TOPOMATCH: TOPOLOGY-CONSTRAINED SEMI-BALANCED TRANSPORT FOR OPEN-WORLD 3D VASCULAR GRAPH PARSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Anatomical vessel labeling is more structured than closed-set branch classification: expected vessels may be absent, one vessel may be fragmented into multiple observations, and unseen anatomy may coexist with extraction artifacts. We introduce TopoMatch, an open-world framework that formulates 3D vascular parsing as structured correspondence between observed branches and an anatomical label space. A reversal-invariant curve tokenizer and an incidence-aware encoder retain ordered geometry and physical branch–junction structure. Topology-aware semi-balanced matching preserves branch-wise assignment, softly constrains the mass of known classes, and leaves separate unknown-anatomy and artifact slots unconstrained in total mass. A learned gate combines local and structured predictions before conditional topology-aware decoding. On CoW-VTP, TopoMatch achieves 94.72% Macro-F1 and 42.0% exact-graph accuracy. Capacity-matched controls remain close on clean recognition, whereas controlled experiments locate the clearest benefit of relaxed target mass in missing-anatomy regimes. Strict open-set evaluation further shows that separate free slots improve the distinction between unknown anatomy and extraction artifacts. Corruption-aware training improves robustness with little change in clean performance. Task-specific retraining yields consistent branch-level gains across graph-construction pipelines, larger vascular graphs, and coronary anatomy, though whole-graph recovery remains challenging on TopBrain. These results support open-world structured matching without a fixed one-to-one anatomical template.

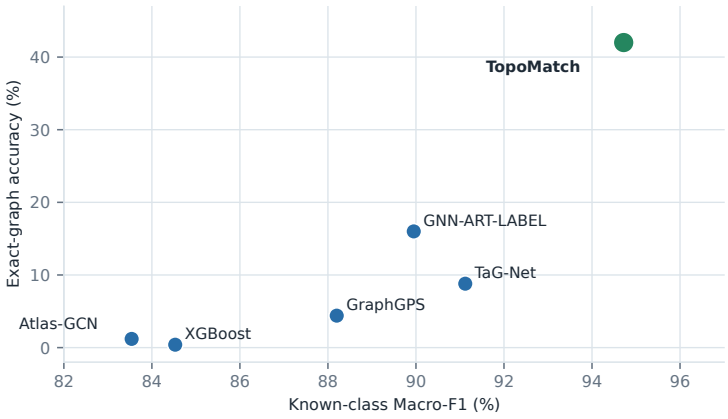

Figure 1: Teaser. CoW-VTP system-level comparison: TopoMatch reaches 94.72% Macro-F1 and 42.0% exact-graph accuracy; the strongest task-adapted prior-method results are 91.12% Macro-F1 (TaG-Net) and 16.0% exact-graph accuracy (GNN-ART-LABEL), respectively.

1 INTRODUCTION

Vessel extraction provides geometry, whereas downstream analysis requires anatomical branch identities for standardized territories, branch-level measurements, and structured reporting. Anatomical labeling has therefore been studied for intracranial, head-and-neck, and coronary vessels (Chen et al., 2020; Yao et al., 2023; Hampe et al., 2024; Vos et al., 2024); standardized coronary reporting likewise

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

organizes findings by named arterial segments (Dehmer et al., 2020). Parsing thus remains distinct from segmentation.

This problem is not ordinary closed-set classification. Local curves can look similar despite different anatomical identities, so their connections to neighboring branches matter. The observed graph may not match a fixed anatomical template: expected vessels can be absent, one vessel can be split into several extracted segments, and unfamiliar but plausible anatomy can coexist with spurious extraction artifacts. With arbitrary centerline traversal direction, parsing must combine geometric invariance and structural context while distinguishing missing classes, fragmented observations, and different causes of rejection.

Graph and topology-aware vessel-labeling models provide relational context (Chen et al., 2020; Yao et al., 2023; Vos et al., 2024), but are typically optimized for a predefined vocabulary. Generic graph Transformers offer stronger contextual encoders (Rampášek et al., 2022) without explicitly specifying how an incomplete graph should map to reusable anatomical classes. Open-set recognition allows rejection (Bendale & Boult, 2016), and unbalanced transport relaxes prescribed correspondence mass (Chizat et al., 2018); neither alone expresses the physical junction relations needed for anatomical parsing. The key question is how to explain every observed branch without forcing the graph into a complete, one-to-one template.

We propose TopoMatch for open-world structured matching. Label-blind branches pass through a shared forward/reverse curve encoder and an incidence-aware encoder that retains physical junctions. The matcher then combines local predictions, anatomical prototypes, graph-level class visibility, and junction compatibility. It normalizes each observed branch but imposes only soft mass constraints on known anatomical classes; unknown anatomy and extraction artifacts receive separate free slots. Adaptive fusion retains reliable local evidence, followed by calibrated, conditional structural decoding.

Our contributions are threefold. First, we formulate anatomical vessel labeling as label-blind open-world vessel-graph parsing, explicitly accommodating missing anatomical classes, repeated fragments of the same vessel, unknown anatomy, and extraction artifacts without requiring a fixed one-to-one correspondence. Second, we develop a reversal-invariant and incidence-aware representation together with semi-balanced topology-aware matching, integrating branch geometry, explicit physical junctions, visibility-conditioned class mass, and anatomical compatibility within a unified structured prediction framework. Third, we provide a comprehensive evaluation across patient-disjoint splits, capacity-matched controls, mechanism-specific stress tests, controlled graph corruptions, strict open-set settings, and task-specific retraining across different vascular anatomies and graph-construction pipelines.

2 Related Work

Vessel labeling and geometric representation. Relational GNNs, atlas-based graph convolutions, and topology-aware centerline networks exploit anatomical context (Chen et al., 2020; Zhu et al., 2022; Vos et al., 2024; Yao et al., 2023). PointNet/PointNet++, DGCNN, and CurveNet provide general 3D encoders (Qi et al., 2017a;b; Wang et al., 2019; Xiang et al., 2021), while GraphGPS combines local propagation and global attention (Rampášek et al., 2022). Vessel curves additionally require invariance to traversal reversal and a representation of shared multi-branch junctions. Our incidence representation retains those physical entities, related in spirit to vertex–hyperedge aggregation (Chien et al., 2022), rather than reducing every junction to independent line-graph edges. Topology-preserving segmentation (Shit et al., 2021) is complementary: it improves extraction but does not assign anatomical identities to an already imperfect graph.

Structured correspondence and open-set recognition. Entropic transport enables differentiable matching (Cuturi, 2013), while graph matching can incorporate relational costs (Xu et al., 2019a). However, these methods are mainly designed for correspondence between observed sets or graphs, whereas vessel parsing maps a variable graph to a reusable anatomical vocabulary in which classes may be absent and multiple fragments may share one label. Relaxed transport marginals (Chizat et al., 2018; Pham et al., 2020) address variable class mass but do not model anatomical branch compatibility. Open-set methods (Hendrycks & Gimpel, 2017; Bendale & Boult, 2016; Liu et al., 2020) can reject unseen categories, but typically treat rejection independently and do not distinguish unknown anatomy from extraction artifacts. These limitations motivate a structured open-world

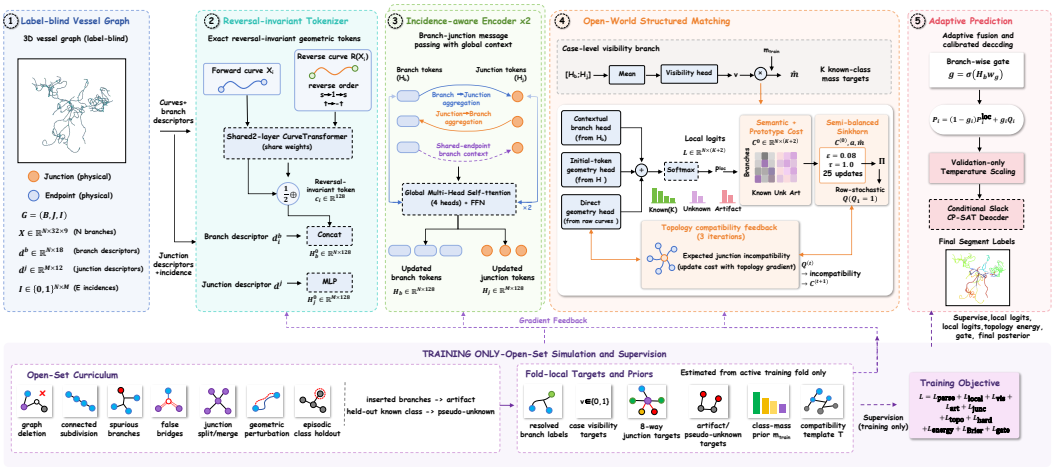

Figure 2: TopoMatch overview. Label-blind curves and physical junctions feed an invariant incidence-aware encoder. Visibility-conditioned semi-balanced matching and junction feedback produce a structured posterior, which is fused with local evidence before conditional decoding. The lower rail is used only during training.

formulation that jointly models variable class mass, vascular topology, and distinct rejection causes. Appendix A provides the expanded literature discussion.

3 Method

3.1 Label-Blind Open-World Parsing

We represent the observed graph as $G = (\mathcal{B}, \mathcal{J}, \mathcal{I})$, with branches $\mathcal{B} = \{b_i\}_{i=1}^{N}$, physical endpoints and junctions $\mathcal{J} = \{j_r\}_{r=1}^{M}$, and incidences $\mathcal{I} \subseteq \mathcal{B} \times \mathcal{J}$. Given K known anatomical classes, each branch receives a known label, an unknown-anatomy label, or an artifact label:

$$\mathbf{P} \in [0,1]^{N\times(K+2)}, \qquad \sum_{c=1}^{K+2} P_{ic} = 1 \quad (i = 1, \dots, N). \tag{1}$$

There is no one-to-one constraint: several fragments may share a known class, and a known class may receive no observations. Unknown anatomy denotes plausible anatomy outside the vocabulary; an artifact denotes a spurious extraction outcome.

For CoW-VTP and the mask-derived graphs, semantic labels do not determine branch boundaries. We first deduplicate undirected centerline edges and delimit maximal chains at nodes with degree other than two. Consecutive edges within each maximal chain are greedily grouped while their cumulative arc length, normalized by the graph-scale diagonal, remains at most $\eta = 0.04$; individual edges are not subdivided. Each resulting segment is resampled to $S = 32$ equidistant points. Semantic labels are projected only after the graph structure has been fixed: a segment retains its length-majority label when its purity is at least 0.8; otherwise, it is marked as *unresolved*. Unresolved segments are excluded from label-dependent supervision and evaluation and are neither an output category nor unknown-anatomy supervision. ImageCAS-X instead uses its released centerline polylines directly as branches, without η-based subdivision. Figure 2 summarizes the pipeline.

3.2 Reversal-Invariant Curves and Incidence-Aware Context

For branch b_i, pointwise features $\mathbf{X}_i$ contain centered, scale-normalized coordinates, normalized arc positions, unit tangents, curvature magnitudes, and radii. Reversal $\mathcal{R}$ reverses sample order, maps arc position s to $1 - s$, and negates tangent direction. Applying a shared curve Transformer to both traversals gives

$$\mathbf{c}_i = \tfrac{1}{2}\big[f_\theta(\mathbf{X}_i) + f_\theta(\mathcal{R}(\mathbf{X}_i))\big]. \tag{2}$$

The same encoder and symmetric average make this token invariant to stored traversal direction: applying $\mathcal{R}$ again swaps the two summands. The token is combined with symmetric branch descriptors; junction tokens are initialized from geometry, degree, and incident-branch statistics (Appendix B).

Each incidence-aware layer first aggregates incident branch features at junctions. A branch then receives information from its endpoint junctions and from branches sharing either endpoint. Separate residual MLPs update branch and junction tokens, followed by multi-head self-attention for global context. Maintaining explicit junction tokens preserves a multi-branch junction as one shared physical event instead of expanding it into independent pairwise edges.

3.3 Topology-Constrained Semi-Balanced Matching

Three complementary paths—contextual graph features, initial curve/descriptor tokens, and direct forward/reverse geometry—produce summed local logits and a posterior $\mathbf{p}_i^{\mathrm{loc}} \in [0,1]^{K+2}$. Known-class matching additionally uses squared distance to a learned anatomical prototype; unknown and artifact costs rely directly on local probabilities. A graph-level visibility head rescales the training-fold class prior, estimated from resolved branch-length shares. Thus the known-class target mass adapts to the observed graph instead of requiring a complete anatomical inventory.

Let $\mathbf{Q} \in [0,1]^{N\times(K+2)}$ denote the structured posterior. For fixed network predictions and graph-level priors, structured assignment is obtained from the inner objective

$$\mathcal{L}_{\mathrm{match}} = \mathcal{L}_{\mathrm{semantic}} + \tau\mathcal{L}_{\mathrm{mass}} + \lambda_T\mathcal{L}_{\mathrm{junction}}, \qquad \sum_{c=1}^{K+2} Q_{ic} = 1. \tag{3}$$

The semantic term uses local evidence and known-class prototypes. The mass term softly encourages the known columns to follow visibility-conditioned priors. Unknown and artifact columns have no prescribed total mass; both remain subject to branch-wise normalization. Relaxing only target mass explains every observation while allowing absent classes and repeated fragments. Balanced transport is not inherently one-to-one: multiple rows may share a semantic class. Its relevant restriction is prescribed target mass, which can favor assigning observations to an absent class. The soft known-class penalty relaxes this restriction; free columns distinguish the two rejection causes.

The training-derived matrix $\mathbf{T}$ records compatibility between known labels incident to a shared physical junction. Unknown and artifact labels are exempt because neither has a fixed anatomical neighborhood. The junction term penalizes unlikely known-label combinations and is normalized by the number of eligible incident branch pairs. Starting from semantic costs, topology feedback accumulates expected incompatibility from neighboring branches and updates known-class costs before entropically regularized generalized Sinkhorn refinement. If $\widetilde{\mathbf{Q}}^{(t+1)}$ is the current refinement, damping gives

$$\mathbf{Q}^{(t+1)} = (1-\alpha)\mathbf{Q}^{(t)} + \alpha\widetilde{\mathbf{Q}}^{(t+1)}, \qquad \alpha \in (0,1]. \tag{4}$$

This convex combination preserves row normalization. An initial 25-iteration log-domain Sinkhorn solve is followed by three topology-feedback rounds, each with 25 additional Sinkhorn iterations. Damped refinement uses $\alpha = 0.5$; the remaining prior-construction and solver settings are detailed in Appendix B.

3.4 Adaptive Prediction and Open-Set Training

The fused posterior $\mathbf{p}_i$ forms the ith row of the prediction matrix $\mathbf{P}$ before conditional decoding. A learned branch-wise gate combines the local and structured posteriors:

$$\mathbf{p}_i = (1-g_i)\mathbf{p}_i^{\mathrm{loc}} + g_i\mathbf{q}_i, \qquad g_i \in [0,1]. \tag{5}$$

The branch-wise gate is initialized to $g_i = 0.25$ and regularized toward this local-favoring value, so that local evidence retains greater weight when the structured prediction is less reliable. Validation-only temperature scaling calibrates the fused scores. If their labels contain an incompatible known-class pair, a slack-based CP-SAT decoder minimizes negative log-probability plus incompatibility penalties. Slack variables preserve feasibility; no class-capacity or visibility constraints are imposed. Already compatible predictions are unchanged. Consequently, soft matching and final discrete decoding are distinct components, evaluated separately below.

Training centers on class-balanced parsing over supervised branches after augmentation, including resolved real branches, synthetic artifacts, and episodic pseudo-unknowns. Auxiliary losses supervise local classification, graph-level visibility, artifact detection, probability calibration, junction classification, known/unknown energy separation, gate regularization, topology consistency, and difficult branches (Appendix B). Stochastic graph editing generates extraction errors, while episodic class holdout provides pseudo-unknown anatomy. Edits include deletion, connected subdivision, spurious branches, false bridges, junction split/merge, and geometric perturbation. Newly inserted branches are artifacts; an episodically selected known class is pseudo-unknown. These training mechanisms are distinct from the strict leave-one-class-out evaluation, where the test-unknown class is excluded throughout training.

4 Experimental Setup

Benchmarks and leakage control. The primary CoW-VTP benchmark contains 250 expert-corrected CTA/MRA centerline graphs from 125 patients and 13 classes (Musio et al., 2026). TopCoW-mask reconstructs graphs from masks of the same 125-patient cohort and vocabulary (Yang et al., 2026a), isolating graph-construction changes rather than an independent patient population. TopBrain has 50 paired graphs from 25 patients and 36 classes (Yang et al., 2026b); ImageCAS-X has 800 coronary CCTA scans and 14 segment categories (Bransby et al., 2026). All methods are retrained separately per benchmark: these experiments test task-adapted generality, not zero-shot transfer.

The cerebral benchmarks use patient-disjoint five-fold cross-validation, keeping paired CTA/MRA acquisitions together. Each CoW-VTP fold has 80 training, 20 validation, and 25 test patients; out-of-fold (OOF) evaluation covers all 125 patients once. ImageCAS-X follows its official split. Class-mass and compatibility priors use only the training partition; calibration uses validation patients only. Geometry is normalized independently per graph. Graph statistics, annotation coverage, and expanded protocols appear in Appendix C.

Table 1: Benchmark size and resolved annotation coverage for the evaluated branch graphs. Coverage is branch / length coverage (%); unresolved segments are not scored as unknown anatomy.

Benchmark	Classes	Graphs / patients	Branches	Coverage
CoW-VTP	13	250 / 125	23,001	94.05 / 93.75
TopCoW-mask	13	250 / 125	33,783	94.01 / 93.26
TopBrain	36	50 / 25	25,255	92.83 / 90.87
ImageCAS-X	14	800 / 800	11,367	99.31 / 99.97

Primary metrics exclude unresolved segments; Table 1 makes this coverage limitation explicit across extraction pipelines.

Baselines and matched controls. We adapt XGBoost (Chen & Guestrin, 2016), GNN-ART-LABEL (Chen et al., 2020), Atlas-GCN (Vos et al., 2024), GraphGPS (Rampášek et al., 2022), and graph-only TaG-Net (Yao et al., 2023) to the same branch graphs, targets, and patient splits. The neural baselines use the same maximum epoch and effective batch-size budget, with adaptation-specific architectures, losses, and optimization details given in Appendix C.1. Separate branch-MLP, GNN, and capacity-matched GraphGPS controls use the same input representation and protocol as TopoMatch; the matched GraphGPS is also evaluated with the same CP-SAT decoder. These controls must be distinguished from the smaller task-adapted GraphGPS baseline.

Training and metrics. TopoMatch uses $S = 32$, hidden width 128, two incidence layers, four attention heads, and dropout 0.1. AdamW training runs for at most 200 epochs with batch size 8, learning rate 5×10^{-4}, weight decay 10^{-4}, and validation-based early stopping. Architecture and optimization settings are shared across benchmarks; only vocabularies and training-fold statistics change. Full settings are in Appendix B.

Primary metrics are known-class Macro-F1 on resolved segments and exact-graph accuracy (all resolved segments correct). Length-weighted F1, junction exact accuracy, and training-template

topology consistency complement them. Confidence intervals use 2,000 patient-cluster bootstrap samples, resampling paired modalities together. Open-set evaluation uses five strict leave-one-class-out (LOCO) tasks—BA, R-PCA, R-Pcom, Acom, and 3rd-A2—with the held-out class excluded from training. Unknown-anatomy and synthetic-artifact AUPRC are reported separately, together with OSCR. Robustness evaluates six edit families at severities S0–S4 using Macro-F1 robustness AUC (RAUC). Exact metric definitions and the distinction between synthetic artifacts and novel anatomy are retained in Appendix C.

5 Results

5.1 Primary Results and Capacity-Matched Controls

Table 2: Main five-fold OOF comparison on CoW-VTP. All values are percentages.

Method	Macro-F1	Length-F1	Exact graph	Junction exact	Topology
XGBoost	84.53	84.54	0.4	87.01	95.84
Atlas-GCN	83.54	83.45	1.2	86.44	97.34
GraphGPS	88.20	88.26	4.4	88.28	96.68
GNN-ART-LABEL	89.95	89.93	16.0	92.43	97.91
TaG-Net (graph)	91.12	90.98	8.8	93.07	98.48
TopoMatch	**94.72**	**94.64**	**42.0**	**96.40**	**99.48**

Table 2 reports five-fold OOF results. TopoMatch reaches 94.72% Macro-F1, exceeding the strongest task-adapted baseline, TaG-Net, by 3.60 points (paired 95% CI $[2.09, 5.18]$). Exact-graph accuracy reaches 42.0%, versus 16.0% for the strongest prior-method result on this metric, GNN-ART-LABEL. These system-level gains motivate coordinated parsing but do not isolate model capacity.

Table 3: Capacity-matched analysis under the same five-fold OOF protocol.

Method	Params	Macro-F1	Length-F1	Exact graph
Branch MLP	0.021M	80.15	80.27	1.2
Matched GNN	0.768M	87.04	87.13	1.2
Matched GPS	1.032M	94.57	94.80	40.0
Matched GPS + decoder	1.032M	94.65	**94.85**	40.4
TopoMatch	1.155M	**94.72**	94.64	**42.0**

The matched controls substantially narrow the gap (Table 3). Matched GraphGPS with the common decoder reaches 94.65% Macro-F1 and 40.4% exact-graph accuracy, only 0.07 and 1.6 points below TopoMatch; it has better length-weighted F1. Three-seed Macro-F1 is 94.62 ± 0.21 for this control versus 94.89 ± 0.24 for TopoMatch (Appendix D). Clean recognition therefore shows only a modest separation between TopoMatch and the strongest capacity-matched control; the stress tests below isolate the regime motivating semi-balanced matching.

5.2 Task-Adapted Generality

Table 4 reports all six metrics after separate retraining. TopoMatch has the highest Macro-F1, length-weighted F1, and balanced accuracy on all three additional benchmarks. On TopCoW-mask, which changes extraction within the same cohort, exact-graph accuracy rises from at most 0.4% among baselines to 3.6%. On coronary ImageCAS-X, Macro-F1 reaches 87.61% versus 77.62% for TaG-Net, and exact-graph accuracy reaches 64.4% versus the strongest baseline's 40.0%. These are task-adapted gains, not zero-shot transfer; paired confidence intervals are in Appendix F.

TopBrain exposes a structural limit despite the best Macro-F1 (67.82%): all methods have zero exact-graph accuracy, and TaG-Net has better junction exact accuracy (72.29% versus 70.39%) and topology consistency (98.11% versus 97.34%). The full table therefore distinguishes improved branch recognition from unsolved large-graph recovery. Figure 3 likewise retains visible TopBrain errors.

Table 4: Full cross-benchmark comparison after task-specific retraining. All values are percentages. Junction denotes junction-exact accuracy; Topology denotes topology consistency. Confidence intervals are in Appendix F.

Method	Macro-F1	Length-F1	Bal. Acc.	Exact graph	Junction	Topology
TopCoW-mask						
XGBoost	80.22	82.13	80.05	0.4	83.33	99.24
GNN-ART-LABEL	80.99	83.05	82.57	0.0	84.47	99.31
Atlas-GCN	81.11	83.15	83.42	0.0	84.14	99.35
GraphGPS	80.48	82.89	82.74	0.0	83.07	98.88
TaG-Net	84.16	86.78	85.85	0.4	86.08	99.53
TopoMatch	**89.01**	**91.13**	**89.55**	**3.6**	**90.05**	**99.87**
TopBrain						
XGBoost	64.60	68.54	64.30	0.0	70.81	97.54
GNN-ART-LABEL	49.17	51.02	49.66	0.0	62.85	93.24
Atlas-GCN	59.05	61.20	59.52	0.0	68.53	98.08
GraphGPS	61.73	65.48	62.96	0.0	69.30	97.08
TaG-Net	65.49	68.97	66.42	0.0	**72.29**	**98.11**
TopoMatch	**67.82**	**72.68**	**69.36**	0.0	70.39	97.34
ImageCAS-X						
XGBoost	63.59	64.89	63.98	8.8	70.08	94.91
GNN-ART-LABEL	76.43	78.43	76.08	40.0	84.98	98.69
Atlas-GCN	48.30	45.52	51.62	0.0	41.39	97.28
GraphGPS	74.10	74.82	75.71	27.5	79.85	97.70
TaG-Net	77.62	77.99	79.05	33.1	82.63	98.52
TopoMatch	**87.61**	**88.51**	**86.53**	**64.4**	**91.89**	**99.86**

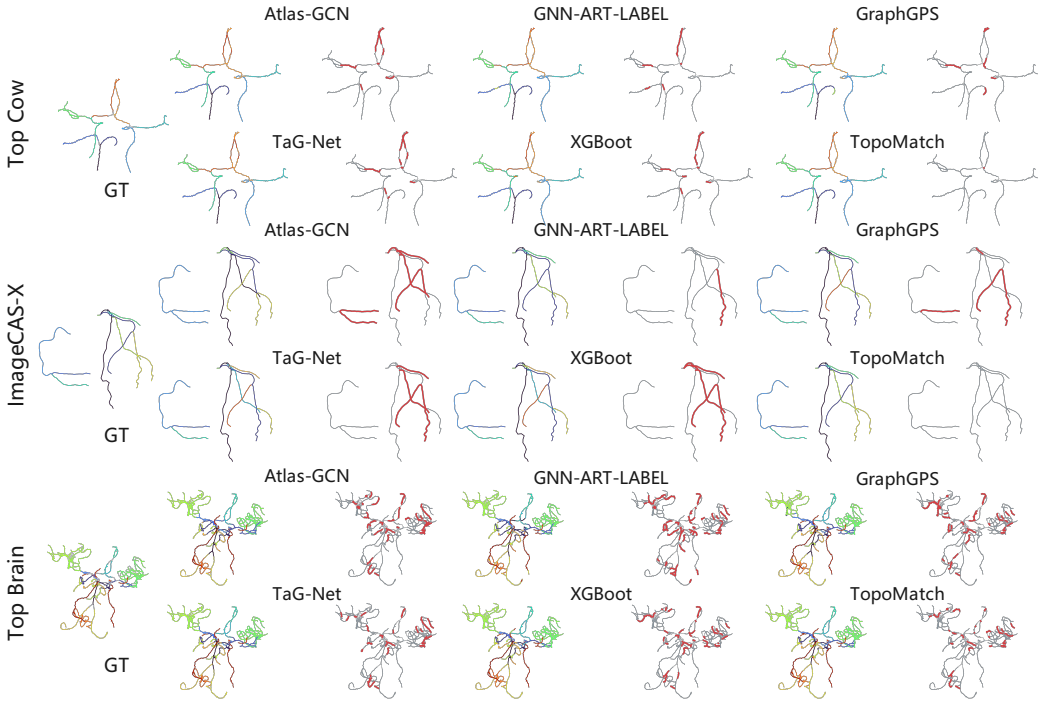

Figure 3: Qualitative comparisons on TopCoW, ImageCAS-X, and TopBrain. Colors denote anatomical labels; paired views overlay errors in red. TopBrain retains residual errors despite improved branch recognition.

5.3 Mechanistic Analysis of Structured Matching

Table 5: Core mechanism ablations under the same five-fold OOF protocol.

Configuration	Macro-F1	Exact graph	Topology
Independent assignment	94.59	39.2	99.23
Balanced OT	**95.09**	38.8	99.27
Line-graph encoder	94.08	38.8	**99.49**
No topology cost	94.27	35.2	99.04
Full TopoMatch	94.72	**42.0**	99.48

Representation, marginal constraints, and topology. In Table 5, a line-graph encoder reduces Macro-F1 from 94.72% to 94.08% and exact-graph accuracy from 42.0% to 38.8%; removing topology cost lowers exact-graph accuracy to 35.2%. Balanced OT has higher clean Macro-F1 (95.09%), but lower exact-graph accuracy (38.8% versus 42.0%). Thus semi-balanced matching is not uniformly superior on clean data. A shuffled compatibility template reaches 36.8% exact-graph accuracy versus 42.0% with anatomical alignment, supporting class-pair compatibility beyond generic smoothing (Appendix E).

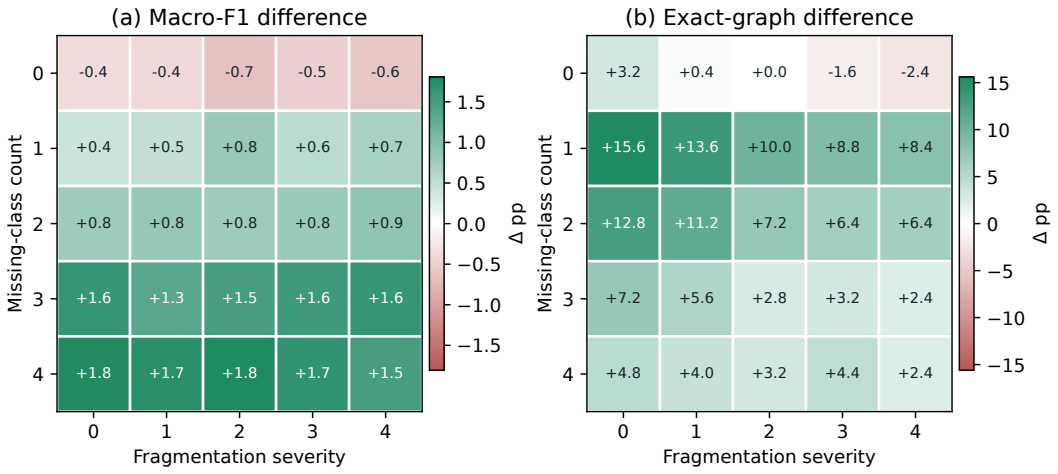

Figure 4: Semi-balanced minus balanced matching under missing anatomy and fragmentation. Entries are percentage-point differences in Macro-F1 (left) and exact-graph accuracy (right). Relaxed target mass helps primarily when expected classes are missing.

Missing anatomy versus fragmentation. Figure 4 tests relaxed target marginals directly. Without missing classes, Macro-F1 differences are negative (-0.4 to -0.7 points), and exact-graph differences vary in sign. With one to four classes missing, semi-balanced matching improves both metrics at every fragmentation level: Macro-F1 gains are 0.4–1.8 points and exact-graph gains are 2.4–15.6 points. Missing anatomy, not fragmentation alone, therefore provides the clearest support for marginal relaxation. Appendix E gives the full grid.

Prediction trajectories. Representative local–structured–fused prediction trajectories are shown in Appendix G. They illustrate three complementary behaviors: branch-level correction by structured matching, preservation of reliable local evidence after fusion, and residual structured failure.

Sparse decoder intervention. The three-seed OOF probability aggregation in Appendix D makes the decoder's role explicit: it changes only 31 of 23,001 branches. Macro-F1 rises from 95.68% to 95.76%, and exact-graph accuracy from 43.6% to 46.4%. This is sparse structural correction rather than a global rewriting of predictions; the aggregated evaluation is separate from the single-seed ablation in Table 5.

5.4 Open-World Recognition and Corruption Robustness

Table 6: Open-set recognition averaged over five strict leave-one-class-out tasks.

Method	Unknown AUPRC	Artifact AUPRC	OSCR
MSP rejection	18.90	95.40	91.40
Energy rejection	21.60	96.10	91.85
Single free slot	11.02	10.51	92.65
Dual free slots	**33.27**	**99.96**	**92.80**

Two rejection causes. Across five strict LOCO tasks, dual slots reach 33.27% unknown-anatomy and 99.96% synthetic-artifact AUPRC (Table 6). OSCR rises only from 92.65% to 92.80% over a single slot, so the main gain is cause-specific rejection. MSP and energy each use one generic score for both positive sets. Unknown-anatomy AUPRC varies from 5.19% (3rd-A2) to 61.42% (BA); near-perfect artifact detection concerns synthetic insertions, not all real extraction errors. Appendix E gives the complete class-wise comparison.

Table 7: Effect of corruption training on clean recognition and robustness.

Training	Clean Macro-F1	Mean RAUC
Without corruption training	**94.81**	87.13
Full TopoMatch	94.72	**92.82**

Robustness versus clean accuracy. Table 7 shows that corruption training changes clean Macro-F1 from 94.81% to 94.72%, but raises aggregate RAUC from 87.13% to 92.82%. Appendix E further shows that branch subdivision and spurious branches retain 99.27% and 99.10% normalized RAUC, whereas geometry perturbation falls to 95.96%. Connectivity-changing edits are less stable than simple subdivision, so robustness to re-segmentation does not imply recovery from altered connectivity. The main-table RAUC of 92.82% is obtained by averaging the six corruption-family RAUCs within each of the five seed-42 OOF test folds, followed by averaging across the five folds. In contrast, the supplementary $93.03 \pm 1.63\%$ summarizes the per-run mean RAUC over 15 fold–seed runs (five folds across seeds 41–43), with the uncertainty reported as the population standard deviation across these runs.

Runtime and calibration. TopoMatch requires 137.3 ms per graph, compared with 15.4 ms for the adapted TaG-Net. The additional latency reflects the full structured inference pipeline, including matching and conditional decoding. TopoMatch has lower AURC (0.207 versus 0.463), but not ECE (0.77 versus 0.58); baselines use native probabilities rather than the same temperature scaling, so these are descriptive diagnostics (Appendix D).

6 Limitations and Conclusion

Limitations. The modest three-seed gain over matched GraphGPS does not establish statistical significance. Unknown-anatomy detection is class-dependent, artifact evaluation is synthetic, and unresolved annotations are excluded. Cross-benchmark results require retraining, and TopBrain remains unsolved at the whole-graph level. Topology consistency is measured against a training-derived compatibility template and is therefore not an independent anatomical ground-truth metric; structural consistency alone does not guarantee anatomical correctness. Inference is also slower, and these results do not establish clinical reliability.

Conclusion. TopoMatch couples reversal-invariant curves and physical junctions with semi-balanced open-world matching. It explains each observed branch while allowing known-class mass to vary and separating unknown anatomy from artifacts. Missing-anatomy stress tests provide the clearest evidence for relaxed target mass, complemented by corruption robustness and task-adapted branch-recognition gains. Reliable novel-anatomy recognition and large-graph recovery remain open challenges.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, prose editing, and LaTeX formatting. The scientific content, methods, experimental results, numerical values, and visual materials were produced by the authors and were not generated by AI. No new experiments or scientific results were produced through this assistance, and all AI-assisted edits were reviewed by the authors.

ETHICS STATEMENT

This work studies anatomical parsing using the vessel-graph benchmarks described in Section 4 and Appendix C. Predictions can mislabel rare anatomy or yield structurally coherent errors, and the reported experiments do not establish clinical safety or suitability for autonomous diagnosis. Any downstream clinical use would require independent validation and appropriate expert oversight. Data use remains subject to the original datasets' access conditions and applicable governance requirements.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model representation, semi-balanced structured matching, adaptive fusion, open-set training, and conditional decoding. Appendix B provides graph-construction rules, feature definitions, prior estimation, solver settings, training objectives, optimization details, and CP-SAT decoding parameters. Appendix C documents the benchmark splits, task-adapted baseline implementations, evaluation metrics, robustness protocol, and strict open-set evaluation. Appendices D, E, F, and G provide multi-seed analyses, class-wise results, mechanism and corruption studies, open-set decompositions, paired cross-benchmark statistics, and qualitative prediction trajectories. All class-mass and compatibility priors are estimated from training partitions only, while calibration parameters are fitted using validation patients only.

REFERENCES

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1563–1572, 2016.

Kit M. Bransby, Esther Øksnebjerg, Kristoffer Kjær, Jacob Kirkeby, Yasmin El Youssef, Aïda Jiménez, Philip R. Pedersson, Martina C. de Knegt, Klaus F. Kofoed, and Rasmus R. Paulsen. ImageCAS-X: A dataset and benchmark for validating coronary vessel segmentation and centerline extraction in computed tomography angiography. *arXiv preprint arXiv:2608.30404*, 2026. doi: 10.48550/arXiv.2608.30404. URL https://arxiv.org/abs/2608.30404.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229, 2020. doi: 10.1007/978-3-030-58452-8_13.

Li Chen, Thomas Hatsukami, Jenq-Neng Hwang, and Chun Yuan. Automated intracranial artery labeling using a graph neural network and hierarchical refinement. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*, pp. 76–85. Springer, 2020. doi: 10.1007/978-3-030-59725-2_8.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are AllSet: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hpBTIv2uy_E.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Gregory J. Dehmer, Vinay Badhwar, Edmund A. Bermudez, Joseph C. Jr. Cleveland, Mauricio G. Cohen, Richard S. D'Agostino, T. Bruce Jr. Ferguson, Robert C. Hendel, Maria Lizza Isler, Jeffrey P. Jacobs, Hani Jneid, Alan S. Katz, Thomas M. Maddox, and David M. Shahian. 2020 aha/acc key data elements and definitions for coronary revascularization: A report of the american college of cardiology/american heart association task force on clinical data standards. *Journal of the American College of Cardiology*, 75(16):1975–2088, 2020. doi: 10.1016/j.jacc.2020.02.010.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1574–1583, 2019.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Nils Hampe, Sanne G. M. van Velzen, Jelmer M. Wolterink, Carlos Collet, José P. S. Henriques, Nils Planken, and Ivana Išgum. Graph neural networks for automatic extraction and labeling of the coronary artery tree in ct angiography. *Journal of Medical Imaging*, 11(3):034001, 2024. doi: 10.1117/1.JMI.11.3.034001.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475. Curran Associates, Inc., 2020.

Agata Mosinska, Pablo Marquez-Neila, Mateusz Koziński, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3136–3145, 2018.

Fabio Musio, Norman Juchler, Kaiyuan Yang, Suprosanna Shit, Chinmay Prabhakar, Bjoern Menze, and Sven Hirsch. Circle of willis centerline graphs: A dataset and baseline algorithm. *Neuroscience Informatics*, 6(1):100265, 2026. doi: 10.1016/j.neuri.2026.100265.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7673–7682, 2020.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas 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*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017b. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d8bf84be3800d12f74d8b05e9b89836f-Paper.pdf.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14501–14515. Curran Associates, Inc., 2022. doi: 10.52202/068431-1054.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4938–4947, 2020.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank–wolfe. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4995–5021, 2022.

Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. clDice: A novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16560–16569, 2021.

Giles Tetteh, Velizar Efremov, Nils D. Forkert, Matthias Schneider, Jan Kirschke, Bruno Weber, Claus Zimmer, Marie Piraud, and Bjoern H. Menze. DeepVesselNet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes. *Frontiers in Neuroscience*, 14: 592352, 2020. doi: 10.3389/fnins.2020.592352.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Iris N. Vos, Ynte M. Ruigrok, Ishaan R. Bhat, Kimberley M. Timmins, Birgitta K. Velthuis, and Hugo J. Kuijf. Graph convolutional networks for automated intracranial artery labeling. *Journal of Medical Imaging*, 11(1):014007, 2024. doi: 10.1117/1.JMI.11.1.014007.

Tao Wang, He Liu, Yidong Li, Yi Jin, Xiaohui Hou, and Haibin Ling. Learning combinatorial solver for graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7568–7577, 2020.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, 38(5):146:1–146:12, 2019. doi: 10.1145/3326362.

Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, and Weidong Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 915–924, 2021.

Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin. Gromov–wasserstein learning for graph matching and node embedding. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6932–6941. PMLR, 2019a. URL https://proceedings.mlr.press/v97/xu19b.html.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019b.

Kaiyuan Yang, Fabio Musio, Yihui Ma, Norman Juchler, Johannes C. Paetzold, Rami Al-Maskari, et al. The TopCoW challenge—topology-aware circle of willis segmentation for CT and MR angiography. *NEJM AI*, 3(8), 2026a. doi: 10.1056/AIdbp2500994.

Kaiyuan Yang, Pengcheng Shi, Houjing Huang, Fabio Musio, Hakim Baazaoui, Orhun Utku Aydin, Adam Hilbert, Rachika E. Hamadache, Cansu Yalcin, Minghui Zhang, et al. TopBrain segmentation challenge for whole brain vessel anatomy. *medRxiv*, 2026b. doi: 10.64898/2026.05.28.26354312. URL https://www.medrxiv.org/content/10.64898/2026.05.28.26354312v1.

Linlin Yao, Feng Shi, Sheng Wang, Xiao Zhang, Zhong Xue, Xiaohuan Cao, Yiqiang Zhan, Lizhou Chen, Yuntian Chen, Bin Song, Qian Wang, and Dinggang Shen. TaG-Net: Topology-aware graph network for centerline-based vessel labeling. *IEEE Transactions on Medical Imaging*, 42(11): 3155–3166, 2023. doi: 10.1109/TMI.2023.3240825.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, pp. 28877–28888, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yaxin Zhu, Peisheng Qian, Ziyuan Zhao, and Zeng Zeng. Deep feature fusion via graph convolutional network for intracranial artery labeling. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 467–470. IEEE, 2022. doi: 10.1109/EMBC48229.2022.9871848.

A Extended Related Work

A.1 Anatomical Vessel Graph Labeling

Anatomical vessel labeling has been approached with geometric rules, atlas registration, and, more recently, graph neural networks. Chen et al. (2020) represent intracranial arteries as an attributed relational graph, jointly classify nodes and edges, and apply hierarchical refinement based on anatomical relations. Zhu et al. (2022) replace the multilayer perceptron backbone with stacked graph convolutions and fuse intermediate features across scales. Vos et al. (2024) investigate graph convolutional operators and atlas-derived spatial features for CoW labeling, demonstrating the value of higher-order structure and anatomical coordinates. These studies show that graph learning can improve over labeling vessel elements independently. However, they are primarily formulated around a predefined anatomical vocabulary and assume that the observed graph can be mapped directly to this vocabulary. They therefore do not explicitly address the simultaneous presence of missing anatomical classes, multiple fragments corresponding to the same vessel, unseen anatomical branches, and non-anatomical extraction artifacts.

Related methods couple centerline reasoning with volumetric segmentation. TaG-Net Yao et al. (2023) constructs a topology-aware graph from vessel centerlines, preserves key topological points during sampling, aggregates point and graph features at multiple scales, and propagates predicted centerline labels back to a vessel mask. Its topology-aware sampling improves geometric representation, while its vessel-completion stage mitigates errors in the initial segmentation. The paired CTA/MRA CoW centerline graphs Musio et al. (2026), derived from the TopCoW benchmark Yang et al. (2026a), enable evaluation of anatomical parsing under a common patient-level protocol. The associated centerline-extraction baseline focuses on topology reconstruction from semantic masks rather than label-blind anatomical assignment of already observed branches. Existing centerline-labeling pipelines therefore remain largely tied to a fixed segmentation or label space and do not directly address parsing when the observed topology is incomplete, fragmented, or contaminated by spurious branches.

Topology-aware vessel extraction is complementary to semantic graph parsing. DeepVesselNet jointly studies volumetric vessel segmentation, centerline prediction, and bifurcation detection Tetteh et al. (2020); topology-preserving losses based on persistent homology or differentiable skeleton overlap penalize broken tubular structures Hu et al. (2019); Mosinska et al. (2018); Shit et al. (2021). These methods improve the structural quality of the graph produced by an upstream image-processing pipeline, but do not solve the subsequent semantic assignment problem once an imperfect graph has been extracted. Anatomical graph parsing must instead operate directly on variable vessel graphs while tolerating missing anatomy, branch fragmentation, and observations that should not be forced into the known anatomical vocabulary.

A.2 Representation Learning on 3D Curves and Junctions

Generic 3D representation learning commonly treats geometry as an unordered point set. PointNet and PointNet++ use symmetric aggregation and hierarchical metric neighborhoods Qi et al. (2017a;b); DGCNN dynamically recomputes feature-space neighborhoods Wang et al. (2019), while CurveNet learns guided walks through a point cloud and pools the resulting curve features Xiang et al. (2021). Deep Sets and Set Transformer establish general permutation-invariant pooling and attention constructions Zaheer et al. (2017); Lee et al. (2019), building on the broader Transformer architecture Vaswani et al. (2017). These models provide general geometric encoders, but vessel centerlines have structure that differs from generic point clouds: samples along a branch have an intrinsic sequential order, the stored orientation of the polyline is arbitrary, and multiple curves may terminate at the same physical junction. Treating these observations as unordered points or relying only on augmentation does not explicitly encode the reversal symmetry of a vessel curve.

At the graph level, graph convolution, attention, and message passing provide standard local operators Kipf & Welling (2017); Veličković et al. (2018); Gilmer et al. (2017); Xu et al. (2019b). Graph Transformers extend their receptive field through global attention; Graphormer introduces structural encodings Ying et al. (2021), while GraphGPS combines local propagation and global attention in a modular architecture Rampášek et al. (2022). These developments belong to the broader program of geometric deep learning Bronstein et al. (2021). A vessel junction, however, may connect more

than two branches, and reducing such a junction to independent pairwise edges can obscure the fact that all incident branches participate in the same physical event. Hypergraph models address an analogous issue by alternating aggregation between vertices and hyperedges; AllSet Chien et al. (2022) formalizes such layers as compositions of learnable multiset functions.

For anatomical vessel parsing, the representation must simultaneously preserve ordered curve geometry, remove the arbitrary choice of polyline traversal direction, and retain physical multi-branch junctions as shared structural entities. Existing point-cloud, line-graph, and generic graph-transformer formulations do not jointly impose these three properties, creating a mismatch between generic geometric representation learning and the structural invariances of vessel graphs.

A.3 Structured Matching and Open-Set Recognition

Differentiable assignment has become a standard mechanism for structured correspondence. Entropic regularization enables efficient Sinkhorn optimization of transport plans Cuturi (2013); modern treatments cover its computational and statistical foundations Peyré & Cuturi (2019); Genevay et al. (2019), with practical implementations provided by the POT library Flamary et al. (2021). SuperGlue combines graph attention with optimal transport and a dustbin state for feature matching Sarlin et al. (2020), while DETR uses bipartite matching for set prediction Carion et al. (2020). Graph matching based on Gromov–Wasserstein discrepancy incorporates relational structure Xu et al. (2019a), and trainable combinatorial solvers learn assignments under one-to-one constraints Wang et al. (2020). These formulations are primarily designed to establish correspondences between two observed sets or graphs. Anatomical vessel parsing differs because the target is a reusable semantic vocabulary rather than another observed instance, multiple extracted branches may correspond to the same anatomical class, and some expected classes may be absent altogether. Balanced permutation-style matching and fixed target marginals are therefore ill-suited to variable anatomical graphs.

Unbalanced optimal transport relaxes marginal constraints and permits the creation or removal of mass Chizat et al. (2018); subsequent analyses characterize Sinkhorn convergence and accelerated translation-invariant updates Pham et al. (2020); Séjourné et al. (2022). These formulations provide a principled way to relax correspondence mass, but they do not by themselves distinguish between different causes of unmatched observations or incorporate anatomy-dependent compatibility between neighboring vessel branches. In particular, a vessel parser must explain every observed branch while allowing the total assignment to individual known anatomical classes to vary with class visibility and fragmentation.

Open-set and out-of-distribution recognition address a complementary problem by identifying samples that do not belong to the known training categories. Representative approaches include confidence baselines, OpenMax, Mahalanobis scoring, outlier exposure, and energy-based detection Hendrycks & Gimpel (2017); Bendale & Boult (2016); Lee et al. (2018); Hendrycks et al. (2019); Liu et al. (2020). Probability calibration, deep ensembles, and selective classification further address uncertainty and reliability Guo et al. (2017); Lakshminarayanan et al. (2017); Geifman & El-Yaniv (2017). These methods generally operate on samples independently and collapse different rejection causes into a common notion of "unknown." For vessel graphs, however, an anatomically plausible but previously unseen branch is qualitatively different from a spurious extraction artifact, and both decisions are coupled to neighboring branch assignments through vascular topology. The resulting problem requires structured anatomical assignment to accommodate variable class mass, distinguish multiple rejection causes, and enforce graph-level compatibility without reverting to a fixed one-to-one anatomical template.

B Graph Construction and Implementation Details

This section collects the graph-construction, representation, prior-estimation, and optimization details reported in the manuscript. The central assignment formulation remains in Section 3.

B.1 Label-Blind Graph Construction and Geometry

After deduplicating undirected centerline edges, nodes with degree other than two delimit maximal topological chains. For CoW-VTP and the mask-derived TopCoW-mask and TopBrain graphs,

semantic labels do not determine branch boundaries. The normalization scale is the diagonal length of each graph's axis-aligned bounding box. Consecutive edges of a maximal topological chain are greedily grouped while their cumulative arc length divided by this scale remains at most $\eta = 0.04$; individual edges are not subdivided. Each resulting segment is resampled to $S = 32$ equidistant points. Semantic labels are projected only after the segment boundaries have been fixed. Label purity is defined as the fraction of the original edge length assigned to the length-majority class, and the majority label is retained when purity is at least 0.8. Otherwise, the segment is marked as *unresolved* and excluded from label-dependent supervision and evaluation. The *unresolved* annotation is neither an output category nor unknown-anatomy supervision. ImageCAS-X instead uses its released centerline polylines directly as branches and does not apply the η-based length subdivision.

For each graph, coordinates are centered using the mean of its original centerline points and scaled by the diagonal length of the corresponding axis-aligned bounding box. A scale of 1 is used for degenerate boxes. Pointwise curve features comprise centered and normalized 3D coordinates, normalized arc positions, unit tangents, curvature magnitudes, and radii. Reversal changes sample order, maps s to $1 - s$, and negates tangent direction before applying the same curve encoder.

The invariant curve token is combined with a branch descriptor containing the centroid, symmetric endpoint features, arc and chord lengths, tortuosity, radius statistics, and endpoint degrees. Junction descriptors contain location, degree, incident-branch statistics, radius statistics, distance to the graph center, and endpoint/bifurcation indicators. Within each incidence layer, junctions aggregate incident branches, and branches receive endpoint-junction and shared-endpoint branch context. Separate residual MLPs update both representations; multi-head self-attention supplies global graph context.

B.2 Semantic Costs, Priors, and Structured Refinement

The local prediction combines logits from contextual graph features, initial curve/descriptor tokens, and a direct forward/reverse geometric path. Let $\mathbf{h}_i \in \mathbb{R}^{128}$ denote the contextual branch embedding produced by the incidence-aware encoder. A learned linear projection maps $\mathbf{h}_i$ to a 128-dimensional prototype space. Each known class is associated with a learnable 128-dimensional prototype initialized from $\mathcal{N}(\mathbf{0}, 0.02^2\mathbf{I})$. The squared Euclidean distance between the projected branch embedding and the corresponding class prototype is added to the known-class matching cost with weight 0.05. Unknown and artifact categories rely on their local probabilities and do not use prototype-distance terms.

Known-class target mass multiplies the graph-level sigmoid visibility prediction by a prior estimated from resolved branch-length shares in the corresponding training fold, with 10^{-6} added for numerical stability. Junction compatibility is estimated from training-fold class-pair co-occurrences with one pseudocount, followed by row-wise maximum normalization, elementwise-maximum symmetrization, and unit diagonal assignment. No test or validation labels enter these priors.

The solver uses entropy weight $\varepsilon = 0.08$, marginal penalty $\tau = 1.0$, and topology weight $\lambda_T = 0.2$. An initial 25-iteration log-domain Sinkhorn solve is followed by three topology-feedback rounds, each with 25 additional Sinkhorn iterations. At each round, the expected incompatibility of neighboring branches updates the known-class costs, while unknown and artifact assignments remain exempt from anatomical-neighborhood constraints. The junction term is normalized by the number of eligible incident branch pairs. Damped refinement follows Equation 4 with $\alpha = 0.5$.

B.3 Optimization, Supervision, and Inference

TopoMatch uses hidden dimension 128, two incidence-aware encoder layers, four attention heads, and dropout 0.1. AdamW training runs for at most 200 epochs with batch size 8, initial learning rate 5×10^{-4}, weight decay 10^{-4}, and gradient-norm clipping at 1. Early stopping has a patience of 30 epochs. Model selection maximizes 0.85 validation Macro-F1 plus 0.15 validation exact-graph accuracy.

The objective in Equation 3 defines the inner structured assignment used to construct $\mathbf{Q}$, whereas the following objective trains the network parameters. Let $\mathcal{S}$ denote the set of supervised branches after augmentation, including resolved real branches, labeled synthetic artifacts, and episodic pseudo-

unknowns. The training objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{parse}} + 0.75\mathcal{L}_{\text{local}} + 0.20\mathcal{L}_{\text{vis}} + 0.50\mathcal{L}_{\text{art}} + 0.02\mathcal{L}_{\text{Brier}} \\ + 0.10\mathcal{L}_{\text{junc}} + 0.05\mathcal{L}_{\text{energy}} + 0.02\mathcal{L}_{\text{gate}} + 0.10\mathcal{L}_{\text{topo}} + 0.15\mathcal{L}_{\text{hard}}.\end{aligned} \tag{6}$$

The parsing term is the weighted negative log-likelihood of the fused posterior over $\mathcal{S}$. For known classes, the weights are proportional to the inverse square root of the training frequency, normalized to have mean one and clipped to $[0.25, 4]$; both free slots use unit weight. $\mathcal{L}_{\text{local}}$ is weighted cross-entropy on the local logits with label smoothing of 0.02. The visibility and artifact terms are binary cross-entropies against graph-level known-class presence and branch-level artifact targets, respectively. $\mathcal{L}_{\text{Brier}}$ averages squared probability error over supervised branches and output classes. $\mathcal{L}_{\text{junc}}$ is cross-entropy over eight coarse junction types defined by the known-class configuration incident to each physical junction: an endpoint with at most one incident known-class branch, a degree-two pass-through configuration, five canonical anatomical triplet configurations, and a residual "other" type covering all remaining configurations, and $\mathcal{L}_{\text{gate}} = N^{-1}\sum_i (g_i - 0.25)^2$.

With $E_i = -\log\sum_{k=1}^{K}\exp(\ell_{ik}^{\text{local}})$, the energy term averages $\text{ReLU}(E_i + 5)^2$ over known branches and $\text{ReLU}(-2 - E_i)^2$ over episodic pseudo-unknown branches; an empty subset contributes zero. If ℓ_i^{parse} denotes the weighted per-branch parsing NLL, the hard-branch term is $\mathcal{L}_{\text{hard}} = 0.25\left[\log\sum_{i\in\mathcal{S}}\exp(\ell_i^{\text{parse}}/0.25) - \log|\mathcal{S}|\right]$. $\mathcal{L}_{\text{topo}}$ is the expected known-class pair incompatibility, normalized by the number of eligible incident branch pairs. It is included in the outer network-training objective, whereas $\mathcal{L}_{\text{junction}}$ denotes the topology term inside the inner structured-assignment objective in Equation 3. Accordingly, the matching coefficient $\lambda_T = 0.2$ and the outer training weight 0.10 are distinct.

The branch-wise gate is $g_i = \sigma(\mathbf{w}^\top \mathbf{h}_i + b)$, with $\mathbf{w} = \mathbf{0}$ and $b = \text{logit}(0.25) \approx -1.099$ at initialization, yielding $g_i = 0.25$.

The open-set curriculum progressively increases graph corruption and episodic class holdout to final probabilities 0.2 and 0.1, respectively. Edits comprise deletion, connected subdivision, spurious-branch insertion, false bridges, junction split/merge, and geometric perturbation. Newly inserted branches are artifacts, whereas an episodically selected known class is pseudo-unknown. The strict LOCO evaluation excludes its test-unknown class throughout training, rather than merely holding it out within an episode.

A scalar temperature is fitted using validation patients. The slack-penalized CP-SAT decoder runs by default at inference when calibrated predictions contain an incompatible known-class pair. It minimizes negative log-probability together with a penalty on incompatible pairs; slack variables preserve feasibility. It imposes neither class-capacity nor visibility constraints, and compatible predictions remain unchanged. The primary model has 1.155M trainable parameters and reported latency of approximately 0.137 s per graph including structured inference (Table A6). The same architecture and optimization settings are used across benchmarks; only the vocabulary and training-fold statistics are adapted.

A pair of known labels c and d incident to the same junction is considered incompatible when $T_{cd} < 0.1$. The slack-based CP-SAT objective sums branch-wise negative log-probabilities and assigns a penalty of 0.2 to each incompatible incident branch pair. Objective costs are scaled by 10^4 and rounded to integers before optimization. The solver is limited to 0.5 s per graph. If its status is neither feasible nor optimal, decoding falls back to the branch-wise argmax of the calibrated fused probabilities.

C Benchmarks, Baselines, and Evaluation Protocol

Benchmarks. We evaluate TopoMatch on four vessel-graph parsing benchmarks with complementary sources of distribution shift, as summarized in Table A1. CoW-VTP is our primary benchmark and is built from the public CoW Centerline Graphs release Musio et al. (2026), containing paired CTA and MRA acquisitions from 125 patients and 250 expert-corrected centerline graphs. Its label vocabulary contains 13 cerebral artery classes. TopCoW-mask is constructed by skeletonizing the multiclass vessel masks from the same 125-patient cohort and projecting the same 13-class vocabulary, providing a controlled evaluation of sensitivity to the graph-construction pipeline while keeping the patient population and anatomical label space fixed. TopBrain Yang et al. (2026b) contains 25

paired CTA/MRA patients with 36 vascular classes and substantially larger graphs, introducing both a more complex label space and more challenging graph topology. ImageCAS-X Bransby et al. (2026) contains labeled coronary centerlines from 800 CCTA scans with 14 segment categories, extending the evaluation from cerebral to coronary anatomy and therefore introducing a distinct vascular territory. Together, these benchmarks probe graph construction, graph scale, label-space complexity, and anatomical domain. Each method is retrained separately on every benchmark to evaluate task-adapted generality.

Evaluation protocol. CoW-VTP, TopCoW-mask, and TopBrain are evaluated using patient-disjoint five-fold cross-validation, with paired CTA and MRA graphs from the same patient always assigned to the same partition. For CoW-VTP, each fold contains 80 training, 20 validation, and 25 test patients, such that the out-of-fold (OOF) evaluation covers all 125 patients and 250 graphs exactly once. ImageCAS-X follows its official train/validation/test split. To avoid information leakage, class-mass priors and junction-compatibility tables are estimated exclusively from the corresponding training partition, whereas calibration parameters are fitted using validation patients only. Geometric normalization is performed independently for each input graph.

Robustness protocol. We evaluate six controlled graph perturbations: segment deletion, connected branch subdivision, spurious branch insertion, false bridge insertion, junction perturbation, and geometric perturbation. For each perturbation family, let F_s denote known-class Macro-F1 at severity level S_s, where $s \in \{0, 1, 2, 3, 4\}$ and S_0 denotes the uncorrupted graph. Using the equally spaced severity axis $x_s = s$, we compute

$$\mathrm{RAUC} = \frac{1}{4} \sum_{s=0}^{3} \frac{F_s + F_{s+1}}{2}. \tag{7}$$

The factor $1/4$ normalizes by the severity range, so RAUC represents severity-averaged Macro-F1 on the same percentage scale as F_s. We average the six family-wise RAUC values to obtain the aggregate RAUC reported in the main comparison. For each perturbation family, normalized RAUC is reported as $100 \times \mathrm{RAUC}/F_0$, measuring robustness relative to its uncorrupted performance.

Open-set protocol. Open-set recognition is evaluated using strict leave-one-class-out training for BA, R-PCA, R-Pcom, Acom, and 3rd-A2. The held-out anatomical class is completely excluded from training and is treated as unknown at test time. Artifact recognition is evaluated separately using synthetically inserted spurious branches and false bridges. We therefore distinguish unknown anatomical structures from synthetic extraction artifacts and report their performance separately.

Table A1: Statistics of the four vessel-graph parsing benchmarks after graph preparation. Coverage reports branch/length coverage.

Benchmark	Setting	Classes	Graphs / Patients	Branches	Coverage (%)
CoW-VTP	CoW / curves	13	250 / 125	23,001	94.05 / 93.75
TopCoW-mask	CoW / masks	13	250 / 125	33,783	94.01 / 93.26
TopBrain	Brain / masks	36	50 / 25	25,255	92.83 / 90.87
ImageCAS-X	Coronary / curves	14	800 / 800	11,367	99.31 / 99.97

C.1 Baselines and Controls

Baselines. We evaluate five task-adapted baselines on the same branch graphs, targets, and patient splits; these adaptations are not intended as exact reproductions of the original systems. XGBoost Chen & Guestrin (2016) classifies standardized branch descriptors using up to 1,200 boosting rounds with maximum tree depth 7 and validation early stopping after 60 rounds. The GNN-ART-LABEL Chen et al. (2020) adaptation uses 16-dimensional node, edge, and global latent states with 10 encode–process–decode steps. Atlas-GCN Vos et al. (2024) uses two graph-convolution layers with 64 hidden channels together with training-fold spatial atlases. The GraphGPS Rampášek et al. (2022) adaptation uses three GPS layers with GCN-based local message passing, 128 hidden channels, four attention heads, and dropout 0.1. The graph-only TaG-Net Yao et al. (2023) adaptation uses a

128-dimensional PointNet curve embedding and aggregates branch-graph features over 1-, 2-, 4-, and 8-hop neighborhoods, with dropout 0.1.

The neural adaptations are trained for at most 200 epochs with an effective batch size of 8 using AdamW with learning rate 5×10^{-4} and weight decay 10^{-4}. Gradients are clipped at norm 5, and a ReduceLROnPlateau scheduler reduces the learning rate by a factor of 0.5 after five validation epochs without improvement. Early stopping uses a patience of 30 epochs, and checkpoints are selected by validation Macro-F1. The adaptations retain their task-specific architectures and losses and do not use TopoMatch's semi-balanced transport, topology objective, corruption curriculum, temperature calibration, or CP-SAT decoder.

Capacity-matched controls. We additionally include a branch MLP, a message-passing network, and a parameter-matched GraphGPS using the same input representation and training protocol as TopoMatch. To separate the contribution of structured matching from the final discrete decoder, we also apply the same CP-SAT decoder to the matched GraphGPS predictions. These controls isolate gains due to representation capacity, graph encoding, and structured inference.

C.2 Evaluation Metrics

Our primary metrics are Macro-F1 over the task-specific known classes on resolved segments and exact-graph accuracy, which requires all resolved segments in a graph to be correctly labeled. We additionally report length-weighted Macro-F1, segment accuracy, and balanced accuracy for branch-level recognition. To evaluate structural correctness, junction exact accuracy compares the set of predicted labels incident to each junction with the corresponding ground-truth set, while topology consistency measures the fraction of predicted known-label incident pairs that agree with the training-fold compatibility template. Junction exact accuracy is computed directly from branch predictions and is therefore distinct from the auxiliary junction-classification head.

For probabilistic evaluation, all branches with resolved targets are pooled across test graphs. Let $\mathcal{S}_{\text{eval}}$ denote this set and let $y_i \in \{1, \dots, K+2\}$ be the target label of branch i; unresolved branches are excluded. Unknown-anatomy and artifact targets are included when they are present. We report $\text{NLL} = -\frac{1}{|\mathcal{S}_{\text{eval}}|} \sum_{i \in \mathcal{S}_{\text{eval}}} \log \max\big(p_{i,y_i}, 10^{-9}\big)$, and the multiclass Brier score $\text{Brier} = \frac{1}{|\mathcal{S}_{\text{eval}}|} \sum_{i \in \mathcal{S}_{\text{eval}}} \sum_{c=1}^{K+2} \left(p_{ic} - \mathbb{1}[y_i = c]\right)^2$. The reported Brier score sums the squared error over classes before averaging over branches, whereas the Brier term used during training additionally averages over output classes. Calibration and selective prediction are further evaluated using 15-bin expected calibration error (ECE) and area under the risk–coverage curve (AURC; lower is better), respectively. For open-set evaluation, we report area under the precision–recall curve (AUPRC) separately for unknown anatomy and artifacts, together with the open-set classification rate (OSCR). Let the known-class confidence of branch i be $s_i = \max_{1 \le c \le K} p_{ic}$, computed using only the K known-class columns. As a threshold θ is swept over s_i, the correct classification rate (CCR) is the fraction of known branches that are both accepted ($s_i \ge \theta$) and correctly classified, whereas the false positive rate (FPR) is the fraction of open-set branches—unknown anatomy or artifacts—that are accepted as known. OSCR is the area under the resulting CCR–FPR curve, with higher values indicating better joint known-class recognition and open-set rejection. Because OSCR depends on known-class confidence, it does not directly evaluate how rejected probability mass is divided between the unknown-anatomy and artifact slots. Confidence intervals are estimated using 2,000 patient-clustered bootstrap samples, with paired CTA/MRA observations resampled together; paired patient bootstrap is used for method comparisons on identical test cases.

D Extended Primary-Benchmark Analysis

The main paper reports the evidence needed for the central system-level claims. This section provides complementary analyses of class imbalance, optimization stability, decoder behavior, class-wise errors, probabilistic quality, and computational cost.

Additional branch-level metrics. Table A2 reports overall accuracy and balanced accuracy on the primary benchmark. TopoMatch remains highest on both metrics, and the balanced-accuracy result indicates that the branch-level improvement is not restricted to the most frequent anatomical classes.

Table A2: Additional branch-level metrics on CoW-VTP.

Method	Accuracy	Balanced accuracy
XGBoost	90.31	83.88
Atlas-GCN	89.54	85.76
GraphGPS	91.07	88.47
GNN-ART-LABEL	94.15	90.94
TaG-Net (graph)	94.66	92.56
TopoMatch	**97.32**	**94.22**

Matched training-seed stability. Table A3 evaluates the strongest external baseline, the capacity-matched GraphGPS control with the common decoder, and TopoMatch across three training seeds. The capacity-matched GraphGPS remains close to TopoMatch across seeds, while TaG-Net remains clearly lower. Thus, the small clean branch-level gap to the strongest matched control is stable across the reported seeds. These runs do not establish a large capacity-independent advantage or a statistically significant difference from the matched control.

Table A3: Three-seed stability of the strongest external and capacity-matched controls.

Method	Seed 41	Seed 42	Seed 43	Mean $\pm$ Std
TaG-Net	90.84	91.12	91.48	91.15 ± 0.32
Matched GPS + decoder	94.39	94.65	94.81	94.62 ± 0.21
TopoMatch	95.16	94.72	94.78	94.89 ± 0.24

Three-seed prediction aggregation and decoder behavior. After aggregating OOF probabilities from three TopoMatch seeds, Table A4 compares direct argmax prediction with structured decoding. The ECE and AURC values in Table A4 correspond to the three-seed aggregated prediction and are therefore distinct from the single-model diagnostics reported in Table A6. This probability-level ensemble is distinct from the mean of the three per-seed Macro-F1 values reported above. The decoder changes only a small fraction of branches and produces a modest change in branch-level metrics, while its effect is more visible on whole-graph correctness. This is consistent with the decoder acting as a sparse structural correction rather than globally rewriting branch predictions. The decoded prediction has a patient-bootstrap 95% confidence interval of $[94.57, 96.73]$ for Macro-F1, and the decoder changes only 31 of 23,001 branches.

Table A4: Three-seed OOF probability aggregation and effect of CP-SAT decoding.

Prediction	Macro-F1	Length-F1	Acc.	Exact	Topology	ECE / AURC
Raw argmax	95.68	95.73	97.71	43.6	99.36	–
CP-SAT decoder	**95.76**	**95.81**	**97.78**	**46.4**	**99.54**	0.47 / 0.127

Per-class analysis. Table A5 decomposes the final three-seed aggregated prediction by anatomical class. The remaining errors are concentrated in a small subset of low-support or difficult classes rather than being distributed uniformly across the vocabulary. Acom and 3rd-A2 remain the most difficult, with 3rd-A2 primarily recall-limited.

Calibration, selective prediction, and efficiency. Table A6 reveals a clear trade-off between probabilistic prediction quality and computational cost. TopoMatch obtains the lowest NLL and Brier score and, more importantly, the lowest AURC, indicating that its confidence estimates are effective for ranking predictions in selective prediction. Its ECE, however, is not the best among the compared methods, with TaG-Net achieving a lower value; the results therefore do not indicate uniformly superior calibration across all criteria. Moreover, TopoMatch incurs substantially higher inference latency than the graph baselines. The latency increase is much larger than the corresponding increase in parameter count, suggesting that structured matching and decoding contribute materially to the additional computational cost. Because TopoMatch applies validation-only temperature scaling whereas the adapted baselines use their native probabilities, NLL, Brier, and ECE should be interpreted

Table A5: Per-class results after three-seed OOF probability averaging and CP-SAT decoding on CoW-VTP.

Class	Patients	Support	Length	Precision	Recall	F1
BA	125	1,463	4,073.1	98.71	99.52	99.12
R-PCA	125	2,540	7,097.5	99.41	98.98	99.19
L-PCA	125	2,607	7,259.9	99.54	98.77	99.15
R-ICA	125	2,096	5,671.9	99.47	99.19	99.33
R-MCA	124	1,644	4,397.8	98.97	99.51	99.24
L-ICA	125	2,089	5,621.6	99.76	99.66	99.71
L-MCA	125	1,575	4,233.4	99.68	99.81	99.75
R-Pcom	63	440	1,321.3	97.25	96.59	96.92
L-Pcom	57	380	1,125.8	95.08	96.58	95.82
Acom	57	101	200.2	80.17	92.08	85.71
R-ACA	125	3,240	9,097.4	95.10	96.45	95.77
L-ACA	125	3,262	9,091.1	95.78	94.67	95.22
3rd-A2	16	196	562.9	89.81	71.94	79.89

as descriptive probabilistic diagnostics rather than strictly matched calibration comparisons. Overall, the clearest result in this analysis is the favorable risk–coverage behavior of TopoMatch, accompanied by a substantial runtime cost from structured inference.

Table A6: Probabilistic quality, selective prediction, model size, and efficiency on the primary benchmark.

Method	NLL ↓	Brier ↓	ECE ↓	AURC ↓	Params	ms/graph
XGBoost	0.247	0.136	2.00	1.396	–	**2.9**
Atlas-GCN	0.293	0.155	5.73	1.743	0.003M	7.8
GraphGPS	0.288	0.137	2.85	1.928	0.453M	9.4
GNN-ART-LABEL	0.216	0.092	2.06	1.122	0.008M	18.3
TaG-Net	0.136	0.076	**0.58**	0.463	0.243M	15.4
TopoMatch	**0.096**	**0.043**	0.77	**0.207**	1.155M	137.3

E Extended Mechanistic and Open-World Analysis

The main paper presents the compact evidence for semi-balanced matching, junction-aware reasoning, open-world recognition, and corruption robustness. This section provides the corresponding stress-test grids and class-wise diagnostics without repeating the main summary tables.

Missing anatomy and fragmentation. Table A7 reports the numerical grid underlying Fig. 4 in the main paper. The relative benefit of semi-balanced matching is weak when the anatomical inventory is complete, even as fragmentation increases. Once classes are missing, the advantage becomes consistently positive across fragmentation levels. The pattern therefore isolates missing anatomy as the dominant factor motivating relaxed target marginals.

Table A7: Controlled missing-anatomy and fragmentation analysis. Each entry reports ΔMacro-F1 / ΔExact in percentage points, where Δ = semi-balanced − balanced matching.

Missing classes	F_0	F_1	F_2	F_3	F_4
0	$-0.4/+3.2$	$-0.4/+0.4$	$-0.7/+0.0$	$-0.5/-1.6$	$-0.6/-2.4$
1	$+0.4/+15.6$	$+0.5/+13.6$	$+0.8/+10.0$	$+0.6/+8.8$	$+0.7/+8.4$
2	$+0.8/+12.8$	$+0.8/+11.2$	$+0.8/+7.2$	$+0.8/+6.4$	$+0.9/+6.4$
3	$+1.6/+7.2$	$+1.3/+5.6$	$+1.5/+2.8$	$+1.6/+3.2$	$+1.6/+2.4$
4	$+1.8/+4.8$	$+1.7/+4.0$	$+1.8/+3.2$	$+1.7/+4.4$	$+1.5/+2.4$

Anatomy-specific junction compatibility. To separate anatomy-specific compatibility from generic pairwise smoothing, Table A8 compares the aligned compatibility matrix with a class-permuted

control that preserves the matrix structure while disrupting anatomical correspondence. The aligned matrix produces the strongest whole-graph result, whereas shuffling weakens exact-graph recovery without producing a corresponding branch-level benefit. This pattern supports the interpretation that the training-derived class-pair alignment contributes primarily to global anatomical coordination.

Table A8: Control for anatomy-specific junction compatibility.

Compatibility	Macro-F1	Exact graph	Junction exact	Topology
None	94.27	35.2	96.13	99.04
Shuffled $\mathbf{T}$	94.10	36.8	95.91	99.15
Aligned $\mathbf{T}$	94.72	42.0	96.40	99.48

Corruption-wise robustness. Table A9 reports corruption-wise robustness over 15 fold–seed runs (five folds across seeds 41–43), complementing the seed-42 aggregate comparison in the main paper. Branch subdivision and spurious-branch perturbations retain the highest normalized RAUC, whereas geometry perturbation is the most challenging condition. Connectivity-changing edits reduce robustness more than simple subdivision, indicating greater sensitivity to connectivity changes than to representational fragmentation under these controlled perturbations.

Table A9: Corruption-wise robustness over 15 fold–seed runs. Values are reported using the RAUC definition in Section C.

Corruption	RAUC (%)	Normalized RAUC (%)
Deletion	92.14 ± 1.49	97.12
Branch subdivision	94.18 ± 1.51	**99.27**
Spurious branch	94.02 ± 1.62	99.10
False bridge	93.24 ± 1.84	98.27
Junction perturbation	93.55 ± 1.56	98.61
Geometry perturbation	91.05 ± 2.32	95.96
Mean	93.03 ± 1.63	98.05

Class-wise leave-one-class-out behavior. Table A10 decomposes the open-set summary from the main paper across five held-out anatomical classes. Unknown-anatomy detection varies strongly with the withheld class, whereas synthetic artifact detection remains consistently high. The disparity shows that semantic novelty is more class-dependent than the controlled extraction-artifact setting and motivates reporting the two rejection modes separately.

Table A10: Strict leave-one-class-out results for the dual free-slot model. Artifact positives are synthetic.

Held-out anatomy	Unknown AUPRC	Artifact AUPRC	OSCR
BA	61.42	99.98	92.81
R-PCA	51.17	99.95	86.06
R-Pcom	36.71	99.93	92.90
Acom	11.86	99.97	96.29
3rd-A2	5.19	99.98	95.96

Single-versus-dual free slots. Table A11 isolates the role of separating unknown anatomy from extraction artifacts. The dual-slot formulation improves both cause-specific AP measures across every held-out class, while OSCR changes only slightly. The main effect is therefore semantic disentanglement of rejection outcomes rather than a large change in the ranking of known classes.

F Cross-Benchmark Statistical Analysis

The complete six-metric comparison is reported in main-paper Table 4. All methods are retrained for each benchmark with its output vocabulary and training partition. TopCoW-mask changes graph

Table A11: Single-versus-dual free-slot comparison over all five LOCO tasks.

Held-out class	Single Unknown AP	Dual Unknown AP	Single Artifact AP	Dual Artifact AP	Single OSCR	Dual OSCR
BA	14.11	61.42	10.26	99.98	92.69	92.81
R-PCA	20.42	51.17	11.80	99.95	85.88	86.06
R-Pcom	12.39	36.71	10.13	99.93	92.72	92.90
Acom	5.72	11.86	9.96	99.97	96.13	96.29
3rd-A2	2.60	5.19	10.55	99.98	95.82	95.96
Mean	11.02	33.27	10.51	99.96	92.65	92.80

construction within the CoW-VTP cohort rather than introducing independent patients; TopBrain and ImageCAS-X extend graph scale, label vocabulary, and anatomical territory. The main paper discusses the TopBrain structural limitations alongside the full results.

Paired cross-benchmark comparison. Table A12 gives patient-cluster bootstrap confidence intervals for the Macro-F1 difference from TaG-Net, the strongest task-adapted competing method on that metric in each benchmark. These intervals remain above zero, with the smallest separation on TopBrain. They are not confidence intervals against the capacity-matched GraphGPS control. Point differences and confidence intervals are retained at their reported precision.

Table A12: Paired Macro-F1 differences between TopoMatch and TaG-Net across four task-adapted benchmarks. Confidence intervals are obtained by paired patient-cluster bootstrap.

Benchmark	ΔMacro-F1	95% CI
CoW-VTP	$+3.60$	$[+2.09, +5.18]$
TopCoW-mask	$+4.84$	$[+3.73, +6.08]$
TopBrain	$+2.34$	$[+0.54, +3.87]$
ImageCAS-X	$+9.99$	$[+5.78, +14.56]$

G Qualitative Comparisons and Prediction Trajectories

G.1 Local, Structured, and Fused Prediction Trajectories

Figure A1 illustrates two representative prediction trajectories through local classification, structured matching, adaptive fusion, and final decoding. In the upper case, the local and structured predictions each contain one error, which is corrected after adaptive fusion. In the lower case, structured matching corrects the local classification error, and the corrected prediction is preserved through fusion and decoding. Both examples reach zero final errors, illustrating two different ways in which structured inference can refine branch-level predictions.

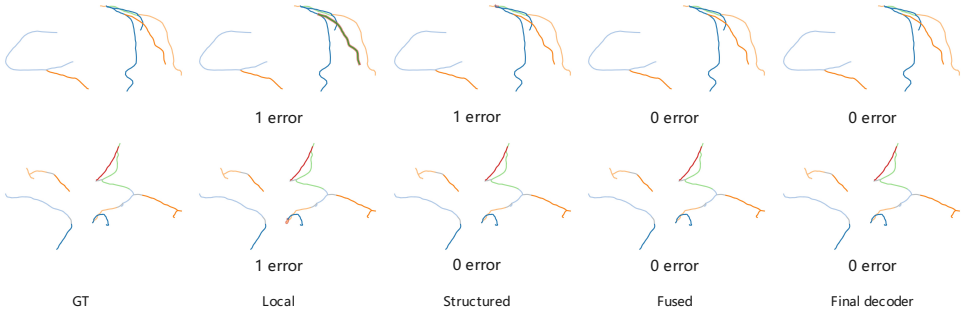

Figure A1: Prediction evolution through TopoMatch. Columns show ground truth, local, structured, fused, and final decoded predictions. The upper example is corrected after fusion, whereas the lower example is corrected during structured matching; both reach zero final errors.