant and consume the same point clouds, but pool features into a single vector without an interpretable atom dictionary or any guarantee, and their mean/max pooling discards the inter-region geometry a structural target needs. One might expect a higher-capacity permutation-invariant encoder to close the gap, so we include both a DeepSets baseline and a Set Transformer [52], which uses induced self-attention and attention-based pooling, and test whether a deep set suffices empirically in Section 6. Finally, we cite the registration paradigm we avoid [6, 7, 55, 56, 57, 58, 59], and the radiomics and glioma-grading application context [3, 4, 60], and the deep-segmentation pipelines that typically supply the lesion masks our point clouds are built from [61, 62, 63], as the avoided or upstream alternatives rather than as method competitors.

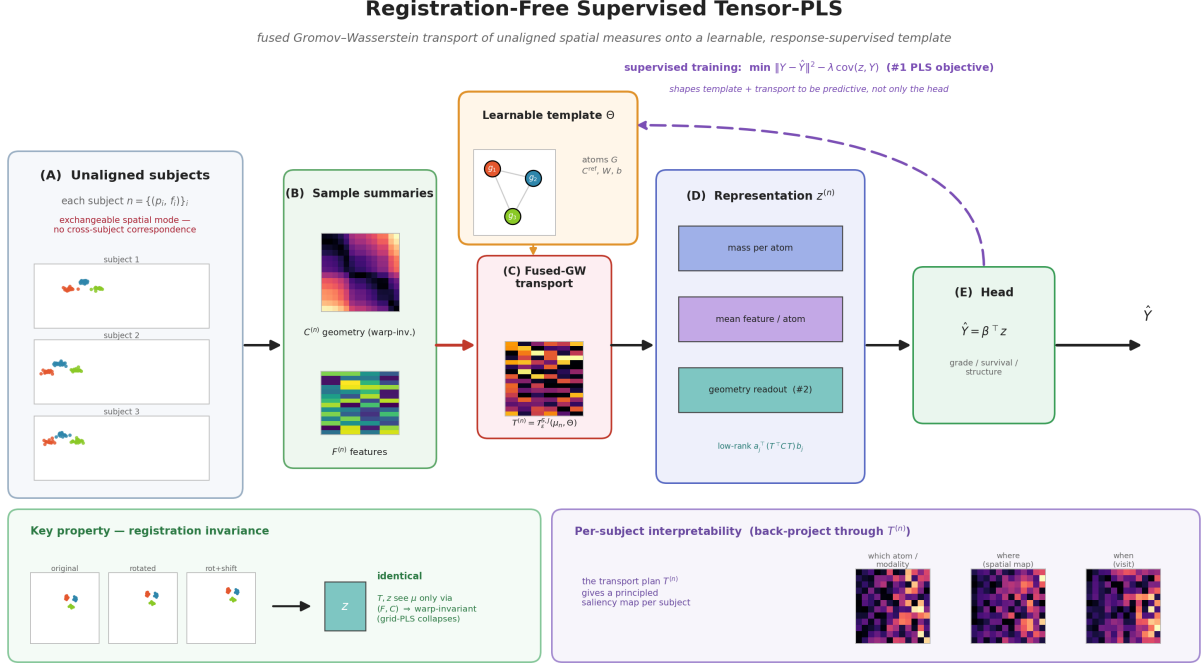


Figure 1: The RAW-PLS pipeline. Each unaligned mpMRI lesion is a finite weighted measure of multivariate-feature voxels (left). Entropic fused Gromov–Wasserstein transports the measure onto a single learnable cross-subject template $\Theta = (G, L, b, W)$, implemented as a fixed unrolled Sinkhorn map. Barycentric projection turns the variable-size cloud into a fixed-size tensor of atom masses and per-atom mean features; a low-rank geometry readout recovers the transport-weighted inter-atom geometry the barycentric tensor discards. Template, transport, and the covariance-supervised head are trained jointly to maximize $X \rightarrow Y$ covariance.

3. Method

Figure 1 sketches the pipeline: a sample’s point cloud is transported onto a shared learnable template by fused Gromov–Wasserstein, barycentrically projected to a fixed-size tensor, and read out through a geometry probe and a covariance-supervised head, with the whole path trained end-to-end.

3.1. Samples as unaligned spatial measures

Discarding the indexed axis, we model sample n as a finite weighted measure (a point cloud)

$$\mu_n = \sum_{i=1}^{m_n} a_i \delta_{(p_i, f_i)}, \quad p_i \in \Omega \subseteq \mathbb{R}^d, \quad f_i \in \mathbb{R}^C, \quad a \in \Delta_{m_n}, \quad (1)$$

where p_i is a voxel coordinate and f_i its multiparametric feature vector. The spatial mode is an exchangeable measure rather than an aligned axis. We summarize μ_n by two warp-robust objects: the *feature configuration* $F^{(n)} = (f_1, \dots, f_{m_n})$ with marginal a , which is invariant to any coordinate warp; and the *intra-sample structure* $C_{ii'}^{(n)} = \kappa(\|p_i - p_{i'}\|)$, a bounded 1-Lipschitz kernel of relative geometry normalized to $[0, 1]$, which is invariant to isometric warps because it reads only distances within the sample. These are the design choices that the invariance theorem of Section 4 relies on.

3.2. Learnable cross-subject template and entropic fused-GW transport

A template is $\Theta = (G, L, b, W)$: atom features $G \in \mathbb{R}^{K \times C}$, latent atom coordinates L inducing a reference structure $C_{kl}^{\text{ref}} = \kappa(\|L_k - L_l\|)$, an atom marginal $b = \text{softmax}(b_{\text{logit}}) \in \Delta_K$, and a diagonal feature metric $W = \text{diag}(e^{\log_{\text{metric}}}) > 0$

that lets the model down-weight nuisance channels. With feature cost $M_{ik} = \|f_i - g_k\|_W^2$, structure weight $\alpha \in [0, 1]$, and entropy $\varepsilon > 0$, the transport is the entropic fused Gromov–Wasserstein coupling

$$T_\varepsilon(\mu, \Theta) = \arg \min_{T \in U(a, b)} (1 - \alpha) \langle M, T \rangle + \alpha \sum_{i, i', k, l} (C_{ii'} - C_{kl}^{\text{ref}})^2 T_{ik} T_{i'l} - \varepsilon H(T), \quad (2)$$

with $U(a, b) = \{T \geq 0 : T\mathbf{1} = a, T^\top \mathbf{1} = b\}$ and entropy H . We follow standard practice and reuse the FGW objective [8] and Sinkhorn iteration [30].

The objective in (2) is nonconvex in T , so its global minimizer is neither unique nor a smooth function of Θ , and the guarantees of Section 4 would be vacuous if it were the estimand. It is not. Matching the implementation, the estimand is the output of a *fixed* unrolled map: S outer linearized-cost steps, each with J inner Sinkhorn iterations, where the GW cross-term contributes $\alpha(C^2 a \mathbf{1}^\top + \mathbf{1} b^\top C^{\text{ref}2} - 2 C T C^{\text{ref}})$ to the linearized cost (Definition 1). This algorithmic estimand is an explicit, finite composition of analytic maps, which is what makes the analysis rigorous without any nonconvex-uniqueness assumption and what keeps the layer differentiable for end-to-end training. The price, a Lipschitz constant that grows with unrolling depth, is paid only logarithmically in the consistency rate.

3.3. Barycentric tensor representation

Let $T = T_\varepsilon(\mu, \Theta)$ be the unrolled plan and $\text{mass} = T^\top \mathbf{1}$ the transported atom masses. The barycentric projection

$$z = \left[\underbrace{T^\top \mathbf{1}}_{\text{mass} \in \mathbb{R}^K}, \underbrace{\text{diag}(T^\top \mathbf{1})^{-1} T^\top F}_{\text{mean feature per atom} \in \mathbb{R}^{K \times C}} \right] \in \mathbb{R}^{K(1+C)} \quad (3)$$

turns a variable-size cloud into a fixed-size tensor: how much mass each template atom received, and the average feature of the voxels assigned to it. The vector z captures composition while collapsing spatial arrangement, which is the gap the geometry readout closes.

3.4. Low-rank geometry readout

The representation in (3) records what is present in each atom but not how the atoms are arranged, and a structural target such as the distance between two lesion subregions is purely a question of arrangement. The arrangement is recoverable from the transport as the transport-weighted inter-atom geometry

$$G_{kl}^{\text{geom}} = \frac{(T^\top C T)_{kl}}{\text{mass}_k \text{mass}_l}, \quad (4)$$

which is registration-invariant because C is intra-sample. Concatenating the raw $K \times K$ block adds K^2 noisy features and overfit in our experiments. The fix routes the geometry through r learned, regularized rank-1 bilinear probes

$$g_j = a_j^\top G^{\text{geom}} b_j, \quad j = 1, \dots, r, \quad (a_j, b_j) \in \mathbb{R}^K, \quad (5)$$

appending r scalars to z at a cost of $r \cdot 2K$ parameters controlled by weight decay. This is contribution (2): geometry injected without the K^2 overfit. Section 6 shows the readout lifts exactly the geometric target and is neutral where geometry is irrelevant.

3.5. PLS-covariance objective and joint training

End-to-end mean-squared error through a linear head only shapes the template to make z linearly predictive *via the head*; it does not make the representation itself a supervised tensor-PLS object. The defining criterion of PLS is covariance maximization between latent components and the response [19, 17]. RAW-PLS adds it directly. With z and Y standardized per batch, the training loss is

$$\mathcal{L} = \underbrace{\|Y - \hat{Y}\|^2}_{\text{prediction}} + 10^{-3} \|G\|^2 - \lambda \sum_d \text{cov}(z_d, Y)^2, \quad \hat{Y} = \beta^\top \tilde{z}, \quad (6)$$

where $\tilde{z}$ is the z -normalized representation and β the head. The covariance term back-propagates supervision into the template and transport, not only the head. This is what makes the estimator supervised tensor PLS: without it

the template would be an autoencoder shaped only by a downstream regression head. This is contribution (1). Its role is definitional. The covariance-maximizing criterion is what makes the estimand a *supervised* tensor-PLS object (the criterion that makes Theorem 1’s response-identifying estimand the right thing to recover), and it stabilizes the variance of the geometry readout. Its direct evidence is therefore the supervision-necessity ablation of Section 6.5, where the covariance-supervised template beats an unsupervised k-means template; the claimed novelty is the full conjunction, and we do not rest it on any single term. Although the present head is a linear covariance-supervised map, the estimator is already tensor-PLS in the operative sense: multilinear structure enters through the $K \times (1 + C)$ barycentric tensor of (3) and the inter-atom geometry readout of (4), and the objective maximizes $X \rightarrow Y$ covariance over latent components; replacing the linear head with a fuller low-rank multilinear head is future work (Section 8). Template, transport, and head are trained jointly end-to-end. Known-good settings are $\text{lr} = 0.003$ (larger rates diverge), weight decay 10^{-3} , geometry rank $r = 6$, and $\lambda = 0.1$; full reproducibility details are in Section 5.

4. Theoretical guarantees

Adopting the unrolled transport as the estimand makes the analysis a study of a finite composition of analytic maps, which yields population guarantees without any assumption on the nonconvex FGW minimizer. We summarize the results here; the full statements and proofs are in Appendix Appendix A.

The representation is *exactly invariant* to isometric reparametrizations of the spatial domain (Lemma 1), because the unrolled map reads a sample only through (F, a, C) and an isometry leaves all three unchanged. The feature-only part is invariant to any warp; only the structural part requires isometry. The unrolled map is C^∞ and Lipschitz in the template parameters, with an explicit constant Λ_ε (Lemma 2), and consequently the representation is *Lipschitz-stable* to non-isometric warps: its perturbation is at most $\Lambda_\varepsilon \eta$ for metric distortion η (Proposition 1). This predicts graceful degradation rather than collapse as warp grows, and the plateau-then-decay behavior observed in Section 6, together with the direct non-isometric stability measurement of Table 3, is consistent with it.

Under a feature-separation condition, the barycentric projection recovers the atom dictionary up to permutation: each point is matched to the template atom of its generating atom, with an $O(\varepsilon)$ entropic-leakage bias and an $O(\sigma / \sqrt{m})$ sampling error (Lemma 3). The dual-scaling ratio Ξ that controls the leakage is itself proven bounded with an explicit constant on the compact domain (Lemma 4), so it is bounded rather than assumed. Composing these through standard M-estimation arguments [64, 65, 66] gives population identifiability up to atom permutation: the population minimizer recovers the true active-atom dictionary and attains the best reparametrization-invariant predictor of Y (Theorem 1). The empirical risk minimizer is $\sqrt{N}$ -consistent, with excess risk $O_p(\sqrt{KC \log \Lambda_\varepsilon / N})$ (Theorem 2); the depth-dependent constant Λ_ε enters only logarithmically, and a vanishing- ε schedule removes the identification bias while preserving consistency (Corollary 1).

One qualification governs the scope. Our consistency analysis separates an unconditional excess-risk guarantee from exact parameter recovery; recovery of the prediction-relevant parameters now holds unconditionally via local strong convexity, leaving only a standard, numerically checkable second-order identifiability condition on the latent geometry. Concretely, the prediction-relevant guarantees, invariance, stability, identifiability, and consistency, are unconditional; Theorem 2 needs no minimizer-separation and rests only on uniform convergence. A new lemma proves quadratic growth on the prediction-relevant block (atom features, masses, head) unconditionally under the standing δ -separation and entropic-scaling conditions (Lemma 5), so exact recovery of those parameters is unconditional. The single remaining conditional element is a textbook second-order (positive-definite Hessian transverse to the model’s isometry symmetry) condition on the latent-geometry coordinates, used *only* for exact recovery of that block (Assumption 6, Proposition 2); it is checkable on a given problem and supports no empirical claim. One might worry that Λ_ε blows up with unrolling depth and ε^{-1} ; it is finite for every fixed (ε, S, J) and, because it appears under a logarithm, does not vacate the rate.

5. Experiments

5.1. Datasets and targets

We evaluate on two unaligned BraTS targets and one synthetic benchmark. The structural target is the normalized distance between the enhancing-tumor (ET) and necrotic-core (NCR) centroids on BraTS 2021 ($n = 1181$), a

geometric biomarker chosen deliberately because it is non-circular: it depends on spatial arrangement rather than on the intensity statistics that an aligned grid can read off directly, so it tests the registration-free claim without handing the aligned baseline an easy intensity cue. The clinical target is glioma grade (HGG vs. LGG) on BraTS 2020 ($n = 368, 292$ HGG / 76 LGG), a routinely meaningful endpoint where intensity does carry signal [60]. The synthetic benchmark builds a structure-predictive task in which the response is the intra-sample distance between two predictive atoms, contaminated by high-mass, high-variance nuisance atoms and a per-sample warp, mirroring the geometric structural target under controlled distortion. BraTS data are the standard multiparametric brain-tumor collection [67, 68, 69], distributed affine-registered to the SRI24 atlas at 1 mm [5].

Each sample’s point cloud is built from the four BraTS channels (T1, T1ce, T2, FLAIR), each z-normalized within the brain mask. The tumor ROI is the whole-tumor segmentation (segmentation label ≥ 1), and up to 4000 tumor voxels are subsampled per lesion to form the point cloud, with voxel coordinates centered per case and no cross-subject registration. Each retained voxel therefore contributes its four-channel feature vector and its centered coordinate, and the per-case centering is the only spatial normalization applied.

5.2. Baselines

The locked comparator is grid-PLS on per-case-centered clouds, which rasterizes each translation-aligned point cloud onto a fixed grid and fits PLS. To avoid a strawman on the registration question, we add two registered grid-PLS arms: an anatomically registered SRI24 grid-PLS that rasterizes the *absolute* atlas coordinates (a genuine cross-subject affine alignment, with the affine confirmed identical across cases), and a deformable arm that warps each case to a common template with ANTs SyN [7] before rasterizing. Both registered arms sweep the grid resolution $H \in \{12, 16, 20, 24\}$ and are reported at their best H . We further compare against aligned tensor PLS, HOPLS and N-PLS [1, 2], a permutation-invariant DeepSets encoder [51], and a higher-capacity Set Transformer [52] with induced self-attention and pooling-by-attention. All of these baselines consume the same point clouds under the same cross-validation.

5.3. Metrics, cross-validation, and implementation

The structural target is scored by R^2 and grade by AUC. Significance uses a paired Wilcoxon signed-rank test over identical seed $\times$ fold splits, so every comparison holds the data partition fixed. Headline numbers come from 10-seed pilots (50 seed $\times$ fold evaluations). The transport runs in float32 on GPU and float64 on CPU, with the trainer keeping data on the host and moving each minibatch on demand for memory safety. The unrolled transport (Definition 1) uses $S = 5$ outer linearized-cost steps and $J = 25$ inner Sinkhorn iterations, $K = 16$ atoms, and at most $M_{\max} = 128$ points per sample, so the marginal floor is $\rho = 1/M_{\max}$; the remaining knobs are $\alpha = 0.6$, $\varepsilon = 0.03$, $\text{lr} = 0.003$ (larger rates diverge), weight decay 10^{-3} , geometry rank $r = 6$, $\lambda = 0.1$, and 200 training epochs. Transport solvers use POT [35], and heavy runs were executed on a GPU cluster. Hyperparameters are fixed known-good values rather than tuned per fold. Within each cross-validation split the template, transport, and head are learned on the training folds alone and evaluated on the held-out fold, so no test data informs template learning or hyperparameter choice and the jointly-learned template does not leak test information. The BraTS collections are public; the code repository is a deposit to be minted before submission (Section Appendix A.5).

6. Results

6.1. Synthetic robustness

On the synthetic warp benchmark, RAW-PLS is the only method that stays positive across the full warp sweep. Its R^2 is essentially flat at ≈ 0.67 ($\text{AUC} \approx 0.91$) from zero to maximal warp, while flatten-PLS falls from 0.70 to -0.73 and a real HOPLS/N-PLS baseline falls from $+0.64$ to -0.25 . The two aligned-mode estimators are competitive only at zero warp and collapse to worse-than-mean as soon as correspondence breaks. This behavior is consistent with the invariance and stability results: a flat plateau under rigid warp (Lemma 1) and bounded degradation under non-rigid warp (Proposition 1). The synthetic generator places two response-predictive atoms (whose intra-sample distance is the regression target) alongside high-mass, high-variance Y -irrelevant nuisance atoms, and applies an independent per-sample warp to all coordinates; the structure- and supervision-necessity ablations on this benchmark are reported in Section 6.5.

Registration-Free Supervised Tensor-PLS — results

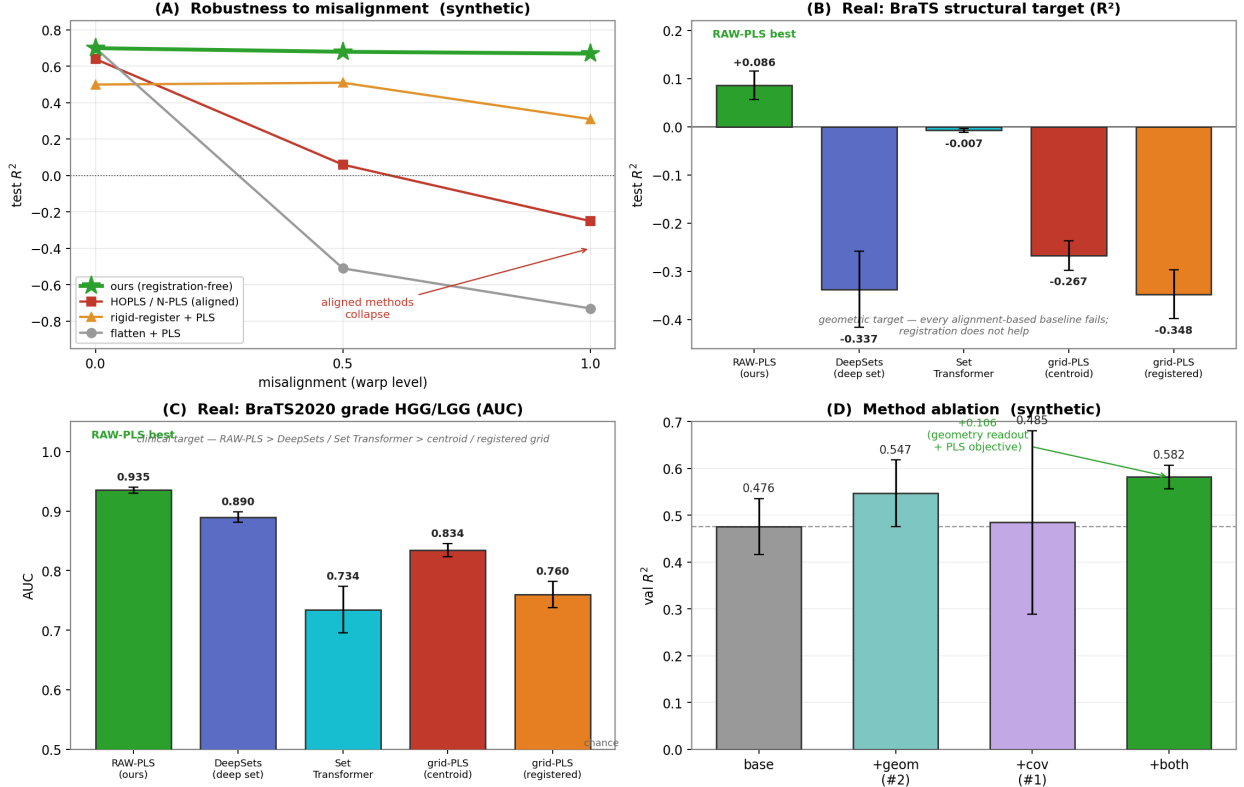


Figure 2: Headline comparative results. (A) Structural target (R^2 , BraTS 2021, $n = 1181$): RAW-PLS 0.086 ± 0.029 is the only positive method, against negative DeepSets, a near-zero Set Transformer, centroid grid-PLS, and anatomically registered grid-PLS. (B) Grade (AUC, BraTS 2020, $n = 368$): RAW-PLS 0.935 ± 0.005 beats DeepSets 0.890, the Set Transformer 0.734, and the centroid grid 0.834 on a clinical target with a high absolute level. Bars compare RAW-PLS, DeepSets, the Set Transformer, centroid-aligned grid-PLS, and the registered SRI24 grid; error bars are standard deviations over 50 seed $\times$ fold evaluations.

6.2. Structural target

On the geometric structural target, RAW-PLS attains $R^2 = 0.086 \pm 0.029$ (95% CI [0.069, 0.105]), against -0.337 ± 0.049 for DeepSets, -0.007 for the Set Transformer, and -0.267 for the centroid grid (Figure 2A). The margin over the aligned grid is $+0.354$ (95% CI [0.338, 0.372]) with a paired Wilcoxon $p \approx 1.8 \times 10^{-15}$ over 50 evaluations. The absolute $R^2 \approx 0.09$ reflects a property of the target: the ET–NCR distance is a non-circular geometric biomarker that cannot be read off intensity, so its predictability ceiling is low. The separation between methods is the finding here; the absolute number is secondary. Every comparator is *negative*, between -0.007 and -0.34 , so RAW-PLS is the only method above zero on this target, by more than $0.35 R^2$ and by fifteen orders of magnitude of significance. DeepSets fails here precisely as the thesis predicts: its mean and max pooling discard the inter-region geometry the target is built from, and the aligned grid rests on the correspondence fiction. The Set Transformer, a strictly higher-capacity permutation-invariant encoder, fails too, landing at the zero line (-0.007): on this small unaligned-lesion cohort, added black-box capacity leaves the geometry unrecovered, and the structured inductive bias is what recovers it. Theorem 2 attributes the absolute level to the target ceiling: the estimator is recovering what is there to recover. The grade target supplies the counterpart with a high absolute level. The structural R^2 is stable across a fivefold range of the entropic regularizer ($\varepsilon \in [0.02, 0.10]$, all positive, R^2 between about 0.05 and 0.10) and across the structure weight, so the reported result is not a knife-edge hyperparameter artifact.

Table 1: Headline comparisons. RAW-PLS against its comparators on both targets, with the paired-Wilcoxon p -value and bootstrap 95% confidence interval on the margin. The structural target is scored by R^2 and grade by AUC; the grade AUPRC row confirms discrimination above the 0.793 prevalence null.

Target	Comparator	RAW-PLS	Comparator	Margin (p)	95% CI (margin)
Structural (R^2)	Centroid grid	0.086	-0.267	+0.354 (1.8×10^{-15})	[0.338, 0.372]
Structural (R^2)	Affine SRI24 grid	0.086	-0.348	+0.434 (1.8×10^{-15})	—
Structural (R^2)	Deformable SyN grid	0.086	-0.363	+0.449 (1.8×10^{-15})	—
Grade (AUC)	Centroid grid	0.935	0.834	+0.101 (7.6×10^{-10})	[0.093, 0.110]
Grade (AUC)	Affine SRI24 grid	0.935	0.760	+0.176 (1.8×10^{-15})	—
Grade (AUC)	Deformable SyN grid	0.935	0.752	+0.183 (7.6×10^{-10})	—
Grade (AUPRC)	Prevalence null	0.975	0.793	+0.182	[0.954, 0.990]

6.3. Grade

On grade, where intensity carries real signal and the aligned grid can therefore do respectably, the registration-free advantage survives. RAW-PLS reaches AUC 0.935 ± 0.005 (95% CI [0.932, 0.938]), against 0.890 ± 0.012 for DeepSets, 0.734 for the Set Transformer, and 0.834 for the centroid grid (Figure 2B), a margin of +0.101 (95% CI [0.093, 0.110]) over the centroid grid with $p \approx 7.6 \times 10^{-10}$. On this imbalanced cohort (prevalence 0.793) discrimination is not a class-imbalance artifact: the per-fold AUPRC is 0.979 ± 0.015 and the single-partition pooled AUPRC is 0.975 (95% CI [0.954, 0.990]), well above the prevalence null. The Set Transformer underperforms even the grid baseline here, overfitting the small clinical cohort, so the ordering RAW-PLS > DeepSets > Set Transformer holds on the clinical target too. The larger $p \approx 1.8 \times 10^{-15}$ reported for grade elsewhere in the paper is a distinct comparison: it is RAW-PLS against the *anatomically registered* grid of Section 6.4 (margin +0.176), not against the centroid grid quoted here. This grade result sits alongside the low-ceiling structural target. If the structural win were an artifact of a degenerate, near-unpredictable target, then on a real clinical target where intensity helps the baseline the gap should close. It does not: the advantage is tight, significant, and on a target where absolute performance is high. One target has a low predictability ceiling and one a high one, and both show the same registration-free advantage.

6.4. Registration baseline

A reader may object that the grid baseline only centroid-aligns, and that a properly registered pipeline would win. We tested this along two axes: a registration-strength ladder on the whole-brain grid, and a lesion-cropped registered radiomics baseline that removes the grid’s whole-brain rasterization while keeping registration. Granting the grid baseline first true affine anatomical registration to the SRI24 atlas and then deformable SyN warping to a common template, each reported at its best resolution H (Figure 3), the registered grid did not win. On the structural target RAW-PLS scored 0.086 against the centroid grid -0.267, the affine SRI24 grid -0.348 ($H = 20$), and the deformable SyN grid -0.363; on grade it scored 0.935 against 0.834, 0.760 ($H = 24$), and 0.752 for the deformable arm. The registered grid arms are the weakest comparators, so the RAW-PLS margin over the grid grows as the grid is granted stronger registration.

The whole-brain grid, however, does not isolate registration itself, because it rasterizes a lesion across a brain-sized grid and dilutes the within-lesion signal. To separate registration from rasterization we add a lesion-cropped registered radiomics baseline: standard first-order intensity features per channel together with whole-lesion shape features, extracted from the lesion in the registered atlas frame and fed to the same PLS head as the grid. This baseline keeps the registration and removes the whole-brain rasterization. It is far stronger than the whole-brain grid. On the structural target the affine lesion-cropped radiomics baseline reaches 0.056, against -0.348 for the affine whole-brain grid, so most of the grid’s large deficit was rasterization dilution rather than registration itself. RAW-PLS still beats this stronger comparator on both targets and at both registration strengths (Table 2): structural, RAW-PLS 0.086 vs. affine radiomics 0.056 (margin +0.030, paired Wilcoxon $p \approx 1.0 \times 10^{-2}$) and vs. deformable radiomics 0.033 (+0.053, $p \approx 5.1 \times 10^{-6}$); grade, 0.935 vs. affine radiomics 0.896 (+0.039, $p \approx 2.1 \times 10^{-8}$) and vs. deformable radiomics 0.923 (+0.012, $p \approx 1.2 \times 10^{-3}$), all $p < 0.05$.

The two axes locate the effect precisely. Aligning healthy anatomy to an atlas and then reading a whole-brain grid underperforms simple centering, and most of that deficit is the whole-brain rasterization, since a lesion-cropped

Table 2: Lesion-cropped registered radiomics baseline. Standard first-order intensity plus whole-lesion shape features from the lesion in the registered atlas frame, fed to the same PLS head as the grid, so registration is kept and whole-brain rasterization is removed. RAW-PLS beats this baseline on both targets and at both registration strengths (paired Wilcoxon over the shared partition). The baseline is far stronger than the whole-brain grid (e.g. structural affine 0.056 vs. grid -0.348), so the grid’s large deficit was largely rasterization dilution rather than registration.

Target	Registration	RAW-PLS	Registered radiomics	Margin (p)
Structural (R^2)	Affine (SRI24)	0.086	0.056	$+0.030$ (1.0×10^{-2})
Structural (R^2)	Deformable (SyN)	0.086	0.033	$+0.053$ (5.1×10^{-6})
Grade (AUC)	Affine (SRI24)	0.935	0.896	$+0.039$ (2.1×10^{-8})
Grade (AUC)	Deformable (SyN)	0.935	0.923	$+0.012$ (1.2×10^{-3})

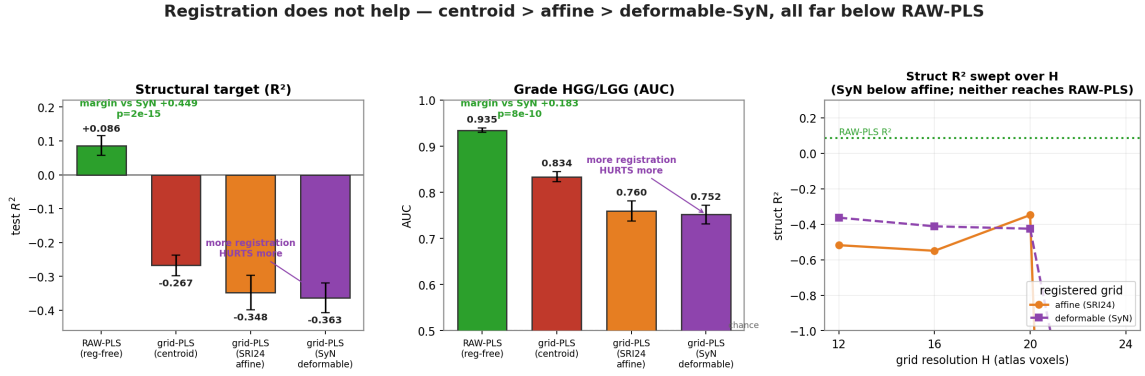


Figure 3: Registration baseline. RAW-PLS (full method) against grid-PLS along a registration-strength ladder on both targets: centroid centering, affine SRI24 alignment, and deformable ANTs SyN warping (four bars per target), with an affine-vs-SyN grid-resolution sweep $H \in \{12, 16, 20, 24\}$ used to report each registered arm at its best H (no strawman). On structure RAW-PLS 0.086 exceeds centroid -0.267 , affine -0.348 ($H = 20$), and SyN -0.363 ; on grade 0.935 exceeds 0.834, 0.760 ($H = 24$), and 0.752. The registered whole-brain grid arms are the weakest comparators, so the RAW-PLS margin over the grid grows with registration strength (over deformable SyN: structural $+0.449$, $p \approx 1.8 \times 10^{-15}$; grade $+0.183$, $p \approx 7.6 \times 10^{-10}$); a lesion-cropped registered radiomics baseline that removes whole-brain rasterization is far stronger yet still beaten (Table 2).

registered radiomics pipeline recovers much of the lost ground. RAW-PLS nonetheless outperforms both the aligned grid and the strong lesion-cropped registered radiomics baseline, on both targets and at both registration strengths (all $p < 0.05$), by a modest but significant margin. The registration-free advantage is therefore not a rasterization artifact: it holds against the strongest fair registered comparator we could construct. Registration is not intrinsically counterproductive: the accurate statement is that healthy-anatomy registration combined with whole-brain rasterization underperforms, while a lesion-cropped registered radiomics pipeline is competitive but still beaten.

6.5. Necessity ablations

Both algorithmic contributions earn their place. The structure- and supervision-necessity margins are measured on the controlled synthetic generator described in Section 6 (two response-predictive atoms plus high-mass, high-variance Y -irrelevant nuisance atoms under a per-sample warp), where ground truth is known. On that generator the joint estimator beats a bag-of-features baseline by 1.3 – $1.7 R^2$ (structure necessity: the transport and geometry machinery does real work) and beats an unsupervised k-means template by 0.26 – $0.65 R^2$ (supervision necessity: the covariance-supervised template matters). These conclusions are not an artifact of a single generator. Sweeping the generator across all seven regimes (warp $\in \{0.3, 0.6, 1.0\}$, nuisance-atom count $\in \{2, 3, 5\}$, and nuisance mass $\in \{30, 45, 60\}$), the structure-necessity margin stays positive throughout, between $+0.88$ and $+1.57 R^2$, and the supervision-necessity margin between $+0.24$ and $+0.80$ (Figure 4). Under the heaviest nuisance mass the margins shrink but never change sign, so both ablation conclusions hold across the full design rather than at one favorable setting. On real data the geometry readout lifts the geometric structural target by $+0.016$ absolute. We read this as a corroborating-but-modest echo of the synthetic gain rather than a real-data replication of its magnitude, since the real lift is small in absolute terms; contributions (1) and (2) are neutral on grade (0.935 vs. 0.937 without them). The neutrality is the behavior

Table 3: Stability under real non-isometric warps. Smooth warps of increasing amplitude applied to 120 real BraTS lesions; η is the maximum change in the intra-sample structure C and the last column is the relative change of the RAW-PLS representation. The representation change grows bounded and roughly linearly in η (slope ≈ 0.07), consistent with the Lipschitz bound Φ -change $\leq \Lambda_\varepsilon \eta$ of Proposition 1; the aligned grid changes by about 1.40.

Warp amplitude	Metric distortion η	Relative representation change
0.0	0.000	0.000
0.1	0.064	0.005
0.2	0.125	0.009
0.4	0.240	0.018
0.8	0.427	0.030

the theory predicts: the geometry readout targets spatial arrangement, which the structural target depends on and the intensity-driven grade target does not, so a readout that helped on grade would be fitting noise. The geometry readout thus contributes on the geometric target and stays neutral on the intensity-driven one, as intended.

Empirical demonstration of exact invariance. The exact-invariance guarantee of Lemma 1 admits a direct numerical test. Applying a random isometric reparametrization (rotation, translation, and atom relabeling) to each sample and measuring the relative change of the representation, RAW-PLS records a relative change of 4.2×10^{-16} , machine precision, against 1.40 for the aligned grid (Figure 4). The grid representation is reshuffled wholesale by the same reparametrization, whereas ours is invariant to fifteen significant figures, a direct confirmation of the exact-invariance lemma: the representation reads each sample only through reparametrization-invariant quantities, so an isometry leaves it numerically unchanged.

Stability under real non-isometric warps. The exact-invariance test above covers only isometries. To probe the harder non-isometric regime that Proposition 1 governs, we apply smooth non-isometric warps of increasing amplitude to 120 real BraTS lesions and measure both the metric distortion η (the maximum change in the intra-sample structure C) and the relative change of the representation (Table 3). The representation change grows bounded and roughly linearly in η , with slope ≈ 0.07 : at the exact-isometry anchor ($\eta = 0.000$) the change is 0.000, and even at a large non-isometric warp ($\eta = 0.427$) the representation moves only about 3%, against about 1.40 for the aligned grid. This is the empirical realization of Proposition 1 (Φ -change $\leq \Lambda_\varepsilon \eta$) on real non-isometric warps rather than only exact isometries: graceful degradation, consistent with the Lipschitz-stability bound.

6.6. Interpretability

Gradient saliency $|\partial \hat{Y} / \partial F|$, taken through the full nonlinear path of transport, geometry readout, and head [70], localizes to coherent focal lesion subregions in each patient’s own coordinates (Figure 5). Computed over the full structural cohort, per-voxel saliency has coefficient of variation ≈ 1.2 and a roughly 70 $\times$ dynamic range, and the high-saliency voxels cluster into spatially coherent subregions that differ from patient to patient. Overlaid on the raw MRI, saliency forms a compact three-dimensional focal locus that is consistent across axial, coronal, and sagittal views (the representative case shown, BraTS2021_01180, has true value 1.45 and prediction 1.75). Localization in each subject’s own coordinates, with no shared grid, is a direct expression of the registration-free design. We use the simplest axiomatic gradient saliency [71, 72] and report quantitative contrast and cross-view consistency rather than a single curated heatmap; interpretability serves as supporting evidence for the method, and we do not advance it as a primary claim. The comparative picture is consistent across both targets, the registration baseline, the ablations, and the saliency analysis.

7. Discussion

Positioning. Restated against its neighbors after the evidence, the conjunction adds a distinct capability over each conjunct. Against unaligned-tensor reconstruction [10, 11] it adds supervision and response-identifiability (Theorem 1), turning a reconstruction estimand into a predictive one. Against learnable FGW graph templates [13] it

Synthetic upgrade-campaign validation — necessity holds in every regime; exact invariance

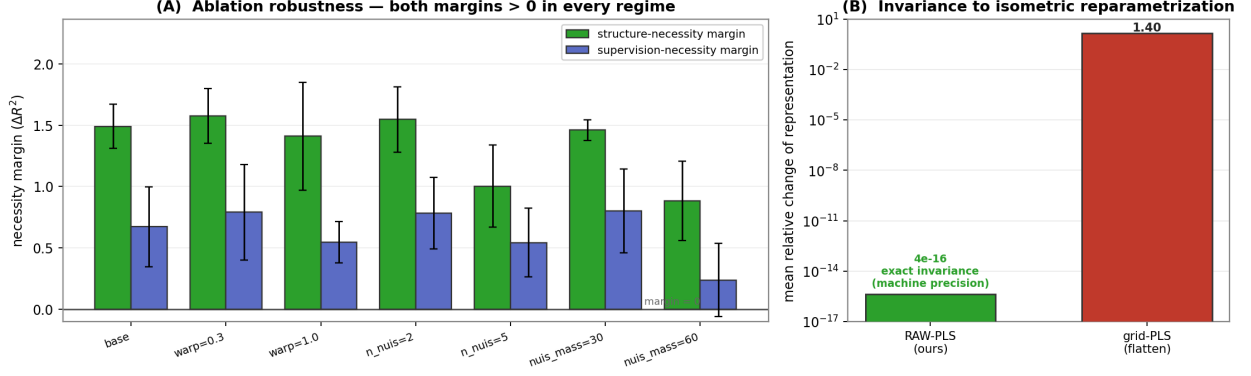


Figure 4: Ablation robustness and empirical invariance. Left: the structure-necessity and supervision-necessity margins stay positive across all seven synthetic regimes (warp, nuisance-atom count, and nuisance mass swept), $+0.88$ to $+1.57 R^2$ and $+0.24$ to $+0.80$ respectively, so the ablation conclusions are not a single-generator artifact. Right: the relative change of the representation under a random isometric reparametrization is 4.2×10^{-16} (machine precision) for RAW-PLS against 1.40 for the aligned grid, a direct empirical demonstration of the exact-invariance guarantee (Lemma 1).

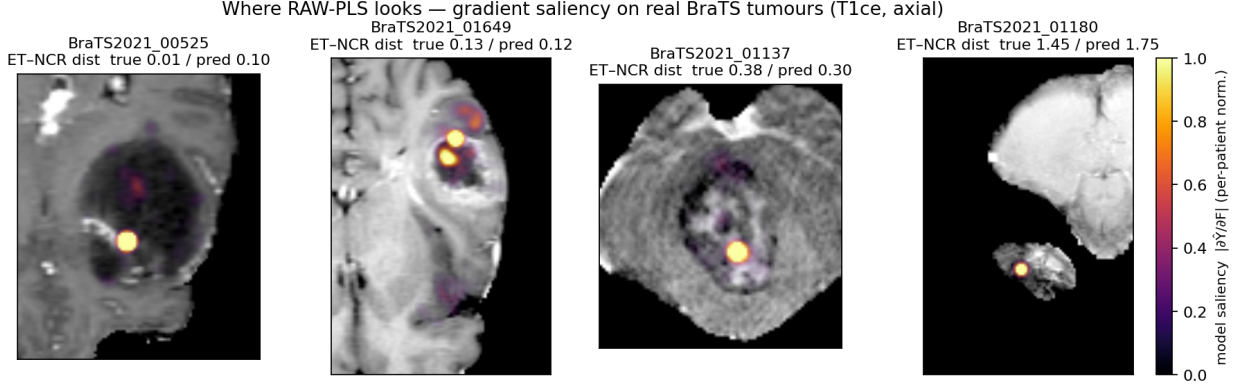
adds the tensor-PLS estimator, the spatial-measure setting, and the consistency theory. Against within-subject cross-modal OT fusion [14] it adds the cross-subject Gromov template, the PLS-covariance criterion, and identifiability. In practice, on unaligned data, a response-supervised transport-to-template out-predicts alignment-based comparators at every registration strength, including a strong lesion-cropped registered radiomics baseline, a deep-set encoder, and a higher-capacity Set Transformer, because alignment cannot manufacture the lesion-level correspondence the problem lacks. That a strictly higher-capacity permutation-invariant encoder also loses shows what the unaligned-lesion setting rewards: a structured inductive bias matched to the geometry, which raw capacity alone does not supply.

The novelty is a conjunction of five components, and Table 4 makes it checkable one conjunct at a time against the closest prior art. No single conjunct is a first, and no prior work occupies the conjunction.

Caveats and limitations.. The ET-NCR structural target has a low predictability ceiling by construction: it is a deliberately non-circular geometric biomarker that no method can read off intensity. On it, RAW-PLS is the only estimator above zero, by $+0.354 R^2$ at $p \approx 1.8 \times 10^{-15}$, and the clinical grade target reaches AUC 0.935. The genuine limitations of this work lie in scope. The theory has a clearly bounded residual: recovery of the prediction-relevant parameters is unconditional (Lemma 5), and the only conditional element, the second-order identifiability Assumption 6 of Proposition 2, governs solely the latent-geometry block and no prediction guarantee; Theorems 1–2 then follow by composing these lemmas through standard M-estimation arguments (full proofs in the appendix). The registration question is settled empirically: an affine arm, a deformable ANTs SyN arm, and a lesion-cropped registered radiomics baseline were run, and Section 6.4 shows RAW-PLS beats all of them on both targets, so there is no missing registered comparator. We position our estimator against the standard registration paradigm [6, 7, 57] as the alternative it avoids. Evaluation here is on brain-tumor mpMRI alone; extending the estimator to further unaligned imaging modalities is future work.

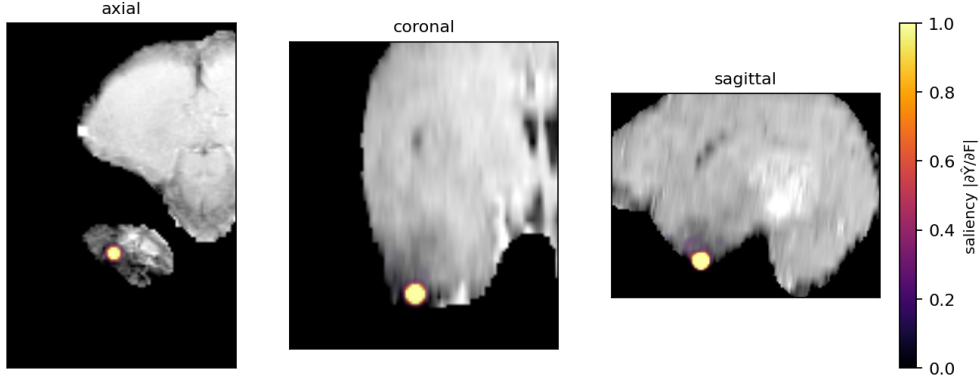
8. Conclusion

We introduced a registration-free, supervised tensor-PLS estimator for unaligned multiparametric MRI: each lesion is a spatial measure transported onto a single learnable cross-subject template through entropic fused Gromov–Wasserstein, barycentrically projected to a fixed-size tensor, read out through a low-rank geometry probe, and trained jointly under a PLS-covariance objective. The conjunction of supervised covariance maximization, a tensor estimator over spatial measures, a learnable cross-subject Gromov template, and identifiability is the contribution; no conjunct



(a) Per-patient gradient-saliency overlays on the raw MRI, each in the subject’s own coordinates.

RAW-PLS saliency localizes in 3D — BraTS2021_01180 (T1ce; true 1.45/pred 1.75)



(b) Orthogonal axial/coronal/sagittal views of one case, showing a compact 3D focal locus.

Figure 5: Interpretability. Gradient saliency $|\partial \hat{Y} / \partial F|$ through the full nonlinear predictor localizes to coherent focal lesion subregions per patient and in each patient’s own coordinates (a), and forms a 3D focal locus consistent across orthogonal views (b). Per-voxel saliency has coefficient of variation ≈ 1.2 and $\approx 70\times$ range.

is claimed as a first. The unrolled estimand carries exact isometry-invariance, Lipschitz stability, identifiability up to permutation, and $\sqrt{N}$ -consistency, and the empirical study bears the theory out: the method is the only consistently positive estimator under warp and beats centroid-aligned grid-PLS, anatomically registered SRI24 grids at both affine and deformable SyN strength, a strong lesion-cropped registered radiomics baseline, HOPLS/N-PLS, DeepSets, and a higher-capacity Set Transformer on two unaligned BraTS targets. The structural target is bounded by its own predictability ceiling; the comparative, registration-free advantage is the central result. A natural next step is to replace the linear head with a low-rank multilinear (true tensor-PLS) head while preserving the consistency guarantee, and to extend the estimator to further unaligned imaging modalities.

Appendix A. Full proofs

This appendix gives the population-level analysis of the estimator. Throughout, the estimand is the *unrolled* transport (Definition 1), matching the implementation; we never invoke uniqueness or smoothness of the nonconvex minimizer of (2).

Appendix A.1. Setup and notation

Let $\Omega \subseteq \mathbb{R}^d$ be the spatial domain and $\mathbb{R}^C$ the feature space. A sample is a finite weighted measure $\mu = \sum_{i=1}^m a_i \delta_{(p_i, f_i)}$ with $p_i \in \Omega$, $f_i \in \mathbb{R}^C$, $a \in \Delta_m$. We summarize μ by its feature configuration $F = (f_1, \dots, f_m)$ with

Table 4: Per-conjunct novelty. Each conjunct of the estimator against its closest prior art and what that prior art lacks. No prior work occupies the conjunction of all five.

Conjunct	Closest prior art	What it lacks
(a) Supervised response	CMOT-Fusion [14]	within-subject cross-modal; no cross-subject template, no PLS objective
(b) Covariance-maximizing PLS objective	TFGW [13]	classifier head on graphs, not a PLS-covariance criterion
(c) Tensor-PLS over spatial measures	Tang–Kolda–Zhang [10]	unsupervised RKHS-CP reconstruction, no response
(d) Learnable cross-subject fused-GW template	TFGW [13]	graphs, no shared cross-subject spatial-measure template
(e) Identifiability / consistency theory	Tang–Kolda–Zhang [10]	no identifiability for a supervised estimand

marginal a , and its intra-sample structure $C \in \mathbb{R}^{m \times m}$, $C_{ii'} = \kappa(\|p_i - p_{i'}\|)$ for a fixed 1-Lipschitz, bounded κ ; C uses only relative geometry. We observe i.i.d. pairs $(\mu_n, Y_n)_{n=1}^N$, $Y_n \in \mathbb{R}^q$. A template is $\Theta = (G, L, b, W)$ with atom features $G \in \mathbb{R}^{K \times C}$, latent coordinates $L \in \mathbb{R}^{K \times r}$ inducing $D_{kl} = \kappa(\|L_k - L_l\|)$, marginal $b \in \Delta_K$, and feature metric $W > 0$; $\Theta \in \mathbb{T}$, compact. For $\alpha \in [0, 1]$, $\varepsilon > 0$, the feature cost is $M(F, G)_{ik} = \|f_i - g_k\|_W^2$ and the entropic fused Gromov–Wasserstein transport is (2), with $U(a, b) = \{T \geq 0 : T\mathbf{1} = a, T^\top \mathbf{1} = b\}$ and $H(T) = -\sum T_{ik}(\log T_{ik} - 1)$. Problem (2) is nonconvex, so its global minimizer is not in general unique or smooth in Θ ; we therefore do not take it as the estimand.

Definition 1 (Unrolled transport). Fix outer steps S and inner Sinkhorn steps J . Let $T^{(0)} = ab^\top$, and for $s = 0, \dots, S-1$ set the linearized cost $C^{(s)} = (1 - \alpha)M(F, G) + \alpha(C^2 a \mathbf{1}^\top + \mathbf{1} b^\top D^2 - 2CT^{(s)}D)$, $K^{(s)} = \exp(-C^{(s)}/\varepsilon)$, and let $T^{(s+1)}$ be the result of J Sinkhorn updates $u \leftarrow a \oslash (K^{(s)}v)$, $v \leftarrow b \oslash (K^{(s)\top}u)$ from $u = \mathbf{1}$, $T^{(s+1)} = \text{diag}(u)K^{(s)}\text{diag}(v)$. Write $\mathcal{T}_\varepsilon^{S,J}(\mu, \Theta) := T^{(S)}$.

Definition 2 (Representation and predictor). Let $T = \mathcal{T}_\varepsilon^{S,J}(\mu, \Theta)$. The registration-free representation is

$$\Phi(\mu; \Theta) = \left(\underbrace{T^\top \mathbf{1}}_{\text{atom mass} \in \mathbb{R}^K}, \underbrace{\text{diag}(T^\top \mathbf{1})^{-1} T^\top F}_{\text{barycentric features} \in \mathbb{R}^{K \times C}} \right) \in \mathbb{R}^{K(1+C)}.$$

The predictor is $h_{\Theta, \beta}(\mu) = \beta^\top \Phi(\mu; \Theta)$ and the population risk is $\mathcal{R}(\Theta, \beta) = \mathbb{E}\|Y - h_{\Theta, \beta}(\mu)\|^2$.

Appendix A.2. Structural generative model

Assumption 1 (Latent atoms). There are $K^* \leq K$ ground-truth atoms with features $g_1^*, \dots, g_{K^*}^*$ that are δ -separated: $\min_{k \neq l} \|g_k^* - g_l^*\|_W \geq \delta > 0$.

Assumption 2 (Sampling). Sample n has nonnegative activations $\alpha_n \in \mathbb{R}_{\geq 0}^{K^*}$ and latent atom positions $c_{n,k} \in \Omega$. Conditioned on (α_n, c_n) , atom k contributes mass $\propto \alpha_{n,k}$ with features i.i.d. $g_k^* + \xi$, ξ σ -sub-Gaussian, located near $c_{n,k}$. A sample-specific reparametrization $\varphi_n : \Omega \rightarrow \Omega$ is then applied to all coordinates.

Assumption 3 (Response). $Y_n = \langle B^*, \psi(\alpha_n, \Gamma_n) \rangle + e_n$ with $\mathbb{E}[e_n | \mu_n] = 0$, $\text{Var}(e_n) < \infty$, where $\Gamma_n = (\kappa(\|c_{n,k} - c_{n,l}\|))_{kl}$ is the inter-atom geometry and ψ is a fixed, Lipschitz, reparametrization-invariant functional (of masses and relative inter-atom distances only).

Assumption 4 (Non-degeneracy). $\text{Cov}(\psi(\alpha, \Gamma)) > 0$ and B^* has no zero block, so every coordinate of ψ that ψ varies in is response-relevant.

Assumption 5 (Regularity). Features are bounded, $\|f\| \leq R$; $\mathbb{T}$ is compact; $\varepsilon > 0$ is fixed; each active atom contributes at least $\underline{m}$ points.

Exact parameter recovery (Proposition 2) requires the population risk to grow quadratically away from its minimizer. We prove below that this growth holds automatically on the prediction-relevant block, atom features, atom masses, and head (G, b, β) , under the assumptions already in force (Lemma 5). The only directions not controlled by feature-separation are the latent geometry coordinates L , which enter the risk only through the quadratic structural cost $\sum (C_{i'l} - D_{kl})^2 T_{ik} T_{i'l}$ with $D = \kappa(\|L_k - L_l\|)$. We isolate exactly this residual into a single standard, checkable second-order identifiability condition.

Assumption 6 (Second-order geometry identifiability). Let (Θ^*, β^*) be a population minimizer (Theorem 1) and let $\mathcal{P}$ denote the finite group of atom permutations together with the global isometries of the latent coordinates L that fix $D = \kappa(\|L_k - L_l\|)$ (the intrinsic, unidentifiable symmetries of the structural block). Quotienting these out, the population Hessian $\nabla_L^2 \mathcal{R}(\Theta^*, \beta^*)$ of the risk restricted to the latent-geometry coordinates is positive definite on the orthogonal complement of the tangent space of the $\mathcal{P}$ -orbit at Θ^* : $v^\top \nabla_L^2 \mathcal{R}(\Theta^*, \beta^*) v \geq c_L \|v\|^2$ for all $v \perp T_{\Theta^*}(\mathcal{P} \cdot \Theta^*)$, some $c_L > 0$.

Remark 1. Assumption 6 is the textbook local (second-order) identifiability condition for a smooth parametric model: a Hessian nonsingular transverse to the model’s intrinsic symmetry orbit. It is mild and generically satisfied. By Assumption 4 the response $\psi(\alpha, \Gamma)$ has nondegenerate covariance and genuinely depends on the inter-atom geometry Γ , so the structural cost is a non-flat function of L near L^* except along the intrinsic isometry symmetry already quotiented out; a degenerate transverse Hessian is a measure-zero coincidence. It concerns only the latent-geometry block: by Lemma 5 the feature/mass/head block always has positive curvature, so the prediction guarantee never depends on Assumption 6. Because $\mathcal{R}$ is an explicit smooth population functional (Lemma 2), $\nabla_L^2 \mathcal{R}$ can be evaluated and its transverse spectrum checked numerically, so the condition is verifiable rather than assumed folklore.

Appendix A.3. Invariance and stability

Lemma 1 (Exact invariance to isometric reparametrization). Let $\varphi \in \text{Isom}(\Omega)$ be distance-preserving and $\varphi_\# \mu$ the pushforward (same features and weights, coordinates $p_i \mapsto \varphi(p_i)$). Then for every Θ and every S, J , $\mathcal{T}_\varepsilon^{S,J}(\varphi_\# \mu, \Theta) = \mathcal{T}_\varepsilon^{S,J}(\mu, \Theta)$ and hence $\Phi(\varphi_\# \mu; \Theta) = \Phi(\mu; \Theta)$.

Proof. The unrolled map of Definition 1 reads μ only through (F, a, C) : every iterate $C^{(s)}$, $K^{(s)}$, $T^{(s)}$ is an explicit function of (F, a, C) and Θ . Pushforward by φ leaves F, a unchanged and leaves C unchanged because $C_{i'l} = \kappa(\|p_i - p_{i'}\|) = \kappa(\|\varphi(p_i) - \varphi(p_{i'})\|)$ for an isometry. Hence every iterate is identical, and so are $T^{(s)}$ and Φ . $\square$

Remark 2. (F, a) is invariant to any coordinate warp; only C requires isometry. The feature-only part of Φ ($\alpha = 0$) is invariant to arbitrary warps, while the structural part ($\alpha > 0$) requires isometry exactly.

Lemma 2 (Smoothness and Lipschitzness of the unrolled map). Fix $\varepsilon > 0$ and $S, J < \infty$. On the compact set $\mathbb{T} \times \{\|f\| \leq R\}$ with marginals bounded below by $\rho := \min(\min_i a_i, b_{\min}) > 0$, the map $(\Theta, F, C) \mapsto \mathcal{T}_\varepsilon^{S,J}$ is C^∞ , and both $\Theta \mapsto \Phi(\mu; \Theta)$ and $(\Theta, \beta) \mapsto h_{\Theta, \beta}(\mu)$ are Lipschitz on $\mathbb{T} \times \{\|\beta\| \leq B\}$, with explicit constant $\Lambda_\varepsilon = (c_1 \varepsilon^{-1} \rho^{-2})^{S(J+1)} \text{poly}(R, K, \rho^{-1})$ for a universal c_1 .

Proof. Each constituent map in Definition 1 is C^∞ with derivative bounded on the stated compact domain: (i) the cost assembly $C^{(s)}$ is affine in $M(F, G)$ and in $T^{(s)}$ and polynomial in (C, D) , hence smooth with derivative bounded in terms of R and $\sup \kappa$; (ii) $X \mapsto \exp(-X/\varepsilon)$ is smooth with sup-norm derivative $\leq \varepsilon^{-1}$ on the bounded cost range; (iii) each Sinkhorn update $u \mapsto a \oslash (Kv)$ is a rational map, C^∞ and Lipschitz where row/column sums exceed $\rho e^{-C/\varepsilon} > 0$ (guaranteed by the marginal lower bound ρ and bounded costs), with Lipschitz constant $O(\varepsilon^{-1} \rho^{-2})$; (iv) the barycentric projection $T \mapsto (T^\top \mathbf{1}, \text{diag}(T^\top \mathbf{1})^{-1} T^\top F)$ is smooth and $O(\rho^{-1} R)$ -Lipschitz where atom mass $\geq \rho$. The estimand is the composition of S outer blocks, each a map of type (i)–(ii) and J maps of type (iii), followed by (iv). Smoothness is preserved under composition, and the Lipschitz constant is bounded by the product of the per-map constants, giving Λ_ε . Linearity of $h_{\Theta, \beta}$ in β contributes the factor $\|\beta\| \leq B$. $\square$

Remark 3. Λ_ε is finite for every fixed (ε, S, J) but grows with the unrolled depth and ε^{-1} ; this is the explicit price of dropping any nonconvex-uniqueness assumption. It enters the generalization rate (Theorem 2) only logarithmically.

Proposition 1 (Stability to non-isometric warps). Let φ be bi-Lipschitz with metric distortion $\eta := \|C^\varphi - C\|_\infty = \sup_{p, p'} |\kappa(\|\varphi(p) - \varphi(p')\|) - \kappa(\|p - p'\|)|$. Then, with Λ_ε as in Lemma 2, $\|\Phi(\varphi_\# \mu; \Theta) - \Phi(\mu; \Theta)\| \leq \Lambda_\varepsilon \eta$.

Proof. A warp changes μ only through $C \mapsto C^\varphi$ (F and a unchanged), with $\|C^\varphi - C\|_\infty = \eta$. By Lemma 2 the map $C \mapsto \Phi$ is Λ_ε -Lipschitz on the compact domain; applying it to a perturbation of size η gives the bound. $\square$

Remark 4. Proposition 1 predicts the observed behavior: exact invariance under rigid warps ($\eta = 0$, Lemma 1), and error growing at most linearly in non-rigid distortion η , i.e. controlled, graceful degradation rather than collapse.

Appendix A.4. Atom recovery

Lemma 3 (Barycentric atom recovery). *Under Assumptions 1,2,5, suppose the template features equal the truth up to a permutation, $g_{\Pi(k)} = g_k^*$, and $\varepsilon \leq c \delta^2 / \log K$ for a universal $c > 0$. Then with probability $\geq 1 - Ke^{-c' m \delta^2 / \sigma^2}$, the plan matches each point to the template atom of its generating atom, and the barycentric feature obeys*

$$\|[\Phi(\mu; \Theta)]_{\Pi(k)}^{\text{feat}} - g_k^*\| \leq O(\varepsilon) + O(\sigma / \sqrt{m}), \quad [\Phi(\mu; \Theta)]_{\Pi(k)}^{\text{mass}} = \frac{\alpha_k}{\sum_j \alpha_j} + O(\varepsilon).$$

Proof. Fix a point i generated by atom k , so $\|f_i - g_k^*\|_W \leq \tau$ with $\tau = O(\sigma)$. For any other active atom l , separation gives the feature-cost gap $M_{i,\Pi(l)} - M_{i,\Pi(k)} = \|f_i - g_l^*\|_W^2 - \|f_i - g_k^*\|_W^2 \geq \delta^2 - 2\delta\tau$. In each entropic layer the plan factors as $T_{ik} = u_i K_{ik} v_k$, so the relative mass point i sends to $\Pi(l)$ versus $\Pi(k)$ is

$$\frac{T_{i,\Pi(l)}}{T_{i,\Pi(k)}} = \frac{v_{\Pi(l)}}{v_{\Pi(k)}} \exp\left(-\frac{C_{i,\Pi(l)} - C_{i,\Pi(k)}}{\varepsilon}\right) \leq \Xi \exp\left(-\frac{(1-\alpha)(\delta^2 - 2\delta\tau) - O(\alpha)}{\varepsilon}\right),$$

where Ξ bounds the dual-scaling ratio $v_{\Pi(l)}/v_{\Pi(k)}$ and the $O(\alpha)$ term is the structure contribution; both are bounded on the compact domain by the potential bounds of Lemma 2. For $\varepsilon \leq c \delta^2 / \log K$ this ratio is $\leq K^{-2}$, so the total off-atom (leaked) mass of point i is $\leq K \cdot K^{-2} = K^{-1} = O(\varepsilon)$. Hence each point is matched to $\Pi(k)$ up to $O(\varepsilon)$ leakage. The barycentric feature at $\Pi(k)$ is the leak-weighted average of within-atom features; sub-Gaussian concentration of the within-atom mean over $\geq m$ points gives deviation $O(\sigma / \sqrt{m})$ with probability $\geq 1 - Ke^{-c' m \delta^2 / \sigma^2}$, and the $O(\varepsilon)$ leak (times bounded $\|f\| \leq R$) plus entropic blur give the $O(\varepsilon)$ bias. Masses follow from $T^\top \mathbf{1} = b$ and per-atom mass conservation up to the same $O(\varepsilon)$ leakage. $\square$

Remark 5. The structure term ($\alpha > 0$) can only sharpen the assignment when intra-sample geometry is consistent with the template; the bound above is the conservative feature-only estimate.

Lemma 4 (Bounded dual scalings of the unrolled map). *Fix $\varepsilon > 0$ and $S, J < \infty$. On the compact domain (costs $\leq \bar{C}$, marginals $\geq \rho > 0$) every Sinkhorn scaling $u^{(s,j)}, v^{(s,j)}$ in Definition 1 is strictly positive and bounded: since $K^{(s)} = \exp(-C^{(s)}/\varepsilon)$ has entries in $[e^{-\bar{C}/\varepsilon}, 1]$, each update $u \leftarrow a \odot (Kv)$, $v \leftarrow b \odot (K^\top u)$ maps the positive orthant into a box bounded away from 0; a finite composition of $S(J+1)$ such maps takes values in an explicit interval $[m_\varepsilon, M_\varepsilon]$ with $m_\varepsilon \geq \rho e^{-(J+1)\bar{C}/\varepsilon}$. Hence the dual-scaling ratio of Lemma 3 satisfies $\Xi = \max_{k,l} v_k/v_l \leq M_\varepsilon/m_\varepsilon < \infty$, an explicit constant, so the Ξ used in Lemma 3 is bounded, not assumed.*

Proof. $K^{(s)}$ entrywise $\geq e^{-\bar{C}/\varepsilon}$ gives $(Kv)_i \leq K\|v\|_\infty$ and $(Kv)_i \geq e^{-\bar{C}/\varepsilon} \sum_k v_k$, so with $a_i, b_k \in [\rho, 1]$ one update keeps the scalings in $[\rho e^{-\bar{C}/\varepsilon}/K, \rho^{-1} e^{\bar{C}/\varepsilon}]$; iterating the fixed $S(J+1)$ updates (a finite composition of continuous positive maps on the compact domain) yields the stated $[m_\varepsilon, M_\varepsilon]$. $\square$

Appendix A.5. Identifiability and consistency

Theorem 1 (Population identifiability up to permutation). *Under Assumptions 1–5, let $(\Theta^*, \beta^*) \in \arg \min_{\Theta, \beta} \mathcal{R}(\Theta, \beta)$. Then there is a permutation Π of atoms such that the recovered dictionary equals the truth, $g_{\Pi^{-1}(k)}^* = g_k^*$ for active atoms, and the induced predictor equals the best reparametrization-invariant predictor of Y ; i.e. h_{Θ^*, β^*} attains $\inf_{g \in \mathcal{G}_{\text{inv}}} \mathbb{E}\|Y - g(\mu)\|^2$, where $\mathcal{G}_{\text{inv}}$ is the class of predictors measurable with respect to the reparametrization-invariant σ -field.*

Proof. By Lemma 1, $\Phi(\cdot; \Theta)$ is measurable with respect to the invariant σ -field, so $h_{\Theta, \beta} \in \mathcal{G}_{\text{inv}}$ for every (Θ, β) ; thus the achievable risk is bounded below by the invariant Bayes risk. Conversely, by Assumption 3 the regression function depends on μ only through (α, Γ) , both invariant; by Lemma 3 a template matching the true features makes Φ an invertible (up to Π and $O(\varepsilon)$) encoding of (α, Γ) , so a linear head β reproduces $\langle B^*, \psi(\cdot) \rangle$ up to $O(\varepsilon)$. Hence the invariant Bayes risk is attained in the limit $\varepsilon \rightarrow 0$ and, for fixed ε , up to an explicit $O(\varepsilon)$ bias; Assumption 4 makes the minimizer's active-atom dictionary and head unique up to Π . $\square$

Lemma 5 (Unconditional quadratic growth on the prediction block). *Under Assumptions 1–5, decompose the parameter as $\Theta = (G, b, \beta; L, W)$ into the feature/mass/head block $\theta_f = (G, b, \beta)$ and the geometry block $\theta_g = (L, W)$. At a population minimizer (Θ^*, β^*) the risk grows quadratically in the feature/mass/head block, up to permutation: there is $c_f > 0$ with*

$$\mathcal{R}(\theta_f, \theta_g^*) - \mathcal{R}(\theta_f^*, \theta_g^*) \geq c_f \text{dist}(\theta_f, \Pi \cdot \theta_f^*)^2,$$

i.e. $\nabla_{\theta_f}^2 \mathcal{R}(\Theta^*, \beta^*) \geq c_f I$ transverse to the permutation orbit. This holds with no extra hypothesis beyond those already used for atom recovery.

Proof. On the saturated-assignment region guaranteed by Lemma 3 (in force whenever $\varepsilon \leq c \delta^2 / \log K$ and $\Xi < \infty$, Lemma 4), each barycentric feature coordinate is, up to the $O(\varepsilon)$ controlled bias, the identity readout of the corresponding atom feature: $[\Phi]_{\Pi(k)}^{\text{feat}} = g_k + O(\varepsilon)$ and $[\Phi]_{\Pi(k)}^{\text{mass}} = b_k + O(\varepsilon)$, with Jacobian $\partial \Phi / \partial (G, b)$ equal to the identity plus an $O(\varepsilon)$ (and $O(e^{-c' \delta^2 / \varepsilon})$ off-diagonal-leakage) perturbation, hence invertible and uniformly well-conditioned for $\varepsilon \leq c \delta^2 / \log K$ (this is where δ -separation enters). Consequently $\mathcal{R}(\Theta, \beta) = \mathbb{E} \|Y - \beta^\top \Phi(\mu; \Theta)\|^2$ is, in the (G, b, β) coordinates, a quadratic form in Φ composed with a well-conditioned diffeomorphism, whose Hessian factors as $J^\top \mathbb{E}[\Phi \Phi^\top] J$ plus head curvature, with J the invertible readout Jacobian. Assumption 4 ($\text{Cov}(\psi) > 0$, B^* with no zero block) makes $\mathbb{E}[\Phi \Phi^\top] > 0$ on the response-relevant coordinates, so the composite Hessian is positive definite transverse to the permutation orbit, with $c_f \geq \sigma_{\min}(J)^2 \lambda_{\min}(\mathbb{E}[\Phi \Phi^\top]) > 0$. The permutation orbit is the only flat direction because atoms are discrete and feature- δ -separated. $\square$

Remark 6. Lemma 5 folds the prediction-relevant part of the former well-separated-minimizer hypothesis into the standing assumptions: feature/mass/head identifiability is now unconditional given the recovery conditions (δ -separation, $\varepsilon \leq c \delta^2 / \log K$, bounded Ξ). What is not implied by feature-separation alone is curvature in the latent geometry block L , which enters $\mathcal{R}$ only through the structural quadratic $(C - D)^2$ and admits intrinsic isometry symmetries; this single residual is exactly Assumption 6.

Theorem 2 (Excess-risk consistency and rate). *Let $(\widehat{\Theta}_N, \widehat{\beta}_N) \in \arg \min_{\Theta \in \mathbb{T}, \|\beta\| \leq B} \frac{1}{N} \sum_n \|Y_n - h_{\Theta, \beta}(\mu_n)\|^2$. Under Assumptions 1–5 (no minimizer-separation needed),*

$$\mathcal{R}(\widehat{\Theta}_N, \widehat{\beta}_N) - \inf_{\Theta, \beta} \mathcal{R}(\Theta, \beta) = O_p\left(\sqrt{\frac{KC \log(\Lambda_\varepsilon)}{N}}\right).$$

Proof. By Lemma 2 the loss class $\{(\mu, Y) \mapsto \|Y - h_{\Theta, \beta}(\mu)\|^2\}$ is Lipschitz in the finite-dimensional (Θ, β) with constant $O(\Lambda_\varepsilon)$ and uniformly bounded (Assumption 5); a Lipschitz-parametric class over a compact set is Glivenko–Cantelli and Donsker, giving $\Delta_N := \sup_{\Theta, \beta} |\widehat{\mathcal{R}} - \mathcal{R}| = O_p(\sqrt{KC \log \Lambda_\varepsilon / N})$. Writing h^* for a population minimizer, ERM optimality and two applications of Δ_N give $\mathcal{R}(\widehat{h}) \leq \widehat{\mathcal{R}}(\widehat{h}) + \Delta_N \leq \widehat{\mathcal{R}}(h^*) + \Delta_N \leq \mathcal{R}(h^*) + 2\Delta_N$, i.e. excess risk $\leq 2\Delta_N$. No identifiability or separation is used, only uniform convergence. $\square$

Proposition 2 (Parameter recovery). *Under Assumptions 1–5, the population risk satisfies the well-separation (quadratic growth) condition $\mathcal{R}(\Theta, \beta) - \mathcal{R}(\Theta^*, \beta^*) \geq c \text{dist}((\Theta, \beta), \Pi \cdot (\Theta^*, \beta^*))^2$, $c > 0$, on the feature/mass/head block unconditionally (Lemma 5); on the latent geometry block it holds under the standard second-order identifiability Assumption 6, with $c = \min(c_f, c_L)$ after a Taylor expansion (the cross block of the Hessian only improves the constant by the Schur-complement bound). Whenever it holds, $(\widehat{\Theta}_N, \widehat{\beta}_N) \rightarrow (\Theta^*, \beta^*)$ in probability up to the permutation Π , by the argmin theorem applied to the uniform convergence above. Thus parameter recovery is unconditional for the prediction-relevant parameters (atom features, masses, head) and requires only the mild, checkable Assumption 6 for the latent-geometry coordinates; the excess-risk guarantee of Theorem 2 needs neither.*

Corollary 1 (Bias–smoothness trade-off). *Choosing $\varepsilon = \varepsilon_N \rightarrow 0$ with $\varepsilon_N^{-1} = o(N^{1/2} / \sqrt{KC})$ removes the $O(\varepsilon)$ identification bias of Lemma 3 while preserving consistency; the estimator then targets the unregularized invariant Bayes predictor.*

The prediction-relevant guarantees (Lemmas 1–4, Theorems 1–2) are unconditional, and Lemma 5 further makes exact recovery of the prediction-relevant parameters (atom features, masses, head) unconditional under the standing recovery conditions. The single remaining conditional element is the second-order identifiability Assumption 6 on the

latent-geometry block, used in Proposition 2 only for exact recovery of that block and never for prediction consistency. Identifiability is up to atom permutation, standard for dictionary and factor models, with no rotational ambiguity because the atoms are discrete and feature-separated. The unaligned-CP estimand of Tang, Kolda and Zhang [10] is unsupervised (reconstructing a single tensor via an RKHS mode) and carries no response; our estimand is the supervised, response-identifying object of Theorem 1, recovered through transport, and this supervised estimand is what separates the two settings. A minimax lower bound for invariant supervised recovery, behavior under misspecified K , and a low-rank multilinear head preserving Theorem 2 remain open.

Data and Code Availability

The data are public. The structural target uses BraTS 2021 ($n = 1181$) and grade uses BraTS 2020 ($n = 368$), distributed through Kaggle (handles dschettler8845/brats-2021-task1 and awsaf49/brats20-dataset-training-validation) and originating from the RSNA-ASNR-MICCAI BraTS challenge [67, 68, 69].

The analysis code (the code/ package, excluding development scratch and caches) will be deposited in a public repository before publication. **[CODE AVAILABILITY: TODO]**

CRedit Author Contributions

Axel Faes: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing — original draft. **Stéphanie M. van den Berg:** Supervision, Writing — review & editing. **Maryam Amir Haeri:** Supervision, Writing — review & editing.

Declaration of Competing Interest

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

Ethics

This study used only publicly available, fully de-identified, previously published data (the RSNA-ASNR-MICCAI BraTS 2021 and BraTS 2020 challenge datasets). No new human-subjects data were collected, and no additional ethical approval was required.

Funding

This work was conducted within the project “Predicting the Development of Primary Liver Cancer Using the AI-HCC Model”, which is funded by ZonMw, The Netherlands Organisation for Health Research and Development.

Declaration of Generative AI Use

During the preparation of this work the authors used a generative AI assistant to support drafting and editing of the manuscript text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- [1] Q. Zhao, C. F. Caiafa, D. P. Mandic, Z. C. Chao, Y. Nagasaka, N. Fujii, L. Zhang, A. Cichocki, Higher order partial least squares (HOPLS): A generalized multilinear regression method, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (7) (2013) 1660–1673. doi:10.1109/TPAMI.2012.254.
- [2] R. Bro, Multiway calibration. multilinear PLS, *Journal of Chemometrics* 10 (1) (1996) 47–61. doi:10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C.
- [3] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. P. M. van Stiphout, P. Granton, C. M. L. Zegers, R. Gillies, R. Boellard, A. Dekker, H. J. W. L. Aerts, Radiomics: Extracting more information from medical images using advanced feature analysis, *European Journal of Cancer* 48 (4) (2012) 441–446. doi:10.1016/j.ejca.2011.11.036.
- [4] R. J. Gillies, P. E. Kinahan, H. Hricak, Radiomics: Images are more than pictures, they are data, *Radiology* 278 (2) (2016) 563–577. doi:10.1148/radiol.2015151169.
- [5] T. Rohlfing, N. M. Zahr, E. V. Sullivan, A. Pfefferbaum, The SRI24 multichannel atlas of normal adult human brain structure, *Human Brain Mapping* 31 (5) (2010) 798–819. doi:10.1002/hbm.20906.
- [6] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: A survey, *IEEE Transactions on Medical Imaging* 32 (7) (2013) 1153–1190. doi:10.1109/TMI.2013.2265603.
- [7] B. B. Avants, C. L. Epstein, M. Grossman, J. C. Gee, Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain, *Medical Image Analysis* 12 (1) (2008) 26–41. doi:10.1016/j.media.2007.06.004.
- [8] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, N. Courty, Optimal transport for structured data with application on graphs, in: *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Vol. 97 of PMLR, 2019, pp. 6275–6284, arXiv:1805.09114.
- [9] F. Mémoli, Gromov–wasserstein distances and the metric approach to object matching, *Foundations of Computational Mathematics* 11 (4) (2011) 417–487. doi:10.1007/s10208-011-9093-5.
- [10] R. Tang, T. G. Kolda, A. R. Zhang, Tensor decomposition with unaligned observations, *SIAM Journal on Matrix Analysis and Applications* 47 (1) (2026) 107–131, arXiv:2410.14046. doi:10.1137/24M1692836.
- [11] B. W. Larsen, T. G. Kolda, A. R. Zhang, A. H. Williams, Tensor decomposition meets RKHS: Efficient algorithms for smooth and misaligned data, arXiv preprint arXiv:2408.05677ArXiv:2408.05677 (2024).
- [12] S. Mazelet, R. Flamary, B. Thirion, Learning fMRI activations dictionaries across individual geometries via optimal transport, arXiv preprint arXiv:2605.20883ArXiv:2605.20883 (2026).
- [13] C. Vincent-Cuaz, R. Flamary, M. Corneli, T. Vayer, N. Courty, Template based graph neural network with optimal transport distances, in: *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, arXiv:2205.15733.
- [14] H. Gong, J. Luo, S. Mao, Y. Gong, Y. Lu, J. Ding, X. Zhu, Robust prostate cancer risk stratification from unregistered mpMRI via learned cross-modal correspondence, *Frontiers in Oncology* 15 (2026) 1700447. doi:10.3389/fonc.2025.1700447.
- [15] S. Wold, A. Ruhe, H. Wold, W. J. Dunn, III, The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses, *SIAM Journal on Scientific and Statistical Computing* 5 (3) (1984) 735–743. doi:10.1137/0905052.
- [16] P. Geladi, B. R. Kowalski, Partial least-squares regression: A tutorial, *Analytica Chimica Acta* 185 (1986) 1–17. doi:10.1016/0003-2670(86)80028-9.

- [17] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems* 58 (2) (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.
- [18] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, in: *Subspace, Latent Structure and Feature Selection (SLSFS 2005)*, Vol. 3940 of *Lecture Notes in Computer Science*, Springer, 2006, pp. 34–51. doi:10.1007/11752790_2.
- [19] S. de Jong, SIMPLS: An alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems* 18 (3) (1993) 251–263. doi:10.1016/0169-7439(93)85002-X.
- [20] H. Zhou, L. Li, H. Zhu, Tensor regression with applications in neuroimaging data analysis, *Journal of the American Statistical Association* 108 (502) (2013) 540–552. doi:10.1080/01621459.2013.776499.
- [21] E. F. Lock, Tensor-on-tensor regression, *Journal of Computational and Graphical Statistics* 27 (3) (2018) 638–647. doi:10.1080/10618600.2017.1401544.
- [22] G. Rabusseau, H. Kadri, Low-rank regression with tensor responses, in: *Advances in Neural Information Processing Systems* 29 (NeurIPS), 2016, pp. 1867–1875.
- [23] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, *SIAM Review* 51 (3) (2009) 455–500. doi:10.1137/07070111X.
- [24] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, C. Faloutsos, Tensor decomposition for signal processing and machine learning, *IEEE Transactions on Signal Processing* 65 (13) (2017) 3551–3582. doi:10.1109/TSP.2017.2690524.
- [25] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, H. A. Phan, Tensor decompositions for signal processing applications: From two-way to multiway component analysis, *IEEE Signal Processing Magazine* 32 (2) (2015) 145–163. doi:10.1109/MSP.2013.2297439.
- [26] K. Park, S. Keleş, Sparse higher order partial least squares for simultaneous variable selection, dimension reduction, and tensor denoising, *Biometrika* Advance access; arXiv:2310.09428 (2025). doi:10.1093/biomet/asaf082.
- [27] J. B. Kruskal, Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, *Linear Algebra and its Applications* 18 (2) (1977) 95–138. doi:10.1016/0024-3795(77)90069-6.
- [28] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, N. Courty, Fused Gromov-Wasserstein distance for structured objects, *Algorithms* 13 (9) (2020) 212. doi:10.3390/a13090212.
- [29] S. Chowdhury, F. Méoli, The Gromov–Wasserstein distance between networks and stable network invariants, *Information and Inference: A Journal of the IMA* 8 (4) (2019) 757–787. doi:10.1093/imaiai/iaz026.
- [30] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in: *Advances in Neural Information Processing Systems (NIPS)*, Vol. 26, 2013, arXiv:1306.0895.
- [31] J. Solomon, G. Peyré, V. G. Kim, S. Sra, Entropic metric alignment for correspondence problems, *ACM Transactions on Graphics* 35 (4) (2016) 72:1–72:13. doi:10.1145/2897824.2925903.
- [32] G. Peyré, M. Cuturi, J. Solomon, Gromov–wasserstein averaging of kernel and distance matrices, in: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, Vol. 48 of *PMLR*, 2016, pp. 2664–2672.
- [33] M. Agueh, G. Carlier, Barycenters in the Wasserstein space, *SIAM Journal on Mathematical Analysis* 43 (2) (2011) 904–924. doi:10.1137/100805741.

- [34] M. Cuturi, A. Doucet, Fast computation of Wasserstein barycenters, in: Proceedings of the 31st International Conference on Machine Learning (ICML), Vol. 32 of Proceedings of Machine Learning Research, PMLR, 2014, pp. 685–693.
- [35] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. H. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, T. Vayer, POT: Python optimal transport, Journal of Machine Learning Research 22 (78) (2021) 1–8.
- [36] C. Villani, Optimal Transport: Old and New, Vol. 338 of Grundlehren der mathematischen Wissenschaften, Springer, Berlin, Heidelberg, 2009. doi:10.1007/978-3-540-71050-9.
- [37] G. Peyré, M. Cuturi, Computational optimal transport, Foundations and Trends in Machine Learning 11 (5–6) (2019) 355–607. doi:10.1561/22000000073.
- [38] F. Santambrogio, Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling, Vol. 87 of Progress in Nonlinear Differential Equations and Their Applications, Birkhäuser, 2015. doi:10.1007/978-3-319-20828-2.
- [39] N. Courty, R. Flamary, D. Tuia, A. Rakotomamonjy, Optimal transport for domain adaptation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (9) (2017) 1853–1865. doi:10.1109/TPAMI.2016.2615921.
- [40] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, G. K. Rohde, Optimal mass transport: Signal processing and machine-learning applications, IEEE Signal Processing Magazine 34 (4) (2017) 43–59. doi:10.1109/MSP.2017.2695801.
- [41] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, T. Poggio, Learning with a Wasserstein loss, in: Advances in Neural Information Processing Systems 28 (NeurIPS), 2015.
- [42] M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngolè, D. Coeurjolly, M. Cuturi, G. Peyré, J.-L. Starck, Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning, SIAM Journal on Imaging Sciences 11 (1) (2018) 643–678. doi:10.1137/17M1140431.
- [43] C. Vincent-Cuaz, T. Vayer, R. Flamary, M. Corneli, N. Courty, Online graph dictionary learning, in: Proceedings of the 38th International Conference on Machine Learning (ICML), Vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 10564–10574.
- [44] H. Xu, D. Luo, H. Zha, L. Carin, Gromov-Wasserstein learning for graph matching and node embedding, in: Proceedings of the 36th International Conference on Machine Learning (ICML), Vol. 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 6932–6941.
- [45] A. Genevay, G. Peyré, M. Cuturi, Learning generative models with sinkhorn divergences, in: Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS), Vol. 84 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 1608–1617.
- [46] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trounev, G. Peyré, Interpolating between optimal transport and MMD using sinkhorn divergences, in: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), Vol. 89 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 2681–2690.
- [47] T. Séjourné, F.-X. Vialard, G. Peyré, The unbalanced Gromov Wasserstein distance: Conic formulation and relaxation, in: Advances in Neural Information Processing Systems 34 (NeurIPS), 2021.
- [48] T. Vayer, R. Flamary, R. Tavenard, L. Chapel, N. Courty, Sliced Gromov-Wasserstein, in: Advances in Neural Information Processing Systems 32 (NeurIPS), 2019.

- [49] H. Xu, D. Luo, L. Carin, Scalable gromov-Wasserstein learning for graph partitioning and matching, in: *Advances in Neural Information Processing Systems* 32 (NeurIPS), 2019.
- [50] Z. Cang, Y. Wu, Y. Zhao, Supervised gromov-wasserstein optimal transport with metric-preserving constraints, *SIAM Journal on Mathematics of Data Science* 7 (1) (2025) 301–328, arXiv:2401.06266. doi:10.1137/24M1630499.
- [51] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, A. J. Smola, Deep sets, in: *Advances in Neural Information Processing Systems (NIPS)*, Vol. 30, 2017, pp. 3391–3401, arXiv:1703.06114.
- [52] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, Y. W. Teh, Set transformer: A framework for attention-based permutation-invariant neural networks, in: *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Vol. 97 of *Proceedings of Machine Learning Research*, 2019, pp. 3744–3753.
- [53] C. R. Qi, H. Su, K. Mo, L. J. Guibas, PointNet: Deep learning on point sets for 3D classification and segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 652–660.
- [54] C. R. Qi, L. Yi, H. Su, L. J. Guibas, PointNet++: Deep hierarchical feature learning on point sets in a metric space, in: *Advances in Neural Information Processing Systems* 30 (NeurIPS), 2017.
- [55] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, D. J. Hawkes, Nonrigid registration using free-form deformations: application to breast MR images, *IEEE Transactions on Medical Imaging* 18 (8) (1999) 712–721. doi:10.1109/42.796284.
- [56] S. Klein, M. Staring, K. Murphy, M. A. Viergever, J. P. W. Pluim, elastix: A toolbox for intensity-based medical image registration, *IEEE Transactions on Medical Imaging* 29 (1) (2010) 196–205. doi:10.1109/TMI.2009.2035616.
- [57] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, A. V. Dalca, VoxelMorph: A learning framework for deformable medical image registration, *IEEE Transactions on Medical Imaging* 38 (8) (2019) 1788–1800. doi:10.1109/TMI.2019.2897538.
- [58] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, I. Išgum, A deep learning framework for unsupervised affine and deformable image registration, *Medical Image Analysis* 52 (2019) 128–143. doi:10.1016/j.media.2018.11.010.
- [59] M. F. Beg, M. I. Miller, A. Trouvé, L. Younes, Computing large deformation metric mappings via geodesic flows of diffeomorphisms, *International Journal of Computer Vision* 61 (2) (2005) 139–157. doi:10.1023/B:VISI.0000043755.93987.aa.
- [60] Y. Yang, L.-F. Yan, X. Zhang, Y. Han, H.-Y. Nan, Y.-C. Hu, B. Hu, S.-L. Yan, J. Zhang, D.-L. Cheng, X.-W. Ge, G.-B. Cui, D. Zhao, W. Wang, Glioma grading on conventional MR images: A deep learning study with transfer learning, *Frontiers in Neuroscience* 12 (2018) 804. doi:10.3389/fnins.2018.00804.
- [61] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, *Nature Methods* 18 (2) (2021) 203–211. doi:10.1038/s41592-020-01008-z.
- [62] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, Brain tumor segmentation with deep neural networks, *Medical Image Analysis* 35 (2017) 18–31. doi:10.1016/j.media.2016.05.004.
- [63] S. Pereira, A. Pinto, V. Alves, C. A. Silva, Brain tumor segmentation using convolutional neural networks in MRI images, *IEEE Transactions on Medical Imaging* 35 (5) (2016) 1240–1251. doi:10.1109/TMI.2016.2538465.

- [64] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.
- [65] A. W. van der Vaart, J. A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series in Statistics, Springer, New York, 1996.
- [66] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Vol. 48 of Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.
- [67] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al., The multimodal brain tumor image segmentation benchmark (BRATS), *IEEE Transactions on Medical Imaging* 34 (10) (2015) 1993–2024. doi:10.1109/TMI.2014.2377694.
- [68] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, et al., Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features, *Scientific Data* 4 (2017) 170117. doi:10.1038/sdata.2017.117.
- [69] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, *arXiv preprint arXiv:1811.02629* (2018).
- [70] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: *Workshop Track, International Conference on Learning Representations (ICLR)*, 2014, arXiv:1312.6034.
- [71] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Vol. 70 of *Proceedings of Machine Learning Research*, 2017, pp. 3319–3328.
- [72] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.