RELATIONAL ALIGNMENT PROFILES: THE GEOMETRY OF GRAPH-LEARNING PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph learning is usually architecture-first: for a new attributed graph, the appropriate computational bias (local smoothing, multihop propagation, heterophily-aware filtering, role-based reasoning, or global attention) is typically discovered only after training several competing models. We ask whether the graph-learning problem itself can be diagnosed first, so that the needed computation can be inferred before model search. We introduce the *Relational Alignment Profile* (RAP), which characterizes how structural, feature, and task geometries align, conflict, and vary across relational scale. RAP provides a compact coordinate for architecture selection and dictionary-free curves that reveal whether relational organization is local, delayed, oppositional, feature-explained, or role-driven. We train RAP-based selectors only on controlled synthetic problems and evaluate them under generator shift and frozen synthetic-to-real transfer. Across 22 real node-classification benchmarks, the label-free native coordinate α is the strongest zero-shot selector among the tested representations, achieving only 2.22 percentage points of regret and outperforming conventional graph statistics. At matched dimension, α also outperforms PCA/NMF compressions of the same native alignment fields. These results support a problem-first view of graph learning: diagnose relational organization before deciding what computation an unseen problem requires.

1 INTRODUCTION

A general-purpose graph learner should not commit to a single way of using graph structure. Some problems reward local smoothing, others require high-pass processing, multihop propagation, structural roles, global attention, or almost no graph propagation at all. Modern graph learning offers all of these computational biases (Wu et al., 2020; Corso et al., 2024), yet given a new attributed graph, deciding which one to use is still typically resolved by training and comparing multiple models (Chami et al., 2022). Expressivity alone does not provide this diagnosis.

The difficulty is fundamental because adjacency has no universal computational meaning. An edge may connect similar nodes, opposing nodes, complementary structural roles, or nodes whose relation is irrelevant to the task. Useful dependence may be local, emerge only after several propagation steps, or be better described by structural equivalence than proximity. The same (G, X) can also support tasks whose labels interact with structure differently. Thus, the relational organization through which information *should* move is itself part of the unknown learning problem.

Existing descriptors capture parts of this organization. Degree and spectral statistics summarize topology; feature smoothness measures structure–attribute compatibility; homophily and label informativeness capture task organization (Platonov et al., 2023; Luan et al., 2023). Graph meta-learning and model-selection methods likewise use dataset characteristics to recommend models (Park et al., 2023; Lee et al., 2024). What is missing is a common problem-side representation of *how structure, attributes, and task interact*: where they agree, conflict, and at what relational scale.

We take a problem-first view. The same nodes carry several relational geometries: adjacency, multiscale propagation, structural role, feature proximity, and, when labels are available, task agreement. Our premise is that a graph-learning problem is characterized by *how these geometries align*. Local agreement, opposition, delayed multihop structure, role correspondence, and diffuse organization are properties of the problem, not the architecture used to process it.

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.

We formalize this idea through the *Relational Alignment Profile* (RAP). For each structural geometry, RAP measures rank-based alignment with feature geometry and, when labels are available, task alignment after accounting for what feature similarity already explains. These multiscale alignment fields record both the form and scale of relational organization. We summarize them by $\gamma(\mathcal{D}) = [\alpha(\mathcal{D}), \beta(\mathcal{D})] \in \mathbb{R}^9$, where the five-dimensional native coordinate α describes organization already present in (G, X) and the four-dimensional task-conditioned coordinate β describes how the prediction task interacts with it. As our experiments show, most transferable architecture-selection signal is already contained in the label-free native component α.

The same construction yields two dictionary-free alignment curves. $\rho_{\text{align}}(t)$ asks whether structure organizes attributes at propagation scale t, while $\rho_{\text{task}}(t)$ asks whether structure carries label signal *beyond what the attributes already explain*. Their sign distinguishes agreement from opposition, and their shape reveals whether organization is local, delayed, diffuse, or scale-dependent. Strong structure–feature alignment therefore need not imply useful task structure, and useful task structure need not be strongly reflected in the attributes.

RAP supplies the missing *problem-side diagnosis*. A future adaptive graph learner could use such a diagnosis to condition among different computational mechanisms; here we test the more basic prerequisite: whether it can predict which computational bias succeeds on an unseen problem. We treat architecture selection as a controlled proxy for problem-conditioned computation, using a bank spanning feature-only, low-pass, high-pass, multihop, role-aware, and local–global processing. This complements work that organizes such choices within a common model space (Navuluru et al., 2026).

We evaluate RAP under increasing distribution shift. Dictionaries and the architecture-performance predictor are learned only from controlled synthetic problems, then tested on disjoint synthetic generators, unseen mixtures, and finally frozen zero-shot transfer to 22 real node-classification benchmarks. No real dataset is used to build the dictionaries or fit the selector. On real graphs, the label-free native coordinate α is the strongest zero-shot selector, achieving Spearman 0.384, Top-3 accuracy 72.7%, and only 2.22 percentage points of regret, close to an in-domain predictor trained on other real datasets. At matched dimension, α also significantly outperforms PCA and NMF compressions of the same native fields ($p = 0.013$).

A useful empirical distinction emerges. The task-conditioned coordinate β adds little to architecture selection, whereas task conditioning is highly informative when read directly through $\rho_{\text{task}}(t)$. Across real datasets, the curves distinguish feature-explained, task-oppositional, long-range, and role-driven regimes that one-hop summaries can obscure. RAP therefore separates two questions often conflated: *what relational organization exists?* and *how does it matter for the task?*

This shifts the unit of analysis from *architecture* to *problem*. Rather than asking only how to build a more expressive graph model, we ask whether an unseen problem can first be diagnosed and then matched to an appropriate computation. RAP provides an operational step toward *problem-aware graph learning*. Our main contributions are:

- We introduce the **Relational Alignment Profile (RAP)**, a compact problem-side representation derived from multiscale alignment between structural, feature, and task geometries.
- We derive **RAP alignment curves**, $\rho_{\text{align}}(t)$ and $\rho_{\text{task}}(t)$, revealing the sign, strength, and scale of relational organization and separating structure–feature alignment from task signal beyond it.
- We formulate **problem-to-computation prediction**: predicting which graph-processing bias is most compatible with an unseen problem before exhaustive target-graph search.
- We demonstrate **cross-generator and frozen synthetic-to-real transfer** to 22 real benchmarks, where the 5D native coordinate α provides the strongest transferable architecture-selection signal and significantly outperforms matched-dimensional generic compressions.

2 Related Work

Characterizing graph-learning problems beyond homophily. Prior work shows edge homophily is a poor indicator of when message passing helps, motivating stronger scalars such as adjusted homophily and label informativeness (Platonov et al., 2023), and finer analyses of heterophily (Luan et al., 2023). These are our strongest conventional baselines, but each remains a single-scale scalar. RAP instead captures *how* structural, feature, and label geometries align across scales. §4.3 shows that similar homophily can produce very different alignment curves, separating feature-explained, oppositional, and role-driven regimes that one-hop summaries collapse.

Architectures for heterogeneous and multiscale relational structure. A large body of work designs GNN variants for specific relational regimes: high-pass and signed-propagation methods for heterophily (Zhu et al., 2020; Bo et al., 2021), generalized PageRank and multihop diffusion (Chien et al., 2021; Klicpera et al., 2019), and role-aware encoders privileging positional similarity over proximity (Ribeiro et al., 2017; Donnat et al., 2018). Each architecture embeds, implicitly, a claim about which relational mechanism it suits; standard practice is discovering which claim holds by training and comparing several. Recent work (Navuluru et al., 2026) organizes this space explicitly, decomposing GNN layers into functional slots separating *where* information moves from *what* moves, descriptive, and silent on which configuration a new problem requires. We supply the complementary problem-side object: the mechanisms $a_1, \ldots, a_M$ evaluated in §4 span exactly the biases this literature has developed one architecture at a time, and $\alpha(\mathcal{D})$ predicts, zero-shot, which one a given problem calls for.

Predicting architecture performance without exhaustive search. A parallel line asks the same question, which computation suits this problem, without training every candidate on the target graph. Graph neural architecture search treats it as a search problem over a design space, via reinforcement learning or evolutionary control (Gao et al., 2021; Zhou et al., 2022); a profile is complementary, since it could initialize or prune such a search rather than replace it. MetaGL (Park et al., 2023) performs evaluation-free selection by meta-learning over a historical corpus of evaluated graph-model pairs, retrieving structurally similar graphs and recommending whatever worked for them, fundamentally a similarity lookup that presumes access to that corpus. NetInfoF (Lee et al., 2024) computes a closed-form, theoretically grounded score answering *whether* a GNN can succeed at all, not *which* one should be used. Outside graphs, AutoML represents datasets by learned meta-feature vectors to transfer model choices across tasks (Jomaa et al., 2021), extracted from tabular statistics with no relational geometry to align against. A relational alignment profile differs from all four: rather than a search procedure, a similarity lookup, a sufficiency score, or a feature-only meta-vector, $\gamma(\mathcal{D})$, its label-free component α especially, is a structured, multiscale diagnosis of *which kind* of relational organization is present and *where*, predicting *which* mechanism a problem needs, zero-shot and without a historical corpus.

Multiscale and statistical dependence measures. The native alignment field (§3.2) is a rank-based departure from independence in the sense of copula theory (Nelsen, 2006), estimated via the standard empirical-copula construction (Deheuvels, 1979; Fermanian et al., 2004), adapted for ties following Genest & Nešlehová (2007). Our propagation distances draw on diffusion geometry (Coifman & Lafon, 2006), made scalable via Chebyshev approximations of graph spectral filters (Hammond et al., 2011) rather than exact eigendecomposition. Our endpoint-disjoint cross-fitting (Appendix B.4) adapts the cross-fitting principle of double/debiased machine learning (Chernozhukov et al., 2018) to dyadic data, where pairs sharing nodes break the usual i.i.d. assumption. To our knowledge, this combination is new: a copula-based, multiscale description of how structural, feature, and task geometries align on the same node population, used not only to characterize a graph-learning problem but to predict which computational bias it requires.

3 METHODOLOGY

Core idea: one node population, multiple geometries. An attributed graph does not define a single notion of similarity. The same nodes can be close because they are adjacent, because information diffuses between them over several steps, because they occupy analogous structural roles, or because their attributes are similar. The learning task introduces another relation: whether two nodes tend to share a label. RAP starts from a simple question: *how do structure, attributes, and task align or conflict, and at what relational scale does that interaction occur?*

This distinction matters computationally. Two graph-learning problems can have similar density, degree distribution, homophily, or feature smoothness and yet require very different processing. Direct neighbors may agree, oppose one another, or be irrelevant; useful dependence may emerge only after several propagation steps; or the relevant relation may be structural role rather than proximity. A local averaging GNN, high-pass filter, multihop diffuser, and role-aware encoder make different assumptions precisely about these cases.

We represent this organization through the *Relational Alignment Profile* (RAP), $\gamma(\mathcal{D}) = [\alpha(\mathcal{D}), \beta(\mathcal{D})] \in \mathbb{R}^9_{\geq 0}$, constructed from richer multiscale *alignment fields* between

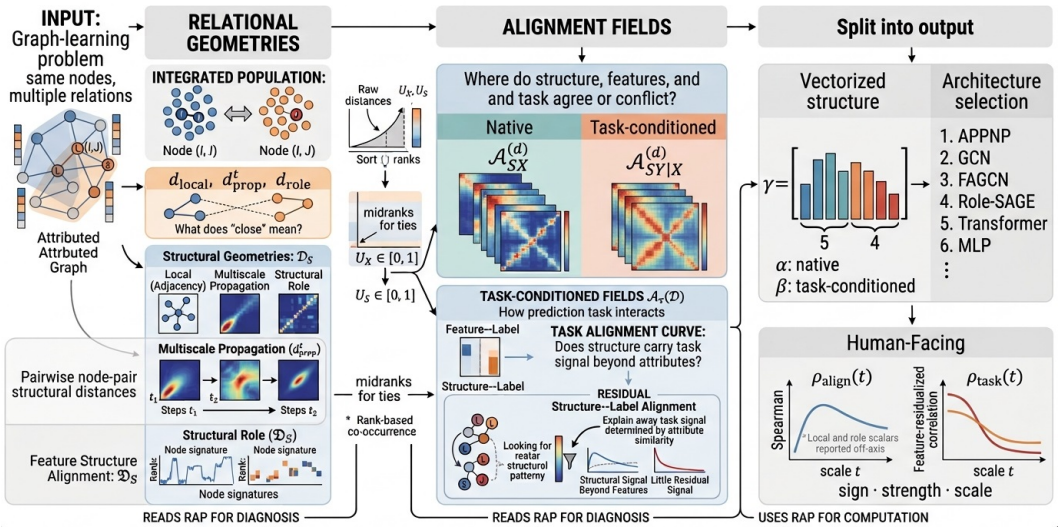


Figure 1: **Relational Alignment Profile (RAP) pipeline.** Structural, feature, and task geometries yield multiscale alignment fields, the compact coordinate $\gamma = [\alpha, \beta]$ for architecture selection, and dictionary-free curves $\rho_{\text{align}}(t)$ and $\rho_{\text{task}}(t)$ for diagnosis.

structural, feature, and task geometries. The native coordinate $\alpha \in \mathbb{R}^5$ describes relational organization already present in (G, X), while the task-conditioned coordinate $\beta \in \mathbb{R}^4$ describes how the prediction task interacts with that organization. From the same underlying alignments we derive two dictionary-free curves, $\rho_{\text{align}}(t)$ and $\rho_{\text{task}}(t)$, which expose their sign, strength, and scale directly. Thus the two views serve complementary purposes:

RAP coordinates for computation	alignment curves for diagnosis.

We develop the construction for node classification. Estimation, cross-fitting, uncertainty, dictionary construction, and scalable approximations are detailed in Appendix B.

3.1 Relational Geometries: What Does "Close" Mean?

Let $\mathcal{D}$ denote a node-classification problem on an attributed graph $G = (V, E)$ with nodes $V = \{1, \ldots, n\}$, attributes $x_i \in \mathbb{R}^p$, and labels y_i observed on $V_L \subseteq V$. Feature proximity is measured by $d_X(i, j) = \|x_i - x_j\|_2$. Graph structure, however, admits several learning-relevant notions of proximity. We use $\mathfrak{D}_S = \{d_{\text{local}}\} \cup \{d^t_{\text{prop}} : t \in \mathcal{T}\} \cup \{d_{\text{role}}\}$. These views deliberately correspond to different relational biases.

Local geometry. d_{local} records direct adjacency and asks whether immediate neighbors exhibit a coherent relational pattern. This is the geometry most closely associated with ordinary message passing and one-hop statistics such as homophily.

Propagation geometry. Adjacency alone can be misleading when the relevant relation is not one-hop. For normalized graph Laplacian L, let $H_t = e^{-tL}$. We compare nodes through their diffusion profiles, $d^t_{\text{prop}}(i, j) = \|H_t(i, \cdot) - H_t(j, \cdot)\|_2$. Small t emphasizes local neighborhoods, whereas increasing t exposes broader relational organization. Sweeping t therefore asks whether a relation is local, delayed, or diffuse rather than fixing a propagation range in advance.

Role geometry. Graph proximity is not the only useful structural relation. Distant nodes may occupy analogous roles, for example, hubs in different communities or corresponding motif positions. We therefore use d_{role}, the distance between multiscale structural-role signatures.

These distances have incomparable units and distributions, so we compare relative rather than absolute proximity. For a sampled node pair (I, J), let $U^{(d)}_S = F_d(d(I, J))$ $U_X = F_X(d_X(I, J))$, where F_d and F_X are empirical marginal CDFs, with midranks for ties. The resulting rank coordinates are invariant to monotone rescalings and place all geometries on a common scale.

3.2 From Geometries to Alignment Fields

A single correlation could summarize whether two geometries tend to agree, but it cannot say *where* that dependence occurs. Two problems can have the same overall correlation even if one aligns only among the closest pairs and another only at intermediate range; positive and negative regions can also cancel. We therefore retain the full departure from independence before compressing it.

Structure–feature alignment. For each structural geometry $d \in \mathfrak{D}_S$, define

$$A_{SX}^{(d)}(r,s) = \Pr\Big(U_S^{(d)} \le r,\, U_X \le s\Big) - \Pr(U_S^{(d)} \le r)\Pr(U_X \le s).$$

If structural and feature proximity are independent, the field vanishes. Positive regions indicate co-occurring closeness; negative regions indicate opposition. The field therefore records not only *whether* structure and attributes are related, but *where in the pair distribution* the relation occurs and *which structural geometry and scale* carries it.

We collect the native fields as $\mathcal{A}_0(\mathcal{D}) = \left\{A_{SX}^{(d)} : d \in \mathfrak{D}_S\right\}$. This object is completely label-free: it describes relational organization already present in (G, X) before specifying how that organization relates to the task.

Why labels require a second layer. Strong structure–feature alignment need not imply added predictive value. If nearby nodes have similar attributes that already determine their labels, the graph may be highly organized yet add little task information.

Let $Z_Y(I,J) = \mathbb{I}\{y_I = y_J\}$, $\bar{Z}_Y = Z_Y - \hat{p}$, where $\hat{p}$ is the baseline same-label probability. We first record raw structure–label and feature–label organization:

$$A_{SY}^{(d)}(r) = \mathbb{E}\Big[\bar{Z}_Y\, \mathbb{I}\{U_S^{(d)} \le r\}\Big], \qquad A_{XY}(s) = \mathbb{E}\big[\bar{Z}_Y\, \mathbb{I}\{U_X \le s\}\big].$$

More importantly, we ask what structural task signal remains after feature similarity has already been accounted for. Because sampled node pairs share endpoints, ordinary pairwise cross-fitting can leak node-level label information. We therefore estimate $\mu(s) \approx \mathbb{E}[\bar{Z}_Y \mid U_X = s]$ using endpoint-disjoint cross-fitting and let $\widetilde{R}(I,J)$ denote the resulting centered residual (Appendix B.4). We then define $A_{SY|X}^{(d)}(r) = \mathbb{E}\Big[\widetilde{R}\,\mathbb{I}\{U_S^{(d)} \le r\}\Big]$. Thus $A_{SY|X}^{(d)}$ asks whether structural proximity organizes label agreement *beyond what feature similarity already predicts*.

We collect the task-conditioned fields as $\mathcal{A}_\tau(\mathcal{D}) = \left[A_{XY}, \left\{A_{SY}^{(d)}, A_{SY|X}^{(d)}\right\}_{d\in\mathfrak{D}_S}\right]$

The native and task-conditioned alignment fields are the rich relational objects from which both RAP and its diagnostic curves are derived.

3.3 RAP Coordinates: A Finite Vocabulary of Relational Mechanisms

The alignment fields are expressive but high-dimensional. To compare many problems and use their relational organization inside a downstream learner, we map them to a common finite vocabulary. Rather than asking an unsupervised compression to discover arbitrary latent directions, we ask: *to what extent does this problem resemble a small set of canonical relational mechanisms?*

Native coordinate α. Let $q_0 = \mathrm{vec}[A_{SX}^{(d)}]_{d\in\mathfrak{D}_S}$ be the concatenated native alignment fields. We construct the dictionary $\Phi = [\phi_{\mathrm{agr}}, \phi_{\mathrm{opp}}, \phi_{\mathrm{multi}}, \phi_{\mathrm{role}}, \phi_{\mathrm{diff}}]$ from controlled problems exhibiting five canonical forms of relational organization: local agreement, local opposition, delayed multihop organization, structural-role correspondence, and diffuse multiscale organization. The native coordinate is $\alpha = \arg\min_{a\ge 0} \|q_0 - \Phi a\|_2^2 \in \mathbb{R}^5_{\ge 0}$. Its entries are therefore nonnegative coordinates with respect to named mechanism templates rather than arbitrary latent axes. Because α depends only on (G, X), it requires no task labels.

Task-conditioned coordinate β. Task organization admits a complementary vocabulary. Let $q_\tau = \mathrm{vec}(\mathcal{A}_\tau)$. We construct $\Psi = [\psi_{\mathrm{feat}}, \psi_{\mathrm{struct}}, \psi_{\mathrm{opp}}, \psi_{\mathrm{role}}]$ around four canonical task-conditioned regimes: *feature-explained*, where structure organizes attributes but adds little task signal beyond them; *structure-explained*, where structure carries additional task signal; *opposed*, where structural proximity is negatively associated with residual label agreement;

and *role-aligned*, where task organization follows structural role rather than ordinary proximity. We define $\beta = \arg\min_{b \geq 0} \|q_\tau - \Psi b\|_2^2 \in \mathbb{R}^4_{\geq 0}$. The resulting Relational Alignment Profile is $\gamma(\mathcal{D}) = [\alpha, \beta] \in \mathbb{R}^9_{\geq 0}$. Conceptually, α describes *what relational organization exists* in (G, X), whereas β describes *how the prediction task interacts with that organization*.

Both dictionaries are constructed exclusively from controlled synthetic meta-training problems. Held-out synthetic generators, mechanism mixtures, and all real datasets are excluded from dictionary construction; their role is to test whether the resulting coordinates recognize relational organization beyond the generators used to define the vocabulary.

3.4 Reading RAP: Alignment Curves

RAP provides a compact coordinate for computation, but compression obscures how a relation changes with scale. For interpretation we therefore return to the underlying rank geometry and derive dictionary-free curves over propagation time.

Structural alignment curve. For propagation scale t, $\rho_{\text{align}}(t) = \text{Corr}\left(U_S^{(t)}, U_X\right)$ is the Spearman correlation between structural and feature proximity. For continuous marginals, $\rho_{\text{align}}(t) = 12 \int_0^1 \int_0^1 A_{SX}^{(t)}(r, s)\, dr\, ds$. Thus the curve is a signed scalar readout of the native alignment field: positive values indicate structure–feature agreement, negative values opposition, and variation with t reveals the scale at which the relation emerges.

In short, $\rho_{\text{align}}(t)$ *asks whether structure organizes the attributes at scale* t.

Task alignment curve. Using the feature-residualized label agreement above, $\rho_{\text{task}}(t) = -\,\text{Corr}\left(U_S^{(t)}, \widetilde{R}\right)$ is signed so that positive values mean structurally close pairs agree in label *more than their attributes alone would predict*. Negative values mean structural proximity is task-oppositional, while values near zero indicate little residual monotone structural signal. Hence, $\rho_{\text{task}}(t)$ *asks whether structure carries task signal beyond the attributes at scale* t.

The two curves can therefore tell qualitatively different stories. Strong $\rho_{\text{align}}(t)$ with negligible $\rho_{\text{task}}(t)$ means that structure organizes the attributes but adds little new task information. Conversely, $\rho_{\text{align}}(t) \approx 0$ with positive $\rho_{\text{task}}(t)$ indicates useful relational task structure invisible to feature smoothness. The canonical task regimes represented by Ψ provide a useful vocabulary for interpreting such patterns, but the curves themselves require no dictionary projection.

For either curve we summarize the dominant relation by $t^\star = \arg\max_{t \in \mathcal{T}} |\rho(t)|$, $\rho^\star = \rho(t^\star)$, where $|\rho^\star|$ measures strength, its sign distinguishes agreement from opposition, and $t^\star$ identifies the characteristic propagation scale. Direct-adjacency and role alignments are reported separately because they represent distinct structural geometries rather than points on the diffusion axis.

The curves are intentionally lossy relative to the underlying fields: they expose sign, strength, and scale, whereas the fields retain the detailed location and shape of the dependence. Appendix D develops a small vocabulary for interpreting these curve shapes.

3.5 From a Problem Profile to Computation

RAP is actionable if relational organization predicts which computational bias will succeed. This yields a direct *problem-to-computation* hypothesis. Let $a_1, \ldots, a_M$ be a fixed architecture bank spanning feature-only, low-pass, high-pass, multihop, role-aware, and local–global processing. We fit a deliberately low-capacity map $f(\gamma(\mathcal{D})) \longrightarrow \left(\widehat{\text{Perf}}(a_1), \ldots, \widehat{\text{Perf}}(a_M)\right)$. The predictor is intentionally simple: the goal is not to solve architecture selection with a powerful meta-model, but to test whether the *problem representation itself* predicts computational compatibility.

Different architectures assume different ways information should move through a graph; RAP describes the relational organization on which those assumptions act. If the profile is meaningful, different problems should induce different architecture rankings. §4 therefore compares γ, its native component α, the alignment curves, raw alignment fields, conventional graph statistics, and matched-dimensional PCA/NMF compressions. We test under increasing distribution shift: first across disjoint synthetic generators and unseen mixtures, then by freezing the synthetic-trained representation and predictor and transferring them unchanged to real benchmarks.

4 EXPERIMENTS

Our experiments address two questions. (i) **Can RAP predict the computation a problem requires?** We evaluate $\alpha(\mathcal{D})$ for architecture selection under synthetic distribution shift and frozen synthetic-to-real transfer, and test whether the task-conditioned extension $\gamma = [\alpha, \beta]$ improves on α. Across the selection experiments and ablations, it does not consistently help (§4.2, Appendix C.5, C.2). (ii) **What do the alignment curves reveal about a problem?** We use $\rho_{\text{align}}(t)$ and $\rho_{\text{task}}(t)$ as dictionary-free diagnostics of whether relational organization is feature-explained, structure-explained, opposed, or role-driven (§4.3). Thus the two parts use RAP for different ends: compact coordinates support automated selection, while the curves explain *why* different computational biases succeed. Code for all experiments can be found here.

4.1 EXPERIMENTAL SETUP

Architecture bank. We use a fixed bank of six architectures spanning distinct processing biases: feature-only MLP, local low-pass (GCN), high-pass (FAGCN), multihop diffusion (APPNP), role-aware SAGE, and a local–global Graph Transformer. Each uses fixed, untuned hyperparameters, with performance averaged over five seeds. Throughout, "best architecture" means the best member of this bank under the shared training protocol; we do not assume the ranking is hyperparameter-invariant. Appendix C.4 tests this sensitivity; full configurations are in Appendix A.

Synthetic problem families. We construct two synthetic families (Appendix A.5): $\mathcal{G}_{\text{train}}$, SBM-based graphs realizing controlled relational mechanisms, used to fit the RAP dictionaries and predictor; and $\mathcal{G}_{\text{test}}$, the same mechanisms realized through disjoint generators and parameter ranges, plus held-out *mixtures* combining two mechanisms in one graph.

Real benchmarks. We evaluate frozen synthetic-trained selectors on 22 real node-classification benchmarks spanning citation, co-purchase/co-authorship, WebKB, heterophilous, social-network, and feature-free airport graphs (Appendix A.4, Table 4), each re-split $60/20/20$ with only training-split labels visible to any label-aware representation.

Representations, baselines, metrics. Every representation passes through the same standardized low-capacity Ridge predictor: CONVENTIONAL graph statistics, the alignment CURVES, the RAW RAP fields, PCA/NMF compressions matched to γ's dimensionality, and RAP α and $\gamma = [\alpha, \beta]$, alongside random/oracle references and, on real data, a **real-fit reference** meta-trained leave-one-dataset-out on the other 21 datasets' true performance (full definitions in Appendix A). We report Spearman rank correlation, selection regret $R(\mathcal{D}) = \max_i p_i(\mathcal{D}) - p_{\hat{\imath}}(\mathcal{D})$, $\hat{\imath} = \arg\max_i \hat{p}_i(\mathcal{D})$, and Top-1/Top-3 hit rate, via a paired regret test.

4.2 PART I: RAP FOR ARCHITECTURE SELECTION

We test whether $\alpha(\mathcal{D})$ predicts relative architecture performance on unseen problems, and whether conditioning on labels ($\gamma = [\alpha, \beta]$) helps. Different relational mechanisms favor different architectures, and no architecture wins uniformly on the real benchmarks either (Sections C.1 and C.3), exactly the heterogeneity a problem-side selector is meant to exploit.

Cross-Generator Architecture Selection. Dictionaries and predictor are trained only on $\mathcal{G}_{\text{train}}$, then frozen and evaluated on the disjoint $\mathcal{G}_{\text{test}}$ and on held-out mechanism mixtures.

On the held-out-generator split (Table 1), α and γ beat CONVEN-

Table 1: **E2: synthetic generalization.** Architecture-selection performance on held-out *synthetic* splits (mean ± SE). *Held-out generator*: a mechanism withheld during fitting. *Held-out mixture*: harder compositional mixtures of mechanisms; RAW/α need a single reference generator and are omitted there.

Predictor	Spearman ρ	Top-1 (%)	Top-3 (%)	Regret (pp)
Held-out generator				
Oracle	1.000±0.000	100.00±0.00	100.00±0.00	0.00±0.00
Random	0.003±0.003	16.91±0.20	50.17±0.27	8.96±0.31
Conven.	−0.177±0.043	12.00±3.78	50.67±5.81	12.34±1.62
Raw	0.509±0.038	21.33±4.76	**66.67**±5.48	2.03±0.31
α	0.534±0.031	**28.00**±5.22	57.33±5.75	**1.14**±0.17
γ	**0.550**±0.033	26.67±5.14	62.67±5.62	1.17±0.16
Held-out mixture				
Oracle	1.000±0.000	100.00±0.00	100.00±0.00	0.00±0.00
Random	0.011±0.003	16.49±0.20	50.17±0.35	10.64±0.59
Conven.	0.245±0.096	6.67±4.63	**70.00**±8.51	11.64±2.38
γ	**0.327**±0.062	**16.67**±6.92	63.33±8.95	**7.18**±1.42

Mean ± SE, SE $= s/\sqrt{n}$ (n=75/30 generator/mixture problems; n=22 real datasets in Table 2), not std. dev.: binary Top-1/Top-3 have raw spread ≈0.45–0.50 regardless of precision. Comparisons use the paired regret test of §4.1.

TIONAL and random on every metric, CONVENTIONAL is in fact worse than random on regret, with γ best on Spearman and α best on top-1 and regret. The mixture split is uniformly harder, and only γ and CONVENTIONAL are defined there (α and RAW require a single reference generator); γ remains clearly ahead, and we treat mixtures as a stress test rather than the setting the dictionary was built for (full results in Appendix C.1).

Frozen Synthetic-to-Real Transfer. The central test freezes exactly this predictor and applies it, unmodified, to all 22 real benchmarks; no real dataset is used to fit Φ, Ψ, f, or architecture hyperparameters. α is the strongest zero-shot predictor by every metric in Table 2, Spearman 0.384, regret 2.22 pp, approaching the in-domain real-fit reference (0.447) without ever seeing real labels, while γ trails on every metric and CONVENTIONAL is worst, with a *negative* mean Spearman. This ordering is unchanged under a five-architecture bank that drops the Transformer, so it is not an artifact of that baseline, and no architecture wins uniformly across the 22 datasets (per-dataset results in Appendix C.4, C.3).

Table 2: **E3: real-transfer.** Aggregate architecture-selection performance across 22 real datasets (mean ± SE). Spearman ρ: rank correlation between predicted and true rankings; Top-1/Top-3: hit rate matching the true best (resp. top-3); Regret: accuracy gap (pp) between predicted- and true-best. Random/Oracle are trivial bounds; Real-fit fits directly on real labels, an in-domain reference rather than zero-shot.

Predictor	Spearman ρ ↑↑	Top-1 (%) ↑↑	Top-3 (%) ↑↑	Regret (pp) ↓↓
Oracle	1.000±0.000	100.00±0.00	100.00±0.00	0.00±0.00
Random	0.003±0.004	16.60±0.37	49.51±0.47	8.61±1.38
Convent.	−0.138±0.084	13.64±7.49	40.91±10.73	6.01±1.72
Raw	0.346±0.085	31.82±10.16	59.09±10.73	5.14±1.95
α	**0.384**±0.083	**36.36**±10.50	**72.73**±9.72	**2.22**±0.75
γ	0.319±0.087	27.27±9.72	59.09±10.73	4.79±1.46
Real-fit	0.447±0.108	40.91±10.73	81.82±8.42	2.32±1.01

What Information in RAP Matters for Selection? Two ablations explain why. Isolating Φ at α's native 5D, α significantly beats matched-dimension PCA/NMF compression of the same raw fields ($p = 0.013$); at γ's combined 9D this advantage disappears, indicating β dilutes Φ's signal rather than Φ itself being redundant with generic compression (E4a). Under a realistic 20-labels/class budget, α and $\alpha + \beta$ both significantly beat a label-aware conventional baseline and a naive selector fit on the uncompressed task field ($p \leq 0.01$ both), so compressing the task field through RAP's dictionaries is clearly preferable to either alternative (E4b; full tables in Appendix C.5). Across every comparison α alone is at least as strong as $\alpha + \beta$, so we center the paper's claims on α rather than γ, a negative result for β as a numeric feature specifically, not for the four task-conditioning categories it is built from, which §4.3 shows are the right vocabulary for reading alignment curves instead.

4.3 PART II: READING GRAPH-LEARNING PROBLEMS WITH ALIGNMENT CURVES

Part I asks whether RAP can help choose a computation. Here we ask a simpler question: *what do the alignment curves reveal about the problem itself?*

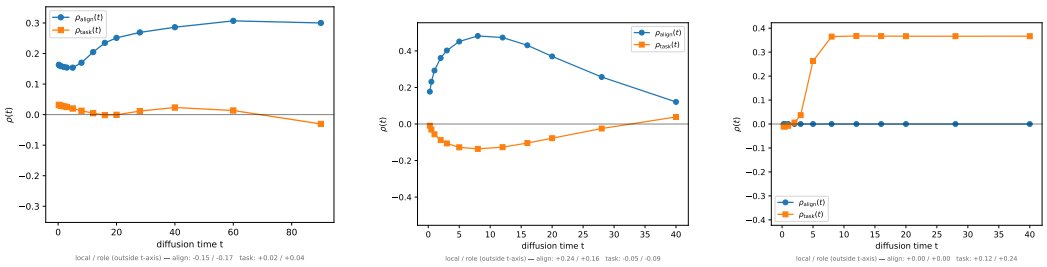
◇ $\rho_{\text{align}}(t)$ asks whether structurally close nodes at scale t have similar features.

◇ $\rho_{\text{task}}(t)$ asks whether structurally close node pairs show excess label agreement beyond what feature similarity explains.

Their sign, magnitude, and shape reveal whether relational organization is local or long-range, aligned or opposed, and feature-explained or genuinely task-relevant. Appendix D gives a compact vocabulary for reading common curve shapes and the two curves jointly (Tables 18–19).

Controlled signatures. We first verify this interpretation where the generating mechanism is known. For the native curve, local agreement appears early, diffuse organization peaks at intermediate scale, and delayed or role-based structure remains weak. For the task curve, delayed and role-driven dependence emerges clearly after conditioning on features. Thus the curves recover not only whether relational signal is present, but also *where in scale* it appears. The full synthetic gallery is in Appendix C.6 (Figure 5).

Three real datasets, three different problems. **CS** shows strong, slowly increasing $\rho_{\text{align}}(t)$ but almost no $\rho_{\text{task}}(t)$. Structure organizes the features, yet structurally close nodes show little excess label agreement beyond what feature similarity explains. Consistently, the feature-only MLP reaches 94.44%, close to the best model at 95.98%. The late-rising curve indicates predominantly long-range structure–feature organization rather than one-hop alignment.



(a) **CS:** strong long-range alignment, little task signal.

(b) **Squirrel-filtered:** aligned features, opposed task.

(c) **Brazil:** no feature alignment, strong structural task signal.

Figure 2: Three qualitatively different graph-learning problems revealed by the alignment curves.

Squirrel-filtered tells a very different story. It has the strongest structure–feature alignment in the suite ($\rho^{\star}_{\text{align}} = 0.482$), despite near-zero homophily, while its task curve is negative over most scales. Thus structure strongly organizes the attributes, but structurally close nodes show *less* label agreement than feature similarity alone would predict. GCN and APPNP, which rely directly on graph propagation, are correspondingly among the weakest architectures. This is a distinction one-hop homophily cannot capture: strong relational organization need not imply task-useful organization.

Brazil shows the complementary regime. It has no meaningful node features, so $\rho_{\text{align}}(t) \equiv 0$. Yet, using the available labels, $\rho_{\text{task}}(t)$ shows strong agreement between structural proximity and label agreement, and the task-role statistic is positive. Thus the observed label organization is structural rather than explained by node features. Consistently, role-aware SAGE wins by a wide margin (60.0% vs. 47.1% for the next-best model).

These examples illustrate what the curves add beyond one-hop summaries: they show *where* relational signal appears, *which direction* it points, and whether it remains task-relevant after accounting for attributes. Across all 22 datasets we observe the same broad regimes—feature-explained, structure-explained, opposed, and role-driven—with the complete dataset-by-dataset analysis in Appendix C.6.

We emphasize that the curves are diagnostics, not deterministic architecture rules. In particular, a late peak $t^{\star}$ indicates long-range relational organization but should not be read literally as a recommended number of GNN layers: our direct depth test finds such a mapping unreliable for stacked GCNs and only weakly informative for decoupled propagation (Appendix D, Table 21). We therefore use curve shape and sign as interpretable evidence about the computational regime, not as a substitute for validation.

5 Conclusion

We introduced the *Relational Alignment Profile* (RAP), a problem-side representation that characterizes how structure, attributes, and task information align across relational scales. RAP supports two complementary uses: compact coordinates for architecture selection and dictionary-free alignment curves for diagnosing whether relational organization is local, multihop, oppositional, feature-explained, or role-driven. Across synthetic generator shift and frozen transfer to 22 real benchmarks, the label-free native coordinate α carries the strongest architecture-selection signal, consistently matching or outperforming the task-conditioned extension $\gamma = [\alpha, \beta]$ and significantly outperforming matched-dimensional generic compressions of the same native fields. The task-conditioned component is instead most informative through the alignment curves, where it reveals forms of task-relevant structure that conventional one-hop summaries can miss.

These results support a problem-first view of graph learning: before deciding how information should be processed, we can first characterize the relational organization of the problem itself. An important next direction is *transfer*. Beyond transferring an architecture selector across problem distributions, RAP could be used to ask whether relational proximity between graph-learning problems predicts which pretrained representations, source datasets, or learned components will transfer successfully. When compatible real sources are unavailable, the same coordinates could also guide the construction of synthetic pretraining problems matched to the relational organization of a target task.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and revising portions of the manuscript, primarily to improve English, clarity, organization, and readability; to generate and edit research code; to implement author-specified procedures for generating the synthetic datasets used in the experiments; and to help create and refine the pipeline figure. The synthetic mechanisms, parameters, and experimental design were specified by the authors, while AI-assisted code was reviewed, tested, and verified before use, and the resulting datasets and experimental outputs were checked for consistency with the intended constructions. We did not use generative AI to independently develop the theoretical framework, formulate or prove mathematical claims, or interpret experimental results. All AI-assisted text, code, data-generation procedures, and visual material were reviewed by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work studies methodological questions in graph learning and does not involve human subjects, private or personally identifiable information, or deployment in high-risk settings. The real-world benchmark datasets used in our experiments are publicly available research datasets, and our synthetic datasets are generated solely for controlled methodological evaluation. We do not release sensitive data or make claims about protected attributes, individuals, or populations. As with any model-selection methodology, RAP could in principle be applied in downstream domains with domain-specific fairness, privacy, or safety concerns; such use would require separate validation and appropriate safeguards. We are not aware of additional ethical concerns specific to the methods or experiments presented here.

REPRODUCIBILITY STATEMENT

We provide complete implementation code and experiment scripts in an anonymous GitHub repository, including RAP construction, synthetic-data generation, architecture training, architecture-selection experiments, ablations, and figure/table generation. The main paper describes the methodology and evaluation protocol, while Appendix A gives the architecture configurations and training procedure, Appendix A.5 specifies the synthetic generators and data construction, and Appendix B provides the RAP estimator, alignment fields, cross-fitting procedure, dictionary construction, uncertainty estimation, and fixed hyperparameters. The real benchmark suite, data splits, evaluation metrics, architecture bank, and robustness checks are also documented in the appendix. All reported experiments use fixed random seeds and specified train/validation/test protocols. Together, the released code and appendix are intended to reproduce the synthetic benchmarks, real-data experiments, ablations, and all reported results. Our implementation is available at
https://anonymous.4open.science/r/RelationalAlignmentProfile-F672

REFERENCES

Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3950–3957, 2021.

Ines Chami, Sami Abu-El-Haija, Bryan Perozzi, Christopher Ré, and Kevin Murphy. Machine learning on graphs: A model and comprehensive taxonomy. *Journal of Machine Learning Research*, 23 (89):1–64, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized PageRank graph neural network. In *International Conference on Learning Representations (ICLR)*, 2021.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

Gabriele Corso, Hannes Stark, Stefanie Jegelka, Tommi Jaakkola, and Regina Barzilay. Graph neural networks. *Nature Reviews Methods Primers*, 4(1):17, 2024.

Paul Deheuvels. La fonction de dépendance empirique et ses propriétés. un test non paramétrique d'indépendance. *Académie Royale de Belgique. Bulletin de la Classe des Sciences*, 65:274–292, 1979.

Claire Donnat, Marinka Zitnik, David Hallac, and Jure Leskovec. Learning structural node embeddings via diffusion wavelets. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1320–1329, 2018.

Jean-David Fermanian, Dragan Radulović, and Marten Wegkamp. Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–860, 2004.

Yang Gao, Hong Yang, Peng Zhang, Chuan Zhou, and Yue Hu. Graph neural architecture search. In *International joint conference on artificial intelligence*. International Joint Conference on Artificial Intelligence, 2021.

Christian Genest and Johanna Nešlehová. A primer on copulas for count data. *ASTIN Bulletin*, 37(2): 475–515, 2007.

David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

Hadi S Jomaa, Lars Schmidt-Thieme, and Josif Grabocka. Dataset2vec: Learning dataset meta-features. *Data Mining and Knowledge Discovery*, 35(3):964–985, 2021.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations (ICLR)*, 2019.

Meng-Chieh Lee, Haiyang Yu, Jian Zhang, Vassilis N. Ioannidis, Xiang Song, Soji Adeshina, Da Zheng, and Christos Faloutsos. NetInfoF framework: Measuring and exploiting network usable information. In *International Conference on Learning Representations (ICLR)*, 2024. Spotlight.

Sitao Luan, Chenqing Hua, Minkai Xu, Qincheng Lu, Jiaqi Zhu, Xiao-Wen Chang, Jie Fu, Jure Leskovec, and Doina Precup. When do graph neural networks help with node classification: Investigating the homophily principle on node distinguishability. *NeurIPS*, 2023.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Sai Karthik Navuluru, Siddhartha Shankar Das, Bo Ni, Hongjie Chen, Yu Wang, Baris Coskunuzer, Nesreen K. Ahmed, Franck Dernoncourt, Mahantesh Halappanavar, Tyler Derr, Ryan A. Rossi, and Lakshman Tamil. Unifying graph neural networks through a common layer equation. *arXiv preprint arXiv:2608.16097*, 2026. URL https://arxiv.org/abs/2608.16097.

Roger B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, 2nd edition, 2006.

Namyong Park, Ryan A. Rossi, Nesreen Ahmed, and Christos Faloutsos. MetaGL: Evaluation-free selection of graph learning models via meta-learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=C1ns08q9jZ.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. In *ICLR*, 2019.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress? *ICLR*, 2023.

Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 385–394, 2017.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *Relational Representation Learning Workshop, NeurIPS 2018*, 2018.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020.

Kaixiong Zhou, Xiao Huang, Qingquan Song, Rui Chen, and Xia Hu. Auto-gnn: Neural architecture search of graph neural networks. *Frontiers in big Data*, 5:1029307, 2022.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *NeurIPS*, 2020.

APPENDIX

A Experimental Setup **14**
A.1 The six architectures . . . 14
A.2 Conventional statistics . . . 14
A.3 Real-fit reference . . . 14
A.4 Real Benchmarks . . . 15
A.5 Synthetic Benchmarks . . . 16

B Relational Alignment Profiles **17**
B.1 Structural Geometries . . . 17
B.2 Empirical Rank Transforms and Alignment Fields . . . 17
B.3 Task-Conditioned Alignment Fields . . . 18
B.4 Endpoint-Disjoint Cross-Fitting . . . 18
B.5 Alignment-Curve Identities . . . 18
B.6 Dictionary Construction . . . 19
B.7 Uncertainty and Stable Peak Summaries . . . 19
B.8 Local RAPs . . . 20
B.9 Scalable Estimation . . . 20
B.10 Hyperparameters and Schedules . . . 20

C Architecture Selection and Alignment-Profile Diagnostics **21**
C.1 Architecture Selection: Synthetic Generalization (E2) . . . 21
C.2 Matched-Summary Counterfactual Diagnosis (E1) . . . 22
C.3 Architecture Selection: Real-Dataset Transfer (E3) . . . 24
C.4 Robustness of Architecture Selection to the Architecture Bank and Hyperparameters 25
C.5 Representation and Label-Budget Ablations (E4) . . . 27
C.6 Interpretation of Alignment Curves and Conventional-Statistic Correlations . . . 29

D Reading Alignment Curves **33**
D.1 Summary statistics . . . 33
D.2 Reading the two curves jointly . . . 34
D.3 From diffusion time to propagation depth . . . 34
D.4 Scope . . . 36

A EXPERIMENTAL SETUP

This appendix gives the full experimental setup referenced from §4.1: the architecture bank and training protocol, the conventional-statistics and real-fit baselines, and the real and synthetic datasets they are evaluated on.

A.1 THE SIX ARCHITECTURES

Table 3 gives the fixed per-architecture configuration used for every synthetic and real problem; hyperparameters are never tuned per dataset. All models are trained with Adam, cross-entropy loss, and early stopping: after every epoch we evaluate the configured metric (accuracy or ROC-AUC, per dataset) on the validation split, keep the best-validation checkpoint, and stop once `patience` consecutive epochs fail to improve it by more than 10^{-8}. Reported performance is always the held-out test-split metric of the restored best-validation checkpoint, averaged over 5 training seeds ($\{0, 1, 2, 3, 4\}$).

Table 3: Fixed per-architecture hyperparameters. `hops`/`alpha` apply only to the multihop (APPNP) architecture; `layers` to the high-pass and Transformer architectures; `heads`/`attn_type` only to the Transformer.

Architecture	Hidden	LR	Weight decay	Epochs	Patience	Dropout	Other
A1 Feature-only MLP	64	0.01	5×10^{-4}	300	50	0.5	—
A2 Local low-pass (GCN)	64	0.01	5×10^{-4}	300	50	0.5	—
A3 High-pass (FAGCN)	64	0.01	5×10^{-4}	300	50	0.5	2 layers
A4 Multihop (APPNP)	64	0.01	5×10^{-4}	300	50	0.5	10 hops, α=0.1
A5 Role-structural (SAGE)	64	0.005	5×10^{-4}	300	50	0.5	—
A6 Graph Transformer (GPS)	64	0.003	1×10^{-4}	300	50	0.3	2 layers, 4 heads, Performer attention

A6 is a GPS-style local–global Graph Transformer using Performer (linear) attention, which avoids the $O(n^2)$ cost of full attention. We inject no extra hand-crafted positional or structural encoding beyond what the other five architectures already receive, so that the input information budget is matched across the bank rather than confounded with encoder-specific features. A5 augments GraphSAGE with the same multiscale structural-role signature $\rho(\cdot)$ used to build d_{role} (Appendix B.1), concatenated to the node features before message passing.

The optional five-architecture bank (`-no-gt` in the reference implementation) drops A6 for a robustness check against the possibility that the Transformer's flexibility, rather than the relational diagnosis itself, drives the real-transfer results of §4.2; results under this variant are reported in Appendix C.4 and leave the ordering unchanged.

A.2 CONVENTIONAL STATISTICS

The CONVENTIONAL baseline of §4.1 is the concatenation of: $\log(1+n)$ and $\log(1+m)$; graph density $2m/(n(n-1))$; degree mean, standard deviation, quartiles, and maximum; mean feature distance between adjacent nodes (subsampled to 5,000 edges when denser), as a feature-smoothness proxy; four spectral summaries of the graph Laplacian; and, whenever labels are available, adjusted homophily (Platonov et al., 2023), label informativeness (Platonov et al., 2023), and label entropy. Label-aware entries use only the labels available to the current split (training-split labels under the low-label budget of E4b, full training labels elsewhere); this is the same convention used for the task-conditioned RAP.

A.3 REAL-FIT REFERENCE

The REAL-FIT REFERENCE of §4.1 meta-trains the standardized Ridge predictor on the other 21 real datasets' (γ, Perf) pairs and evaluates leave-one-dataset-out. It never touches synthetic data, but it does see other real datasets' true architecture performance — information the frozen, synthetic-trained predictor used everywhere else in the paper never has access to. We report it strictly as an in-domain soft upper reference, not as a zero-shot competitor to α, γ, or any baseline.

Table 4: Dataset info. h_{node} (Pei et al., 2019) is the mean, over nodes, of each node's same-class neighbor fraction; h_{edge} is the fraction of edges connecting same-class nodes; $h_{\text{edge,adj}}$ (Platonov et al., 2023) is the class-imbalance-corrected edge homophily. All three are computed on the full (non-split) label set; #Edges is the undirected, self-loop-free edge count.

Dataset	#Nodes	#Edges	#Classes	#Features	h_{node}	h_{edge}	$h_{\text{edge,adj}}$
Cora	2,708	5,278	7	1,433	0.825	0.810	0.771
CiteSeer	3,327	4,552	6	3,703	0.717	0.736	0.671
PubMed	19,717	44,324	3	500	0.792	0.802	0.686
Amazon-Computers	13,752	245,861	10	767	0.802	0.777	0.682
Amazon-Photo	7,650	119,081	8	745	0.849	0.827	0.785
CS	18,333	81,894	15	6,805	0.832	0.808	0.785
BlogCatalog	5,196	171,743	6	8,189	0.391	0.401	0.272
Flickr	7,575	239,738	9	12,047	0.243	0.239	0.137
Texas	183	279	5	1,703	0.057	0.061	−0.294
Cornell	183	277	5	1,703	0.111	0.123	−0.220
Wisconsin	251	450	5	1,703	0.155	0.178	−0.173
Chameleon-Filtered	890	8,854	5	2,325	0.244	0.236	0.030
Squirrel-Filtered	2,223	46,998	5	2,089	0.191	0.207	0.009
Amazon-Ratings	24,492	93,050	5	300	0.376	0.380	0.140
Tolokers	11,758	519,000	2	10	0.634	0.595	0.093
Questions	48,921	153,540	2	301	0.898	0.840	0.021
Minesweeper	10,000	39,402	2	7	0.683	0.683	0.009
Roman-Empire	22,662	32,927	18	300	0.046	0.047	−0.047
Actor	7,600	26,659	5	932	0.220	0.217	0.003
USA	1,190	13,599	4	–	0.373	0.698	0.263
Brazil	131	1,003	4	–	0.248	0.431	−0.033
Europe	399	5,993	4	–	0.220	0.405	−0.074

A.4 Real Benchmarks

We evaluate on 22 node classification benchmarks. Table 4 reports statistics for all 22 datasets, ordered by broad category. The suite spans the full homophily spectrum and is drawn from: citation networks (Cora, CiteSeer, PubMed; McCallum et al., 2000); co-purchase and co-authorship networks (Amazon-Computers, Amazon-Photo, Coauthor-CS; Shchur et al., 2018); WebKB webpage graphs (Cornell, Texas, Wisconsin) and the Actor co-occurrence network (Pei et al., 2019); the heterophilous benchmark suite (Roman-Empire, Amazon-Ratings, Minesweeper, Tolokers, Questions) and its filtered variants (Chameleon-Filtered, Squirrel-Filtered), which correct duplicate-node artifacts in the original Chameleon/Squirrel graphs (Platonov et al., 2023); attributed social networks (BlogCatalog, Flickr); and struc2vec-style airport networks (Brazil, Europe, USA), which are role-recovery benchmarks: PyG ships them with $x = \text{identity}(n)$ (a one-hot node id, not a real attribute vector) and no predefined split. Because these three carry no real node features, the native/feature-alignment channel (α, $\rho_{\text{align}}(t)$) is expected to sit near zero on them by construction — that is the correct reading ("there is nothing here for features to align with"), not a bug, and any recoverable signal has to come from the task-conditioned channel (β, $\rho_{\text{task}}(t)$) instead. `ogbn-arxiv` is loaded by the reference implementation but excluded from this suite for the reason given in Appendix A.

Splits. Every dataset in Table 4 is re-split with a fresh, class-stratified $60\%/20\%/20\%$ train/validation/test partition (`stratified_masks` in the reference implementation), *regardless of what split it ships with*: Planetoid's Cora/CiteSeer/PubMed are not run with the standard 20-labels-per-class semi-supervised "public" split, and the heterophilous suite's 10 predefined $50/25/25$ folds are likewise ignored, so that every dataset in the suite is evaluated under the same split convention. Splits are seeded by a fixed base seed plus a split index, so they are reproducible but resample independently for each split index used. Only training-split labels are ever passed to task-conditioned RAP (β) or to any label-aware baseline statistic; validation and test labels are used exclusively for architecture evaluation and are never seen by any problem representation.

Table 5: **Synthetic benchmark statistics and construction.** Edge counts are mean $\pm$ standard deviation over the generated benchmark instances. Every problem has 300 nodes, 8 node features, and two classes. The same relational mechanism is realized using structurally different generators in $\mathcal{G}_{\text{train}}$ and $\mathcal{G}_{\text{test}}$.

Family	Mechanism	# graphs	# nodes	# edges	# classes	Construction
$\mathcal{G}_{\text{train}}$	Local agreement	30	300	2173 ± 42	2	Two-block assortative SBM. Nodes of the same latent class connect with probability 0.09 and nodes of different classes with probability 0.008. The latent block assignment is the class label, and features contain a moderate class-dependent shift.
$\mathcal{G}_{\text{train}}$	Local opposition	30	300	2268 ± 50	2	Two-block disassortative SBM. Cross-class edges occur with probability 0.095, compared with 0.006 within class, producing strong one-hop opposition. Features remain moderately class informative.
$\mathcal{G}_{\text{train}}$	Delayed multihop	30	300	1742 ± 34	2	Four-block cyclic SBM. Edges occur primarily between consecutive blocks, while labels are determined by block parity. One-hop interactions therefore primarily cross classes, whereas two-hop paths recover same-class relationships.
$\mathcal{G}_{\text{train}}$	Role correspondence	30	300	1030 ± 60	2	Six-block SBM formed from three communities and two structural roles per community. Core-like and peripheral nodes have different connection patterns, and the prediction label is structural role rather than community membership.
$\mathcal{G}_{\text{train}}$	Diffuse organization	30	300	716 ± 32	2	Eight ordered SBM blocks form a weak chain. Features vary continuously along the block ordering, while the binary task separates the two halves of the chain. Relational information is therefore distributed across several scales rather than concentrated locally.
$\mathcal{G}_{\text{test}}$	Local agreement	15	300	1243 ± 10	2	Two-dimensional geometric kNN graph ($k = 7$). Labels are defined by which side of a spatial boundary a node occupies, and features vary smoothly with the same latent coordinate.
$\mathcal{G}_{\text{test}}$	Local opposition	15	300	1291 ± 16	2	Nodes have random two-dimensional positions and binary labels. Each node is connected to nearby *opposite-label* nodes, creating local opposition through a geometric construction rather than an SBM.
$\mathcal{G}_{\text{test}}$	Delayed multihop	15	300	997 ± 10	2	Nodes lie on a ring with alternating labels. Each node connects to nearby opposite-label nodes, so one-hop connectivity is oppositional while two-hop connectivity returns to the same class.
$\mathcal{G}_{\text{test}}$	Role correspondence	15	300	299 ± 0	2	Repeated four-node path motifs. Path endpoints form one class and interior nodes the other, with sparse bridges connecting motifs. The task is therefore determined by structural role rather than geometric proximity.
$\mathcal{G}_{\text{test}}$	Diffuse organization	15	300	1238 ± 10	2	Two-dimensional geometric kNN graph ($k = 7$). Labels depend on one spatial coordinate while features vary according to a smoother combination of both coordinates, producing broad multiscale rather than sharply local alignment.

A.5 Synthetic Benchmarks

Our synthetic benchmark is designed to control the *relational mechanism* of a graph-learning problem while allowing the underlying graph generator to change. We construct two disjoint families: $\mathcal{G}_{\text{train}}$, used for meta-training, and $\mathcal{G}_{\text{test}}$, used only for evaluation. The two families instantiate the same five relational regimes but use different graph constructions, making $\mathcal{G}_{\text{train}} \to \mathcal{G}_{\text{test}}$ a genuine generator-shift experiment rather than interpolation within a single synthetic model.

All main synthetic problems contain 300 nodes, 8-dimensional node features, and two balanced or approximately balanced classes. We generate 30 independent $\mathcal{G}_{\text{train}}$ problems per mechanism (150 total) and 15 independent $\mathcal{G}_{\text{test}}$ problems per mechanism (75 total). Edge counts vary with the relational mechanism and generator; Table 5 reports the realized mean and standard deviation over the exact benchmark corpus.

Node features. Across all mechanisms, features consist of isotropic Gaussian noise plus a low-dimensional mechanism-dependent signal,

$$x_i = \epsilon_i + s_i v, \qquad \epsilon_i \sim \mathcal{N}(0, I_8),$$

where v is a randomly sampled direction and s_i depends on the latent class, role, or geometric coordinate of the corresponding construction. Thus the synthetic tasks are not purely structural: features contain moderate predictive information, allowing RAP to distinguish whether graph structure agrees with, opposes, or contributes information beyond the attributes.

Train/validation/test splits. Each generated problem is split independently using class-stratified $60\%/20\%/20\%$ train/validation/test partitions — the same split convention used for the real benchmarks (Appendix A.4). Task-conditioned RAP quantities and all label-aware graph statistics use training labels only; validation and test labels are used exclusively for architecture evaluation.

Held-out mixtures. We additionally generate 30 mixture problems with 300 nodes each. Two distinct mechanisms are sampled and independently generated using $\mathcal{G}_{\text{test}}$, each on approximately half of the nodes. The two regions are combined and connected by three random inter-region edges. The resulting graphs contain heterogeneous relational organization that is never presented as a single mechanism during meta-training. Their realized edge count is approximately 978 ± 234, with two classes and the same 8-dimensional feature space.

B Relational Alignment Profiles

This appendix gives the structural geometries, estimation details, endpoint-disjoint cross-fitting, alignment-curve identities, dictionary construction, uncertainty procedures, and scalable approximations for the RAP methodology of §3, followed by the hyperparameters and schedules used throughout.

B.1 Structural Geometries

Local geometry. The local structural relation distinguishes adjacent from non-adjacent node pairs. Because this geometry contains many exact ties, all rank-based quantities use empirical midranks.

Propagation geometry. For normalized Laplacian L, the heat kernel is

$$H_t = e^{-tL}.$$

For each $t \in \mathcal{T}$,

$$d^t_{\text{prop}}(i,j) = \|H_t(i,\cdot) - H_t(j,\cdot)\|_2.$$

The schedule $\mathcal{T}$ differs between the two uses of RAP in this paper; exact values are given in §B.10.

Role geometry. For each node i, let $\rho(i)$ be a multiscale vector of structural statistics, including degree and clustering summaries over fixed hop neighborhoods. We use

$$d_{\text{role}}(i,j) = \|\rho(i) - \rho(j)\|_2.$$

This geometry captures structural equivalence rather than graph proximity.

B.2 Empirical Rank Transforms and Alignment Fields

For m sampled node pairs (I_k, J_k), we compute empirical midranks $\widehat{U}_S^{(d)}$ and $\widehat{U}_X$ from the geometries above. The native field is evaluated on a fixed $K \times K$ grid:

$$\widehat{A}^{(d)}_{SX}(r_a, s_b) = \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}\{\widehat{U}_S^{(d)}(I_k,J_k) \le r_a,\ \widehat{U}_X(I_k,J_k) \le s_b\} - \widehat{P}(\widehat{U}_S^{(d)} \le r_a)\widehat{P}(\widehat{U}_X \le s_b). \tag{1}$$

We subtract the empirical marginal product rather than $r_a s_b$. This matters for discrete structural geometries — especially d_{local} and d_{role} — where ties imply that empirical rank marginals are not exactly uniform. In the continuous tie-free case the estimator reduces to the standard empirical-copula departure from independence (Deheuvels, 1979; Fermanian et al., 2004; Genest & Nešlehová, 2007).

By the Fréchet–Hoeffding bounds,

$$-\frac{1}{4} \le A^{(d)}_{SX}(r,s) \le \frac{1}{4},$$

with extremal magnitude possible only at $r = s = \frac{1}{2}$.

B.3 Task-Conditioned Alignment Fields

Let

$$Z_Y(I,J) = \mathbb{1}\{y_I = y_J\}, \qquad \bar{Z}_Y = Z_Y - \hat{p},$$

where $\hat{p}$ is the empirical same-label probability. In addition to the residual field used in the main text, we retain

$$A_{SY}^{(d)}(r) = \mathbb{E}[\bar{Z}_Y \mathbb{1}\{U_S^{(d)} \le r\}], \tag{2}$$

$$A_{XY}(s) = \mathbb{E}[\bar{Z}_Y \mathbb{1}\{U_X \le s\}]. \tag{3}$$

The full task-conditioned field collection is therefore

$$\mathcal{A}_\tau = \left[A_{XY}, \{A_{SY}^{(d)}, A_{SY|X}^{(d)}\}_{d\in\mathfrak{D}_S}\right],$$

with $A_{SY|X}^{(d)}$ — the residual field beyond feature similarity — constructed via the endpoint-disjoint cross-fitting described next.

B.4 Endpoint-Disjoint Cross-Fitting

Naively splitting sampled pairs is insufficient because two pairs may share a labeled endpoint. A regression fit on one pair can therefore indirectly contain information about an endpoint appearing in a nominally held-out pair.

We partition the labeled nodes into κ folds $V_L^{(1)}, \dots, V_L^{(\kappa)}$. For every unordered fold pair (v, w), including $v = w$, we fit

$$\hat{\mu}_{-(v,w)}(s) \approx \mathbb{E}[\bar{Z}_Y \mid U_X = s]$$

using only pairs whose two endpoints both lie outside $V_L^{(v)} \cup V_L^{(w)}$.

For an evaluation pair (I, J) with endpoint folds (v, w),

$$R(I,J) = \bar{Z}_Y(I,J) - \hat{\mu}_{-(v,w)}(U_X(I,J)).$$

We center the resulting evaluation residuals,

$$\widetilde{R}(I,J) = R(I,J) - \frac{1}{|\mathcal{P}_{\text{eval}}|} \sum_{(I',J')\in\mathcal{P}_{\text{eval}}} R(I',J'),$$

and estimate

$$\widehat{A}_{SY|X}^{(d)}(r) = \frac{1}{|\mathcal{P}_{\text{eval}}|} \sum_{(I,J)\in\mathcal{P}_{\text{eval}}} \widetilde{R}(I,J)\mathbb{1}\{U_S^{(d)}(I,J) \le r\}.$$

These centered residuals feed both the field above and, via the identity of §B.5 below, the task alignment curve $\rho_{\text{task}}(t)$.

The procedure requires $O(\kappa^2)$ low-capacity regressions; we use $\kappa = 3$ folds throughout (§B.10).

B.5 Alignment-Curve Identities

For continuous marginals, $U_S^{(t)}, U_X \sim \text{Unif}[0,1]$, and therefore

$$\text{Corr}(U_S^{(t)}, U_X) = 12\,\text{Cov}(U_S^{(t)}, U_X).$$

Using

$$A_{SX}^{(t)}(r,s) = \Pr(U_S^{(t)} \le r, U_X \le s) - rs,$$

the standard copula identity gives

$$\rho_{\text{align}}(t) = 12 \int_0^1 \int_0^1 A_{SX}^{(t)}(r,s)\, dr\, ds.$$

With ties we compute the correlation directly from empirical midranks rather than using this continuous-marginal normalization.

For the task field, because $\widetilde{R}$ is centered,

$$\int_0^1 A^{(t)}_{SY|X}(r)\,dr = \mathbb{E}\left[\widetilde{R}\int_0^1 \mathbb{1}\{U_S^{(t)} \le r\}\,dr\right] \tag{4}$$

$$= \mathbb{E}[\widetilde{R}(1-U_S^{(t)})] \tag{5}$$

$$= -\operatorname{Cov}(U_S^{(t)}, \widetilde{R}). \tag{6}$$

Hence $\rho_{\text{task}}(t)$ is the normalized signed integral of the residual task field.

B.6 Dictionary Construction

Native dictionary. Let

$$q_0 = \operatorname{vec}\left[\widehat{A}^{(d)}_{SX}\right]_{d\in\mathfrak{D}_S}$$

denote the concatenated native RAP, estimated as in §B.2. We fit a joint dictionary

$$\Phi = [\phi_{\text{agr}}, \phi_{\text{opp}}, \phi_{\text{multi}}, \phi_{\text{role}}, \phi_{\text{diff}}]$$

from controlled synthetic training problems representing five canonical relational regimes. The columns are normalized before NNLS fitting. We then compute

$$\alpha = \arg\min_{a\ge 0} \|q_0 - \Phi a\|_2^2 \in \mathbb{R}^5.$$

The joint construction allows each dictionary atom to encode how its corresponding mechanism appears across the entire structural-geometry bank, while producing a fixed five-dimensional native representation that requires no labels at all.

Task dictionary. We collect the task-conditioned alignment fields into

$$q_\tau = \left[A_{XY}, \{A^{(d)}_{SY}, A^{(d)}_{SY|X}\}_{d\in\mathfrak{D}_S}\right].$$

From synthetic training problems representing feature-explained, structure-explained, oppositional, and role-aligned regimes, we construct

$$\Psi = [\psi_{\text{feat}}, \psi_{\text{struct}}, \psi_{\text{opp}}, \psi_{\text{role}}]$$

and define

$$\beta = \arg\min_{b\ge 0} \|q_\tau - \Psi b\|_2^2 \in \mathbb{R}^4.$$

Together with the native coordinate α, this gives the full RAP coordinate

$$\gamma = [\alpha, \beta] \in \mathbb{R}^9.$$

We evaluate both γ and its native component α; empirically, α provides the stronger architecture-selection representation (§4.2, Appendix C.5). The reconstruction residuals $\|q_0 - \Phi\alpha\|_2$ and $\|q_\tau - \Psi\beta\|_2$ indicate how well the canonical mechanism dictionaries represent a given problem.

Guarding against template circularity. Both dictionaries are constructed exclusively from controlled problems in $\mathcal{G}_{\text{train}}$. We therefore do not treat recovery on that generator family as evidence of generalization. All transfer evaluations use either the structurally distinct $\mathcal{G}_{\text{test}}$ or real datasets, none of which enters dictionary construction. We further compare the RAP coordinates against the uncompressed alignment fields and matched-dimensional PCA/NMF compressions in Appendix C.5.

B.7 Uncertainty and Stable Peak Summaries

Pairs sampled from the same graph share endpoints and are not independent. We therefore quantify uncertainty by resampling rather than by i.i.d. pair-level standard errors.

Native uncertainty. We repeat pair sampling B times and recompute the native RAP, $\rho_{\text{align}}(t)$, and α. The empirical distribution gives pointwise bands for the alignment curve and coordinate uncertainty.

Task uncertainty. To reflect uncertainty in the available labeled population, we resample labeled nodes rather than individual pairs. For every resample we reconstruct the induced pairs, endpoint folds, residual fields, task curve, and β.

Peak stability. Because

$$t^\star = \arg\max_t |\rho(t)|$$

is discontinuous, nearby scales with similar magnitudes can produce unstable peak locations. We therefore record the resampling distribution of $t^\star$. A characteristic scale is reported only when a single scale is selected sufficiently consistently; otherwise only the full curve is reported.

Coordinate reliability. A large NNLS reconstruction residual signals that the current mechanism dictionary does not adequately explain the observed RAP. A large resampling spread indicates that a coordinate is weakly determined by the available graph pairs or labels. We report these quantities as reliability diagnostics rather than forcing an interpretation in such cases.

B.8 Local RAPs

Global RAPs can hide spatially heterogeneous mechanisms. For node i and candidate neighborhood N_i, we construct a local alignment field from pairs

$$\{(i, j) : j \in N_i\}$$

while retaining the same graph-global rank transformations used by the full RAP. This yields local coordinates α_i. Dispersion of $\{\alpha_i\}$ provides a diagnostic for spatial mixtures in which different regions of the graph exhibit different relational regimes.

B.9 Scalable Estimation

Pair subsampling. RAP does not require all $O(n^2)$ node pairs. We use $m \ll n^2$ sampled pairs, stratified across structural-distance ranges. The same pair set is reused across propagation scales so that changes in the alignment curves cannot be attributed to different sampled populations.

Approximate propagation geometry. Rather than constructing every full heat-kernel row, we select $L \ll n$ reference nodes $\ell_1, \ldots, \ell_L$ and use

$$z_t(i) = (H_t(i, \ell_1), \ldots, H_t(i, \ell_L)), \qquad \widetilde{d}^t_{\text{prop}}(i, j) = \|z_t(i) - z_t(j)\|_2.$$

The reference embeddings are computed with truncated Chebyshev polynomial approximations to e^{-tL} (Hammond et al., 2011). The full RAP estimation algorithm can be found in Algorithm 1.

B.10 Hyperparameters and Schedules

Table 6 collects the RAP estimator's fixed hyperparameters, used identically for every synthetic and real problem in this paper.

Table 6: RAP estimator hyperparameters.

Quantity	Value
Native field grid resolution K	6
Propagation scales for α/γ, $\mathcal{T}$	$\{1, 3\}$
Structural-pair budget m	3,000
Local-geometry pair budget	2,500
Labeled-pair budget (task fields)	5,000
Cross-fitting folds κ	3
Chebyshev landmark count L	20
Native dictionary templates per mechanism	20
Task dictionary templates per mechanism	12
Template graph size	220 nodes

Algorithm 1 Relational Alignment Profile Estimation

Require: Graph G, attributes X, optional labels y on V_L, structural geometries $\mathfrak{D}_S$, dictionaries Φ, Ψ, pair budget m

Sample node pairs and compute structural and feature rank coordinates
for $d \in \mathfrak{D}_S$ **do**
    Estimate native field $\widehat{A}^{(d)}_{SX}$
end for
Assemble native RAP $\mathcal{R}_0$
Concatenate native fields into q_0
Solve NNLS for $\alpha \in \mathbb{R}^5$
if $V_L \neq \emptyset$ **then**
    Construct endpoint-disjoint folds and cross-fitted residuals $\widetilde{R}$
    for $d \in \mathfrak{D}_S$ **do**
        Estimate $\widehat{A}^{(d)}_{SY}$ and $\widehat{A}^{(d)}_{SY|X}$
    end for
    Estimate $\widehat{A}_{XY}$
    Assemble task RAP $\mathcal{R}_\tau$
    Concatenate task fields into q_τ
    Solve NNLS for $\beta \in \mathbb{R}^4$
    Set $\gamma = [\alpha, \beta] \in \mathbb{R}^9$
end if
for $t \in \mathcal{T}$ **do**
    Compute $\widehat{\rho}_{\text{align}}(t) = \text{Corr}(\widehat{U}^{(t)}_S, \widehat{U}_X)$
    if $V_L \neq \emptyset$ **then**
        Compute $\widehat{\rho}_{\text{task}}(t) = -\text{Corr}(\widehat{U}^{(t)}_S, \widetilde{R})$
    end if
end for
Estimate uncertainty by pair/node resampling
return raw RAP $\mathcal{R}$, coordinates α (and γ if labeled), alignment curves and uncertainty diagnostics

The machine-facing coordinate α (and its extension γ) uses a deliberately small structural-geometry bank (d_{local}, d^1_{prop}, d^3_{prop}, d_{role}) so that the resulting representation stays low-dimensional and cheap to compute at meta-learning scale. The human-facing alignment curves of §3.4 and the diagnostic reported in Appendix C.6 instead sweep a much denser diffusion-scale grid, $t \in \{0.25, 0.5, 1, 2, 3, 5, 8, 12, 16, 20, 28, 40\}$ (extended to 60/90 for CS, Computers, PubMed, and Photo's ρ_{align} and for USA's ρ_{task}, none of which had peaked at $t = 40$; Appendix C.6), so that the shape of the curve — not just two representative points on it — can be read off directly. Both instantiations use the same estimator (§B.2); only the queried schedule $\mathcal{T}$ differs.

C Architecture Selection and Alignment-Profile Diagnostics

This appendix reports full architecture-selection results on both synthetic (E2) and real (E3) data, the matched-summary counterfactual diagnosis (E1), the representation and label-budget ablations (E4), and a detailed reading of the relational alignment profile (RAP) diagnostic on all 22 real benchmarks used in E3, including its correlation structure against conventional graph statistics.

C.1 Architecture Selection: Synthetic Generalization (E2)

E2 tests whether an architecture-selection predictor fit on one region of the synthetic generator distribution generalizes to graphs it has not seen, *before* the predictor is ever exposed to real data. Two evaluation splits are used. The *held-out generator* split withholds one synthetic mechanism entirely during fitting and evaluates on graphs drawn from it; the *held-out mixture* split evaluates on synthetic graphs that are compositions of multiple mechanisms, a harder compositional-generalization setting with no single ground-truth generator to reference. Table 1 reports aggregate rank correlation (Spearman ρ between predicted and true architecture rankings), hit rate at which the predictor's top

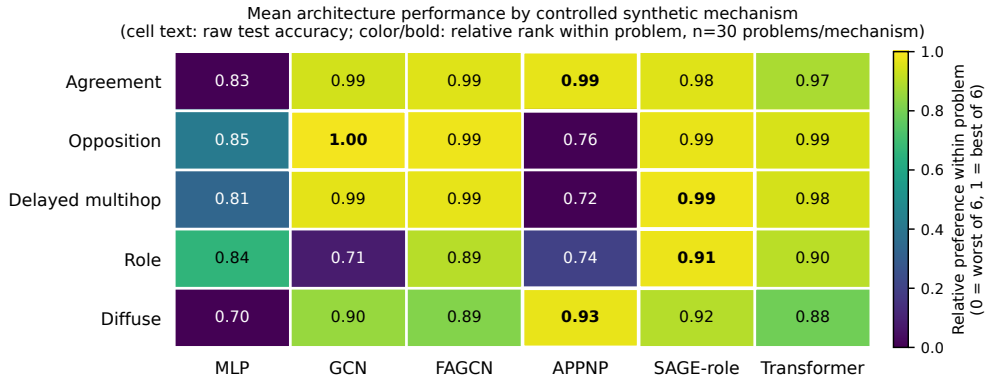


Figure 3: **Different relational problems favor different computations.** Mean architecture performance on the $\mathcal{G}_{\text{train}}$ corpus (30 problems per mechanism); cell text is raw test accuracy, color and boldface mark each problem's relative rank among the six architectures (normalized within problem, emphasizing relative preference over absolute task difficulty). The feature-only MLP is the uniform loser, confirming structure carries real signal in every regime; beyond that, no single architecture dominates. GCN and FAGCN are both strong under agreement and opposition; APPNP (multihop diffusion) is surprisingly the weakest non-MLP architecture under opposition and delayed multihop specifically, consistent with a fixed low-pass diffusion kernel over-smoothing exactly where structure and label disagree; role-aware SAGE is the most consistent performer, winning outright under delayed multihop and role correspondence. This heterogeneity is the sanity check referenced in §4.2: if one architecture dominated the whole problem collection, a problem-side selector would be unnecessary.

pick (resp. top-3 picks) matches the true best architecture, and regret (the accuracy gap, in percentage points, between the predicted-best and true-best architecture), for each candidate predictor. The full table (Table 1) already appears in the main text (§4.2); we do not repeat it here and instead discuss it in full.

Two things stand out. First, no single predictor dominates every metric on the held-out-generator split the way α does later on real data (§C.3): γ has the best rank correlation (0.550), α has the best top-1 hit rate and lowest regret (28.0%, 1.14 pp), and RAW has the best top-3 hit rate (66.7%). All three of RAW, α, and γ comfortably beat both the random baseline and the conventional-statistics baseline on every metric, and the conventional baseline is in fact substantially *worse* than random on regret (12.34 pp mean regret vs. 8.96 pp for random) while nominally having a higher top-3 rate — its top-1 selections are simply bad. Second, the mixture split is uniformly harder: every predictor's top-1 rate drops and regret roughly doubles relative to the generator split, and only CONVENTIONAL and γ are even defined for this split (RAW and α require a single reference generator and are not applicable to mixtures). γ remains the more reliable of the two here, winning Spearman, top-1, and regret; CONVENTIONAL only edges it out on top-3, again driven by a comparatively weak top-1 rate (6.7%, below even its own top-3-inflated appearance and below the random baseline's 16.5%). The overall picture from E2 is that curve-derived predictors (α, γ, and to a lesser extent RAW) transfer within the synthetic distribution substantially better than a fixed-vocabulary conventional graph-statistics baseline, and that this gap persists — if anything widens in relative terms — under the harder compositional (mixture) split.

C.2 MATCHED-SUMMARY COUNTERFACTUAL DIAGNOSIS (E1)

E1 asks a narrower and more targeted question than E2 or E3: holding a single attributed graph (G, X) fixed and changing only the task mechanism that generates the labels — one-hop agreement, opposition, or delayed two-hop dependence — can a problem representation tell these three learning problems apart, and does the representation's predicted-best architecture track the true best architecture as the mechanism changes? Because (G, X) is identical across the three conditions of a triple, the native coordinate α (which depends only on the structure–feature relationship) is identical to machine precision across all three rows of a triple by construction; only the four-dimensional

Table 7: **E1: mechanism identification and retrieval.** Grouped cross-validation (groups = triple) over 30 matched-summary counterfactual triples (90 rows total), 3-way classification, chance = 33.3%.

Representation	Mechanism-ID Acc. (%)	NN-Retrieval Acc. (%)
Conventional (adj. hom., LI)	**41.1**	**44.4**
Raw fields	27.8	27.8
RAP (γ)	26.7	41.1
Chance	33.3	—

Table 8: **E1: counterfactual example.** First of 30 triples; (G, X) fixed across all three rows, so α is identical to machine precision in every row ($\alpha = [0.494, 0.000, 0.000, 0.395, 0.230]$) and is omitted from the table. β shown is its dominant (first, "feature-explained labels") component; the other three components are at or near zero in every row (one-hop agreement's secondary components are 0.012 and 0.016; all other off-diagonal components are exactly 0). Regret is the accuracy gap between the predicted-best and true-best architecture for that row.

Mechanism	Adj. hom.	Label inf.	β (dom.)	Predicted best	Actual best	Regret (pp)
One-hop agreement	0.050	0.0018	0.218	Role-SAGE	Local low-pass (GCN)	5.8
Opposition	−0.027	0.0005	0.000	Role-SAGE	Graph Transformer	1.3
Delayed two-hop	0.012	0.0001	0.078	Role-SAGE	Role-SAGE	**0.0**

task coordinate β, the raw task-conditioned field, or the two label-aware conventional statistics can possibly distinguish them. We generate 30 such triples on graphs of $n = 240$ nodes — an order of magnitude smaller than the real benchmarks in §C.3 — using the frozen dictionaries and the frozen γ predictor from E2.

Table 7 reports two metrics computed on the pooled 90 (triple, mechanism) rows with grouped cross-validation (groups = triple, so no triple's rows appear in both train and evaluation folds): mechanism-ID accuracy, a 3-way classification of which of the three mechanisms generated a given row's labels, and NN-retrieval accuracy, the accuracy of a nearest-neighbor lookup that excludes all rows from the query's own triple. Chance on the 3-way task is 33.3%.

This result cuts directly against a reading in which the RAP representation is what reveals relational structure that conventional summaries conflate. Here, the two-dimensional conventional vector (adjusted homophily, label informativeness) is the best discriminator of the three mechanisms by a wide margin on both metrics (mechanism-ID 41.1%, retrieval 44.4%), while γ is the *worst* of the three representations at mechanism-ID (26.7%, barely above chance and below even the raw, unprojected fields at 27.8%), and only closes part of the gap to conventional on retrieval (41.1% vs. 44.4%). The ordering is worth taking at face value rather than explaining away: because α is constant within a triple, γ's entire discriminative capacity for this task must come from its four-dimensional β block, and β appears to be a lossier summary of the task-conditioned field than the raw field itself, let alone than the two conventional label-aware scalars — despite the fact that the three generating mechanisms here are exactly the canonical templates the task dictionary Ψ was built from. This suggests the projection step that produces β discards information that is specifically useful for identifying *which* mechanism produced a task, even where (§C.3, §C.5) the same α machinery is comparatively strong at predicting *architecture performance*. Mechanism identification and performance prediction are evidently not the same task, and this diagnostic should not be read as evidence against the performance-prediction results elsewhere in the paper — but it is a genuine negative result for β's specific compression, and one reason (among the others in Appendix C.5) that we center the paper's claims on α rather than γ.

Table 8 shows the fully-detailed example from the first triple. It illustrates the same tension concretely. The frozen γ predictor selects ROLE_STRUCTURAL as the predicted-best architecture for all three mechanisms, despite β changing substantially between them (its dominant component is 0.218 under one-hop agreement, 0 under opposition, and 0.078 under delayed two-hop) — i.e. the predictor's induced architecture ranking is essentially insensitive to the β variation it is given. The true best architecture does change across the three rows (local low-pass, graph Transformer, and role-structural,

Table 9: Per-dataset true architecture performance (accuracy %, mean ± std over seeds). Best architecture per dataset in **bold**. Architectures: feature-only MLP, low-pass GCN, high-pass FAGCN, multi-hop APPNP, role-structural SAGE, and a graph Transformer.

Dataset	MLP	GCN	FAGCN	APPNP	SAGE-role	Transformer
Cora	75.04±1.18	89.89±0.52	89.89±0.32	**90.26**±0.38	89.89±0.63	85.37±1.48
CiteSeer	72.02±0.69	75.39±0.49	**75.66**±0.31	75.60±0.30	75.18±0.34	74.31±0.52
PubMed	87.63±0.15	88.07±0.15	88.15±0.15	87.91±0.12	**89.51**±0.08	88.28±0.23
Amazon-Computers	83.88±1.51	90.05±0.23	84.52±6.26	89.73±0.41	**91.17**±0.44	90.46±0.52
Amazon-Photos	91.37±0.27	92.79±0.32	92.47±1.36	93.29±0.31	94.90±0.21	**94.96**±0.32
CS	94.44±0.09	94.03±0.12	94.70±0.10	94.61±0.15	95.13±0.07	**95.98**±0.18
BlogCatalog	93.71±0.23	78.00±0.54	93.36±1.36	95.01±0.51	**96.48**±0.26	94.80±0.55
Flickr	91.95±0.28	66.33±0.42	90.14±0.80	92.62±0.31	**92.76**±0.15	90.28±0.30
Texas	**81.14**±2.91	60.57±2.80	70.86±2.14	58.29±1.40	74.86±2.14	61.71±3.88
Cornell	**79.46**±3.24	49.19±3.15	54.05±2.96	54.05±1.71	72.43±3.15	72.97±1.71
Wisconsin	**84.80**±0.98	50.00±1.26	60.80±7.33	51.60±1.50	81.60±2.94	74.80±6.27
Chameleon-Filtered	36.98±3.54	41.56±1.95	45.59±2.33	39.55±1.47	44.92±1.30	**47.04**±3.03
Squirrel-Filtered	33.96±1.62	32.52±1.83	36.40±1.82	32.75±0.37	37.07±0.76	**38.06**±1.95
Amazon-Ratings	40.63±1.09	43.04±0.15	41.06±0.19	42.60±0.27	47.42±0.30	**50.92**±0.90
Tolokers	74.31±0.05	76.25±0.08	78.01±0.44	71.20±0.14	84.12±0.13	**85.64**±0.12
Questions	71.58±0.56	60.26±0.19	50.59±1.96	45.85±2.86	**78.78**±0.21	77.29±0.59
Minesweeper	49.85±0.96	70.57±0.08	84.12±0.76	66.13±0.01	90.38±0.13	**90.89**±0.08
Roman-Empire	65.82±0.08	48.62±0.19	56.51±0.96	52.53±0.27	79.29±0.19	**82.35**±0.37
Actor	36.59±0.34	30.00±0.58	**37.29**±0.78	33.84±0.44	35.18±0.93	33.63±1.02
USA	25.25±0.33	59.50±3.41	64.50±1.38	51.00±1.50	**67.83**±1.07	42.92±5.37
Brazil	27.86±2.67	39.29±3.91	47.14±2.67	32.86±4.16	**60.00**±3.50	46.43±7.49
Europe	26.42±1.67	49.63±1.98	48.89±2.54	29.63±1.91	**58.77**±1.85	42.96±4.51

respectively), so the predicted-best matches the true-best only once out of three (delayed two-hop). The practical consequence is softer than the raw match rate suggests, because the six architectures' true accuracies are tightly clustered in this small-n regime (0.45–0.58): the regret of following the constant ROLE_STRUCTURAL prediction is 5.8 pp on one-hop agreement, 1.3 pp on opposition, and 0.0 pp on delayed two-hop, for a mean regret of 2.4 pp across the triple — not catastrophic, but not the clean "prediction tracks the changing mechanism" illustration a headline figure would want either. We present it here as the honest, representative (`base_id==0`, not hand-picked) draw from the 30 generated triples, rather than searching for a more flattering one; $n = 240$ is also an order of magnitude smaller than any real dataset in §C.3 and may simply be too small for six architectures to differentiate cleanly.

C.3 Architecture Selection: Real-Dataset Transfer (E3)

E3 is the central transfer test: predictors fit entirely on synthetic data (never touching real labels, with the single exception of the REAL-FIT REFERENCE ablation described in Appendix A) are applied zero-shot to select among six candidate architectures — a feature-only MLP, a low-pass GCN, a high-pass FAGCN, a multi-hop APPNP, a role-structural SAGE variant, and a graph Transformer — on all 22 real datasets. Table 9 reports the true (fully-trained, architecture-by-architecture) accuracy on every dataset, establishing the ground truth that the predictors are being scored against. Table 2 then reports each predictor's aggregate selection quality across all 22 datasets.

Table 2, already shown in the main text (§4.2), reports each predictor's aggregate selection quality across all 22 datasets.

No architecture is uniformly best (Table 9): the role-structural SAGE variant wins outright on 8 of 22 datasets (PubMed, Amazon-Computers, BlogCatalog, Flickr, Questions, USA, Brazil, Europe), the graph Transformer wins another 8 (Photo, CS, Chameleon-Filtered, Squirrel-Filtered, Amazon-Ratings, Tolokers, Minesweeper, Roman-Empire — concentrated among the Platonov heterophilous benchmarks together with the two co-purchase/co-authorship graphs whose alignment curves peak latest, CS and Photo), the feature-only MLP wins the three WebKB datasets by a wide margin, FAGCN wins CiteSeer and Actor, and APPNP wins Cora. The WebKB result is worth calling out

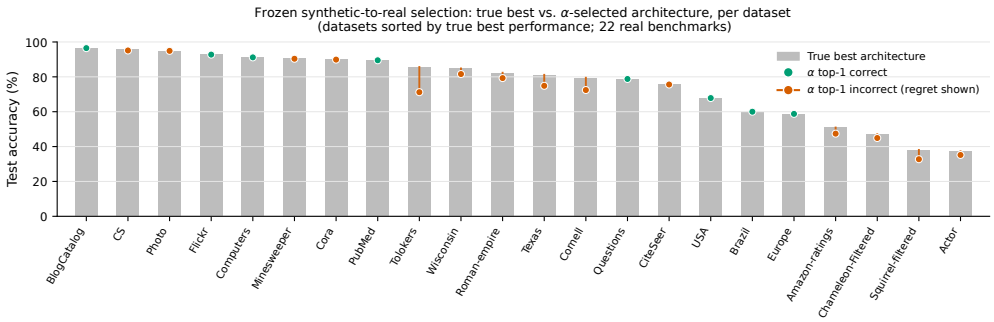


Figure 4: **Frozen synthetic-to-real selection at the dataset level.** Gray bars show the true best architecture's test accuracy on each of the 22 real datasets (sorted by this value); the marker shows the test accuracy of the architecture α actually selects, with a green dot marking an exact top-1 match and an orange dot (connected by a stem to the true best) marking a miss, whose length is the regret. The predictor sees only α of each real problem, never a real architecture-performance observation. Most misses are small; the three largest (Tolokers, Cornell, Texas) are visible as the longest stems.

specifically: on Texas, the MLP's 81.14% beats the next-best architecture (SAGE-role, 74.86%) by more than 6 points, and the same pattern holds on Cornell and Wisconsin — consistent with these three datasets' strongly negative adjusted homophily (Table 4, -0.29 to -0.17), where graph structure is not merely uninformative but actively anti-correlated with label agreement, so the correct architectural choice is to disregard structure entirely. This heterogeneity is precisely what motivates treating architecture selection as a per-dataset prediction problem rather than defaulting to a single architecture.

Against this heterogeneous ground truth, α is the strongest zero-shot predictor by every metric in Table 2: Spearman 0.384, top-1 36.4%, top-3 72.7%, and regret 2.22 pp. CONVENTIONAL performs worst among the real predictors, with negative mean Spearman correlation (-0.138). Notably, α approaches the REAL-FIT REFERENCE in rank correlation (0.384 vs. 0.447), even though the latter is fit in-domain using architecture performance from the other real datasets (Appendix A); we therefore report it only as an in-domain reference, not a zero-shot competitor. The task-conditioned extension does not improve selection here: $\gamma = [\alpha, \beta]$ trails α on every metric (Spearman 0.319 vs. 0.384; top-1 27.3% vs. 36.4%; regret 4.79 vs. 2.22 pp). Thus the strongest real-transfer result is carried by the label-free structure–feature coordinate α.

C.4 ROBUSTNESS OF ARCHITECTURE SELECTION TO THE ARCHITECTURE BANK AND HYPERPARAMETERS

E1–E4 and §C.3 use one fixed, six-architecture bank with one fixed, untuned hyperparameter set per architecture (Table 3). Two checks address whether the architecture-*selection* results above are an artifact of that specific setup rather than of the relational diagnosis itself.

Five architectures instead of six. The six-architecture bank includes a GPS-style Graph Transformer (A6) whose flexibility could in principle absorb outlier tasks and make α's selection quality look better than it would for a bank of more homogeneous, purely message-passing architectures. We re-ran E2 and E3 end-to-end on the five-architecture bank that drops A6 (`-no-gt`), with a freshly-fit Φ/Ψ and predictor (a bank change alters what "the best architecture" means per problem, so the frozen E2 state cannot be reused as-is).

Dropping the Transformer does not weaken the result — if anything it strengthens it. On E3, α's regret *improves* from 2.22 pp (six architectures, Table 2) to 1.68 pp (five architectures), and its top-1 rate rises from 36.4% to 59.1%; γ improves too (4.79→3.41 pp) but remains clearly behind α, exactly reproducing the ordering in §4.2. The same holds on E2: both α and γ's held-out-generator regret improve slightly (1.14→0.96 pp and 1.17→0.97 pp) with the ordering unchanged. One honest caveat: a smaller bank mechanically lowers the maximum possible regret, so the absolute point drops should not be over-read — but the *relative* ordering the reviewer's concern is actually about (conventional

Table 10: E2 on the **five**-architecture bank (MLP, GCN, FAGCN, APPNP, SAGE-role; no Graph Transformer), directly comparable to Table 1's six-architecture numbers.

Predictor	Spearman ρ	Top-1 (%)	Top-3 (%)	Regret (pp)
Held-out generator				
Random	0.001±0.022	20.12±1.66	59.68±2.03	10.27±3.02
Conventional	0.078±0.467	6.67±24.94	60.00±49.0	9.67±12.73
Raw	0.582±0.324	30.67±46.11	85.33±35.38	1.85±3.17
α	0.566±0.296	**34.67**±47.59	85.33±35.38	**0.96**±1.61
γ	**0.627**±0.282	37.33±48.37	**86.67**±34.0	0.97±1.63
Oracle	1.000±0.000	100.00±0.00	100.00±0.00	0.00±0.00
Held-out mixture				
Random	0.001±0.021	20.01±1.93	59.47±2.45	11.83±3.16
Conventional	0.409±0.397	**33.33**±47.14	83.33±37.27	7.66±9.45
γ	**0.437**±0.338	30.00±45.83	**93.33**±24.94	**6.06**±7.16
Oracle	1.000±0.000	100.00±0.00	100.00±0.00	0.00±0.00

Table 11: E3 on the **five**-architecture bank, all 22 real datasets, directly comparable to Table 2's six-architecture numbers.

Predictor	Spearman ρ	Top-1 (%)	Top-3 (%)	Regret (pp)
Random	−0.002±0.022	20.16±1.80	59.41±1.75	8.81±7.05
Conventional	−0.158±0.436	18.18±38.57	45.45±49.79	5.47±7.37
Raw	0.351±0.405	45.45±49.79	**90.91**±28.75	8.04±12.17
α	0.397±0.454	**59.09**±49.17	77.27±41.91	**1.68**±3.22
γ	**0.399**±0.419	50.00±50.00	81.82±38.57	3.41±5.99
Oracle	1.000±0.000	100.00±0.00	100.00±0.00	0.00±0.00
Real-fit reference	0.469±0.471	72.73±44.54	90.91±28.75	1.66±4.60

$\ll$ raw/α/γ, and $\alpha \lesssim \gamma$) is what is being tested here, and it is fully preserved. The one exception is RAW, whose E3 regret gets *worse* without the Transformer (5.14→8.04 pp); RAW is never the paper's recommended representation, and with only 22 real datasets a single high-dimensional, uncompressed baseline moving against the trend is not surprising.

Architecture ranking under a local hyperparameter grid. A separate concern is that the single fixed hyperparameter set (Table 3) disadvantages some architectures — FAGCN and the Graph Transformer are the most plausible candidates — relative to others, so that the six-way ranking α/γ are being scored against is itself an artifact of that choice rather than a stable ground truth. We grid-searched learning rate $\in \{0.003, 0.01, 0.03\}$ and hidden width $\in \{32, 64, 128\}$ (9 configurations) for FAGCN and the Transformer only, substituted each dataset's best-found configuration for the fixed default, and recomputed the full six-way ranking, comparing it against the fixed-hyperparameter ranking via Spearman correlation and whether the argmax (predicted-best architecture) changes.

Table 12: **Architecture ranking under a local hyperparameter grid, all five real datasets tested.** ρ is the Spearman rank correlation between the two six-way rankings (fixed default vs. best-found FAGCN/Transformer configuration); "argmax" is the single best-predicted architecture under each setting.

Dataset	ρ(fixed, tuned)	argmax (fixed → tuned)	Changed?
Cora	0.90	multihop → high_pass	yes
CiteSeer	0.83	high_pass → high_pass	no
PubMed	0.14	role_structural → graph_transformer	yes
Roman-empire	1.00	graph_transformer → graph_transformer	no
Tolokers	1.00	graph_transformer → graph_transformer	no

Across all five datasets, the ranking is stable on three (60%) and flips on two (40%), and the two flips are not the same failure mode as each other, nor as the three stable cases. Roman-empire and Tolokers

Table 13: **E4a: representation ablation.** Raw alignment fields vs. PCA and sign-preserving NMF compressed to γ's own dimensionality (9), vs. γ itself, on the E3 real corpus (all 22 datasets). p is the paired regret test against γ as reference (OURS); the leading two PCA components explain 79.7% of the raw fields' variance ($53.3\% + 26.4\%$).

Representation	Dim	Spearman ρ	Top-1 (%)	Top-3 (%)	Regret (pp)	p vs. γ
Raw	198	0.346	31.8	59.1	5.14	0.50
PCA	9	**0.369**	**36.4**	**72.7**	**3.10**	0.80
NMF	9	0.342	27.3	68.2	3.92	0.66
RAP (γ)	9	0.319	27.3	59.1	4.79	—

($\rho = 1.00$ both) are stable because the Graph Transformer wins by a comfortable margin under the fixed default (82.95% vs. 80.26% for the next-best role-aware SAGE on Roman-empire; 83.92% vs. 81.54% on Tolokers), and tuning only widens that gap (to 85.28% and 84.39% respectively) rather than closing it. CiteSeer ($\rho = 0.83$) is stable at the top despite reshuffling below it: the Transformer overtakes role-aware SAGE for second place under tuning (74.25% vs. 74.19%, a 0.06-point swap), but neither comes close to high_pass, which wins both ways at 74.58%. The two flips look different from each other. Cora's is a coin-flip among near-ties: multihop (89.96%), high_pass, and role-structural (89.81% each) sit within 0.15 points of one another under the fixed default, and tuning raises high_pass by only 0.37 points (to 90.19%) — enough to overtake multihop by a 0.22-point margin, not evidence of a real capability difference. PubMed's is a genuine, non-marginal change: the Graph Transformer is the *worst* of the six architectures under the fixed default (86.38%) but gains 2.71 points under tuning (to 89.09%), overtaking role-aware SAGE (88.92%) for the top spot — the fixed hyperparameters specifically handicapped the Transformer on this dataset, and tuning corrects that. The honest conclusion is not that RAP's predictions are wrong, but that "the best architecture under one fixed hyperparameter set" — the ground truth every experiment in this paper is scored against — is itself hyperparameter-dependent on 2 of the 5 real datasets checked here, most clearly PubMed. This is a limitation of the shared evaluation protocol (best addressed by per-dataset architecture tuning in future work), not a failure specific to α or γ: both are evaluated against the same six-way ground truth on every dataset alike, and the ranking is reliable exactly where the best architecture holds a comfortable margin over the rest.

C.5 Representation and Label-Budget Ablations (E4)

E4 asks two further questions about *why* the RAP representation performs as it does on real transfer, using the same frozen E2 dictionaries and predictors and the same real corpus built for E3: (E4a) does the mechanism-aware compression in γ do anything that a generic low-rank compression of the same raw fields could not, at matched dimensionality; and (E4b) under a realistic low-label-supervision budget, does the label-free α (or $\alpha + \beta$) beat conventional label-aware statistics and a naive direct use of the raw task field.

E4a: representation ablation. Table 13 compares the raw alignment fields (198 dimensions), a PCA compression of the raw fields to 9 dimensions (matching γ's dimensionality), a sign-preserving NMF compression to the same 9 dimensions, and γ itself, all evaluated with the same standardized Ridge meta-predictor on the E3 real corpus. The result does not support a framing in which γ's value comes from being a *mechanism-aware* compression rather than merely a low-dimensional one: generic PCA at matched dimensionality beats γ on every metric (Spearman 0.369 vs. 0.319; top-1 36.4% vs. 27.3%; top-3 72.7% vs. 59.1%; regret 3.10 pp vs. 4.79 pp), and sign-preserving NMF beats it on three of the four metrics (Spearman 0.342, top-3 68.2%, regret 3.92 pp), tying on top-1. Even the uncompressed raw fields edge out γ on Spearman rank correlation (0.346 vs. 0.319). The paired regret test against γ as reference gives $p = 0.50$ (raw), $p = 0.80$ (PCA), and $p = 0.66$ (NMF): none of these differences reach significance in either direction at $n = 22$, so the numerical gap in PCA and NMF's favor should not be over-read as a proven advantage for generic compression — but it equally means there is no evidence in this experiment that γ's mechanism dictionary is doing anything a data-driven low-rank approximation could not. The two leading PCA components already explain 79.7% of the raw fields' variance ($53.3\% + 26.4\%$), consistent with the real corpus at this sample size having only two or three directions of real variation for any compression method, mechanism-aware

Table 14: **E4b: label-budget ablation.** 20 labels/class, sampled only from each dataset's official training pool; all label-aware representations recomputed under this same budget. p is the paired regret test against $\alpha + \beta$ as reference (OURS).

Representation	Spearman ρ	Top-1 (%)	Top-3 (%)	Regret (pp)	p vs. $\alpha + \beta$
α only	**0.419**	**36.4**	**72.7**	**2.22**	0.841
Conventional (label-aware)	0.172	13.6	77.3	7.28	0.010
Direct task-field selector	−0.022	9.1	31.8	11.49	0.0005
$\alpha + \beta$	0.385	36.4	63.6	2.38	—

Table 15: **E4c: 5D native-field ablation.** α (5D NNLS on the native field A_{SX} alone) vs. 5D PCA and 5D sign-preserving NMF fit strictly on the same native field — matched at α's own dimensionality, unlike Table 13's 9D match to γ. p is the paired regret test against α as reference.

Representation	Dim	Spearman ρ	Top-1 (%)	Top-3 (%)	Regret (pp)	p vs. α
Raw native (A_{SX})	144	0.427	40.9	81.8	3.82	0.19
PCA (native)	5	0.320	31.8	77.3	4.81	**0.013**
NMF (native)	5	0.325	31.8	72.7	4.81	**0.013**
α (5D NNLS)	5	**0.384**	**36.4**	72.7	**2.22**	—

or not, to find. This is a genuinely negative result for a "mechanism-aware compression distinguishes meaningful structure from generic dimensionality reduction" reading of γ, and we report it as tested rather than as a demonstrated advantage.

E4b: label-budget ablation. Table 14 restricts every label-aware representation to the same 20-labels-per-class budget, sampled only from each dataset's official training pool, and compares α alone (label-free, included as a reference point that is unaffected by the label budget), the two conventional label-aware statistics (adjusted homophily and label informativeness, computed on the label-limited set), a "direct" baseline that fits the same low-capacity Ridge predictor straight on the uncompressed, label-limited raw task field instead of any compressed coordinate, and $\alpha + \beta$ with β recomputed from the same limited labels. Here the result is unambiguously favorable to the RAP representation: both α alone and $\alpha + \beta$ decisively and significantly beat the conventional label-aware baseline (regret 2.22/2.38 pp vs. 7.28 pp; paired $p = 0.0104$ against $\alpha + \beta$) and the direct task-field selector (regret 11.49 pp; paired $p = 0.00048$), which is in fact worse than random selection on rank correlation ($\rho = -0.022$). Under a realistic low-label budget, compressing the task field through RAP's dictionaries — rather than feeding a high-capacity representation to a necessarily low-capacity predictor, or falling back to generic label-aware statistics — is clearly the right choice.

That said, the same pattern seen in E1 and in the full real-transfer result of §C.3 reappears here: α alone is at least as strong as $\alpha + \beta$ on every metric (Spearman 0.419 vs. 0.385; top-3 72.7% vs. 63.6%; regret 2.22 vs. 2.38 pp), and the paired regret test between them is nowhere near significant ($p = 0.841$). Adding the task-conditioned β component does not measurably help architecture selection under a limited label budget either.

E4c: 5D native-field ablation. E4a's null result is measured at γ's combined 9D, mixing the native (Φ, 5D) and task (Ψ, 4D) dictionaries together; it does not by itself say whether Φ specifically is doing useful work. Table 15 re-runs the same comparison at α's own native dimensionality (5D), isolating Φ from any interaction with Ψ: here α significantly beats both 5D PCA and 5D NMF fit on the identical raw native field ($p = 0.013$ both), roughly halving regret (2.22 vs. 4.81 pp) and improving Spearman correlation (0.384 vs. 0.320/0.325), and is statistically indistinguishable from the fully uncompressed 144D field ($p = 0.19$). This is the result Weakness 2's own proposed remedy asks for: at matched dimensionality, the mechanism-aware native dictionary Φ is empirically validated, not merely tied with generic compression. Reconciling this with E4a: the 9D null result is attributable specifically to β/Ψ (the weaker, task-conditioned half of γ) diluting Φ's signal once the two are concatenated, not to Φ itself being redundant with unsupervised compression — consistent with, and a direct empirical confirmation of, the §3.3 framing of β/γ as an ablation rather than a co-pillar.

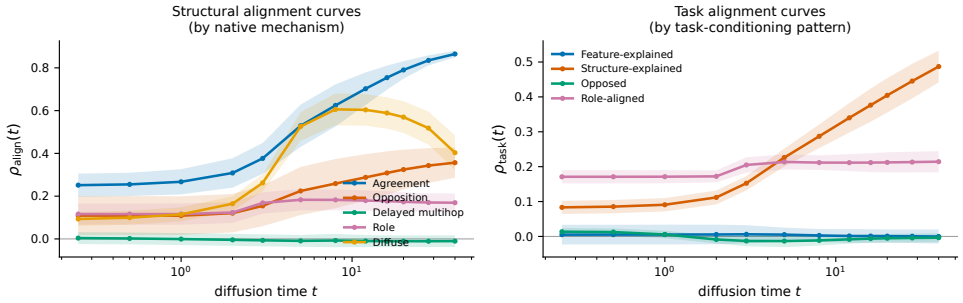


Figure 5: **Canonical signatures of relational mechanisms.** Mean ± one std. alignment curves across 8 resampled seeds per mechanism on held-out synthetic problems. *Left:* $\rho_{\text{align}}(t)$ for the five native (structure–feature) generator mechanisms. Agreement rises monotonically to the strongest long-range alignment in the gallery; opposition rises far more slowly to a much lower plateau (weaker, not sign-reversed, structure–feature correspondence); diffuse organization is the one non-monotonic signature, peaking at intermediate scale ($t^\star \approx 8$) before decaying; delayed multihop and role correspondence stay closest to zero, since both are constructed as label-side rather than feature-side dependencies and are largely invisible to a structure–feature curve. *Right:* $\rho_{\text{task}}(t)$ for the four task-conditioning patterns in the β dictionary. Structure-explained labels are exactly the delayed, long-range task signal absent from the native curves, rising to the largest task peak in the gallery only after $t \gtrsim 5$; role-aligned labels carry a smaller but already-saturated signal from short scales; feature-explained and opposed patterns stay near zero throughout, as intended by construction. The two panels are deliberately not row-matched: α's five native templates and β's four task templates are separate dictionaries (§3.3), so the mechanisms diagnostic to one need not be the same mechanisms diagnostic to the other.

Reading E1, E3, and E4 together. Across five independent tests — the full real-transfer comparison (§C.3), the 9D representation ablation (E4a), the matched-5D native-field ablation (E4c), the label-budget ablation (E4b), and (in a different sense) the mechanism-identification result of E1 — a consistent picture emerges: the label-free native coordinate α carries the useful signal for architecture selection and its native dictionary Φ is specifically validated against generic compression at matched (5D) dimensionality (E4c), while the task-conditioned β component never measurably improves on α alone and sometimes makes both selection and mechanism identification worse, diluting Φ's otherwise-demonstrable advantage when the two are concatenated into γ at 9D (E4a). β is nonetheless clearly better than not compressing the label-limited raw task field at all (E4b) or than relying on conventional graph statistics (§C.3) — it is redundant with α, not useless in isolation. None of this undermines the paper's central claim that a label-free, dictionary-based structure–feature coordinate predicts which architecture a real graph-learning problem needs — α's performance in §C.3 is the strongest result in the paper, and its native dictionary is the specific piece of RAP's construction E4c validates — and it is why the main text (§3, §4) centers its claims on α, presenting β and $\gamma = [\alpha, \beta]$ as a tested ablation rather than the headline representation.

C.6 Interpretation of Alignment Curves and Conventional-Statistic Correlations

Controlled synthetic signatures. §4.3 verifies that the curves behave as intended under controlled mechanisms; Figure 5 shows the full per-mechanism gallery underlying that discussion.

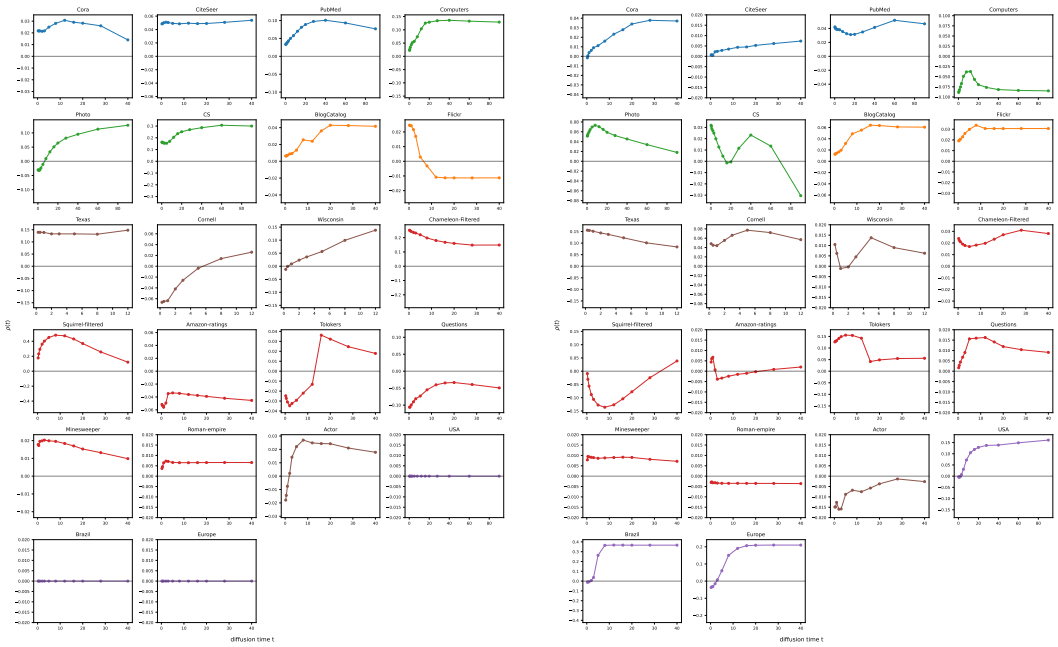
Real-dataset curves. The relational alignment profile (RAP) diagnostic computes two curves per real dataset over a diffusion-scale grid $t \in \{0.25, 0.5, 1, 2, 3, 5, 8, 12, 16, 20, 28, 40\}$, extended to $60/90$ for CS, Computers, and PubMed's ρ_{align}, for Photo's ρ_{align}, and for USA's ρ_{task} (Appendix B.10): each of these five curves was still rising at the original grid's edge $t = 40$, and the extension locates a true peak at $t = 60$ for CS and PubMed, confirms Computers' peak sits at $t = 40$ itself (the curve decays cleanly at $60/90$), and finds Photo's ρ_{align} and USA's ρ_{task} still climbing even at $t = 90$, not yet at their asymptote. $\rho_{\text{align}}(t)$ is a structure–feature correspondence statistic comparing feature similarity at diffusion scale t between connected vs. randomly-paired nodes, and

Table 16: Alignment-profile curve summary statistics on all 22 real datasets. $t^*_{\text{align}}/\rho^*_{\text{align}}$ are the diffusion scale and value at which $\rho_{\text{align}}(t)$ peaks in magnitude (structure–feature correspondence); $t^*_{\text{task}}/\rho^*_{\text{task}}$ are the same for $\rho_{\text{task}}(t)$ (structure–label correspondence beyond what features already predict). local and role are the two fixed diagnostic scalars (balanced edge-vs-non-edge comparison, and structural-role correspondence) reported alongside each curve.

	$\rho_{\text{align}}(t)$				$\rho_{\text{task}}(t)$			
Dataset	t^*	ρ^*	local	role	t^*	ρ^*	local	role
Cora	12.00	0.0308	0.2460	−0.0592	28.00	0.0378	0.0135	0.0229
CiteSeer	40.00	0.0538	0.3717	0.0107	40.00	0.0074	0.0320	−0.0033
PubMed	40.00	0.1007	0.2882	−0.0335	60.00	0.0517	0.0092	0.0335
Amazon-Computers	40.00	0.1359	−0.0547	−0.0571	0.25	−0.0887	0.0678	0.0902
Amazon-Photos	90.00	0.1271	−0.0004	−0.0864	8.00	0.0733	0.0782	0.1473
CS	60.00	0.3069	−0.1454	−0.1676	0.25	0.0319	0.0192	0.0415
BlogCatalog	20.00	0.0429	−0.0478	−0.0204	16.00	0.0640	0.0351	0.0474
Flickr	0.25	0.0246	−0.0187	−0.0202	8.00	0.0338	0.0049	0.0349
Texas	12.00	0.1482	−0.0492	0.0327	0.25	0.1558	−0.0693	0.0647
Cornell	0.25	−0.0667	−0.0267	−0.0106	5.00	0.0772	−0.0639	0.0772
Wisconsin	12.00	0.1382	−0.0669	−0.0887	5.00	0.0137	−0.0193	0.0849
Chameleon-Filtered	0.25	0.2524	−0.2306	−0.1121	28.00	0.0310	0.0014	0.0086
Squirrel-Filtered	8.00	0.4820	0.2412	0.1586	8.00	−0.1359	−0.0468	−0.0854
Amazon-ratings	1.00	−0.0560	0.2250	0.0212	1.00	0.0067	−0.0129	−0.0056
Tolokers	16.00	0.0362	0.2628	0.1796	5.00	0.1556	−0.0170	0.0762
Questions	0.25	−0.1063	0.2245	0.0558	12.00	0.0163	−0.0000	0.0014
Minesweeper	3.00	0.0203	0.0454	0.0004	0.50	0.0095	−0.0084	0.0002
Roman-empire	2.00	0.0072	0.0289	0.0205	40.00	−0.0036	−0.0000	0.0289
Actor	8.00	0.0270	−0.0018	−0.0354	2.00	−0.0159	−0.0135	−0.0133
USA	0.25	0.0000	0.0000	0.0000	90.00	0.1609	0.1108	0.1283
Brazil	0.25	0.0000	0.0000	0.0000	12.00	0.3676	0.1158	0.2380
Europe	0.25	0.0000	0.0000	0.0000	28.00	0.2101	0.1153	0.1062

$\rho_{\text{task}}(t)$ is the analogous statistic for structure–label correspondence *after* conditioning out what node features alone already predict about the label. Each curve is summarized by its peak location t^* and peak value ρ^*, together with two fixed-scale diagnostic scalars: local (a balanced 1-hop edge-vs-non-edge comparison) and role (a structural-role correspondence scalar). Table 16 reports all eight values for every one of the 22 real datasets. Figure 6 shows the full curve grid; Figure 2 reproduces the three regimes highlighted in §4.3. The rest of this subsection reads every one of the 22 datasets against the same four-pattern vocabulary—feature-explained, structure-explained, opposed, role-aligned—that β's dictionary Ψ was built to compress numerically (§4.2) but that survives much better as a description of curve shape than as a nine-dimensional feature.

Three structural regimes. The citation/co-purchase/co-authorship graphs split into two families rather than behaving uniformly. CiteSeer and Cora peak early and locally: CiteSeer's alignment is essentially flat across the whole grid with by far the largest local scalar in this family (+0.372), meaning almost all of its structure–feature signal already lives at the immediate neighbor; Cora peaks at $t^* = 12$ with a substantial local scalar (+0.246) and genuinely decays afterward. CS, Computers, and Photo behave oppositely: their curves rise monotonically to very late peaks ($t^* = 60$, 40, and 90 respectively) with local scalars that are at or below zero (CS: -0.145, Computers: -0.055, Photo: -0.0004) — direct neighbors are *not* more feature-similar than random pairs in these three graphs, and the entire structure–feature correspondence is a long-range, multi-hop phenomenon invisible to 1–2-hop aggregation. PubMed sits between the two families, rising smoothly like the second group while retaining a sizeable local component (+0.288). The heterophilous benchmarks behave as designed, with $|\rho_{\text{align}}|$ small at every scale for Roman-empire, Minesweeper, and Questions; Amazon-ratings is the interesting exception, showing genuine local opposition (feature-*dis*similar neighbors, $\text{local} = +0.225$ read against its negative peak $\rho^* = -0.056$). Airports (Brazil/Europe/USA) reads as an exact zero on ρ_{align} at every diffusion scale — these graphs carry no real node features, so this is a clean negative control rather than a null result to worry about.



(a) $\rho_{\text{align}}(t)$, all 22 datasets.

(b) $\rho_{\text{task}}(t)$, all 22 datasets.

Figure 6: Alignment-profile curves across the full diffusion-scale grid for every real dataset. Structure–feature alignment (left) is strongest and broadest among the citation/co-purchase/co-authorship graphs; structure–label alignment beyond features (right) is dominated by the airport networks, which carry no real node features.

Where structure carries information beyond features. Airports carries by far the strongest ρ_{task} signal in the entire suite, and because ρ_{align} is exactly zero there, this signal is unconfounded by any feature-mediated pathway. Brazil peaks at $\rho^* = +0.368$ ($t^* = 12$), Europe at $+0.210$ ($t^* = 28$), and USA — the largest and sparsest of the three — more slowly at $+0.161$ ($t^* = 90$); all three have large positive role scalars, the textbook signature of a structural-role mechanism (airport traffic volume is a function of hub-ness, a global multi-hop property) rather than a local one. Among the citation/co-purchase graphs, Computers is the standout: its task curve is negative almost everywhere (worst at $t \approx 0.25$, $\rho = -0.089$), meaning structure is mildly *anti*-informative once features are conditioned out, a genuinely different story from "no signal."

The remaining datasets extend, and in one case sharpen, this picture. Actor is close to a null result on both curves (align $\rho^* = 0.027$, task $\rho^* = -0.016$, all four scalars indistinguishable from zero at conventional significance), placing it alongside Roman-empire and Minesweeper as one of the hardest benchmarks by this diagnostic. BlogCatalog shows a moderate, late-peaking alignment curve ($t^* = 20$, consistent with the "slow multi-hop" family above) combined with the strongest task signal of any of these five ($\rho^* = 0.064$ at $t^* = 16$, $p < 0.0001$, with both local ($+0.035$) and role ($+0.047$, $p = 0.0008$) scalars positive and significant) — structure here carries real information about labels beyond both features and 1-hop locality, genuinely delayed rather than immediately local and not purely role-driven either: this is the clearest instance in the suite of β's fourth pattern, *structure-explained*. Flickr shows the same pattern more weakly: a low but real task signal ($\rho^* = 0.034$ at $t^* = 8$) with a significant role-mediated component ($+0.035$, $p = 0.014$) riding on top of a much smaller alignment curve than BlogCatalog's. Chameleon-Filtered is a case worth flagging rather than smoothing over: its alignment curve peaks early and strongly ($\rho^* = +0.252$ at $t^* = 0.25$, $p < 0.0001$) while its local scalar is *strongly negative* (-0.231, $p < 0.0001$) at essentially the same diffusion scale — the balanced edge-vs-non-edge diagnostic and the diffusion-kernel curve disagree in sign at $t \approx 0.25$, which is worth a second look (e.g. against the underlying pair sample) before treating either number in isolation as the dataset's "local" behavior. Squirrel-filtered is the most striking result in the full 22-dataset suite: its $\rho^*_{\text{align}} = 0.482$ at $t^* = 8$ is the single largest alignment value

observed anywhere, more than 50% larger than the next-strongest (CS, 0.307) — despite Squirrel-filtered having near-zero adjusted homophily (0.005) and near-zero label informativeness (Table 4). This is a clean demonstration that the RAP diagnostic detects structure–feature correspondence that is entirely invisible to homophily-style statistics: the graph's structure strongly predicts *feature* similarity even though it says nothing about *label* similarity. Consistent with this, Squirrel-filtered's ρ_{task} is significantly *negative* ($\rho^* = -0.136$, $p < 0.0001$, local and role scalars both negative and significant) — structure is actively anti-informative about labels beyond features, matching the adversarial relabeling construction used to build the Platonov et al. (2023) filtered benchmarks.

Tolokers is the one heterophilous-suite dataset whose curves are not simply flat. Its $\rho_{\text{align}}(t)$ starts negative (dipping to -0.035 at $t = 2$, despite a positive local scalar of $+0.263$), crosses zero between $t = 12$ and $t = 16$, and ends positive ($+0.018$ at $t = 40$) — a genuine sign flip with scale that none of Roman-empire, Amazon-ratings, Minesweeper, or Questions show. Its $\rho_{\text{task}}(t)$ has the strongest peak of any non-airport dataset in the suite ($\rho^* = 0.156$ at $t^* = 5$, role scalar $+0.076$ dominating a negative local scalar of -0.017), but that peak does not decay smoothly: between $t = 12$ (0.142) and $t = 16$ (0.042) it drops more than threefold in a single grid step, a cliff rather than a slope. This is plausibly a real regime change — a role-driven signal that is strong at moderate t giving way to something else at longer range — but a sharp single-seed discontinuity like this one is exactly the kind of feature worth confirming against a second seed or the underlying labeled-pair sample before treating its precise location as more than approximate. Either reading is consistent with the architecture result: role-augmented SAGE and the Transformer are both comfortably ahead of GCN and APPNP (84–86% vs. 71–76%, Table 9), i.e. the role-mediated signal the curve identifies is real regardless of exactly where its strength changes.

The three WebKB webpage graphs (Texas, Cornell, Wisconsin) behave distinctly from every other family. They are the most strongly heterophilous datasets in the suite by adjusted homophily (-0.33 to -0.21), and their small size (under 300 nodes each) limits the diffusion-scale grid to $t \leq 12$. Texas's $\rho_{\text{align}}(t)$ is nearly flat and small across the whole grid (0.13–0.15, local scalar -0.049); its $\rho_{\text{task}}(t)$ is largest at the shortest scale tested ($\rho^* = 0.156$ at $t^* = 0.25$) and decays smoothly as t grows (to 0.084 by $t = 12$) — a real but purely local task signal that gets *weaker*, not stronger, with propagation range, the opposite shape from Brazil's role-aligned curve. Cornell's $\rho_{\text{align}}(t)$ is negative at short range (-0.067 at $t = 0.25$: direct neighbors are *less* feature-similar than random pairs, consistent with its strongly negative homophily) and crosses to weakly positive by $t = 12$ ($+0.026$); its $\rho_{\text{task}}(t)$ peaks at intermediate scale ($\rho^* = 0.077$ at $t^* = 5$) with a positive role scalar ($+0.077$) despite a negative local scalar (-0.064), so what little task signal there is is role-mediated rather than local. Wisconsin shows the same short-range-opposition-to-longer-range-alignment crossover as Cornell, more pronounced (ρ_{align} from -0.012 at $t = 0.25$ to $+0.138$ by $t = 12$, still rising — structurally the same late-rising family as CS, Computers, and Photo, just compressed onto a much smaller grid by the graph's size), while its $\rho_{\text{task}}(t)$ stays close to zero at every scale ($\rho^* = 0.014$). Despite this real, if modest, structure in the curves, all three datasets are won by the feature-only MLP by a wide margin (Appendix C.3, Table 9): whatever task signal the curves detect at short or intermediate range is too weak, or too entangled with genuinely negative homophily at 1–2 hops, to be worth the risk of propagating at all. The curves correctly identify these as low- rather than merely negative-value cases for structure, even where the raw sign of ρ_{align} briefly flips.

Across all 22 datasets, every one of β's four task-conditioning patterns appears in at least one real problem. *Feature-explained* (strong ρ_{align}, near-zero ρ_{task}) is cleanest in CS, and present more mildly in CiteSeer, PubMed, and Cora. *Structure-explained* (a genuinely delayed, long-range task signal that is neither purely local nor purely role-driven) is clearest in BlogCatalog, and present more weakly in Flickr. *Opposed* (ρ_{task} significantly negative) is unambiguous in Squirrel-filtered and present more mildly in Amazon-Computers. *Role-aligned* (task signal carried by the role scalar, largely or entirely absent any structure–feature alignment) is exact in Brazil, Europe, and USA, and present as a component — alongside other structure — in Tolokers and Cornell. This is the sense in which β's dictionary construction, despite adding nothing as a compressed nine-dimensional feature (§4.2), correctly anticipates the qualitative diversity this 22-dataset suite actually contains: the categories were the right idea, just the wrong output format for a low-capacity selection predictor.

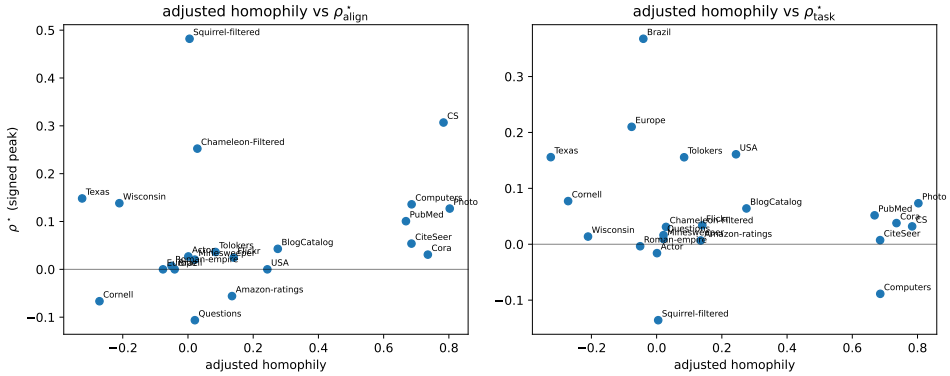


Figure 7: **Alignment curves are related to, but not determined by, adjusted homophily.** Peak $\rho^{\star}_{\text{align}}$ (left) and $\rho^{\star}_{\text{task}}$ (right) vs. adjusted homophily, over all 22 real datasets. Squirrel-filtered has near-zero homophily but among the largest alignment peaks; Brazil/Europe/USA have near-zero ρ_{align} (no real features) but the largest ρ_{task} peaks in the suite.

Correlation with conventional statistics. Table 17 reports Spearman correlations, over all $n = 22$ real datasets, between the eight curve-derived statistics and the two conventional statistics motivated by prior theory (adjusted homophily and label informativeness), pre-registered before the dataset suite was extended to its current size; Figure 7 shows the corresponding scatter for the two peak statistics. Of the six pre-registered pairs, three survive Benjamini–Hochberg FDR correction at $q < 0.05$ when corrected only across these six tests: t^{*}_{align} correlates positively with both label informativeness ($\rho = 0.617$) and adjusted homophily ($\rho = 0.576$) — datasets with more homophilous, more label-informative structure tend to have *later*-peaking alignment curves, which, combined with the regime split above, reflects that the specific homophilous graphs in this suite with the strongest label homophily (CS, Computers, Photo, PubMed) happen to be exactly the ones whose structure–feature correspondence is slow and multi-hop-built, not a general "homophily causes early alignment" relationship (CiteSeer and Cora, in the same homophily range, peak early). The role scalar correlates negatively with label informativeness ($\rho = -0.490$) and survives correction; the same relationship against adjusted homophily ($\rho = -0.392$) does not survive at $n = 22$, though it did at an earlier, smaller pre-registration — worth reporting honestly as a relationship that weakened, not disappeared, under a larger and more diverse dataset suite. For contrast, the second panel of Table 17 shows the strongest hits from the full, unrestricted 168-pair sweep (all 8 curve statistics against all 21 conventional statistics); when FDR is computed across all 168 tests instead of just the 6 pre-registered ones, *nothing* survives at $q < 0.05$ (best $q = 0.188$). Two of the top exploratory hits are in fact the pre-registered pairs reappearing at ranks 2 and 3, which is reassuring, but any claim drawn only from the exploratory sweep — e.g. the feature–edge-distance correlations at ranks 1 and 4 — should be reported as suggestive rather than significant, and ideally validated on a further-extended dataset suite before being treated as a finding in its own right.

D Reading Alignment Curves

The two alignment curves of §3.4 are meant to be read, not merely computed. This appendix fixes a small vocabulary for that reading: the summary statistics we extract, the shapes a curve can take, and what each implies about the computation a problem supports. Appendix C.6 applies this vocabulary dataset by dataset. Throughout, the reading is a *prior over architecture choice*, not a proof: it is meant to narrow a candidate set before training, not to replace validation.

D.1 Summary statistics

Each curve is summarized by a signed peak and the scale at which it occurs,

$$t^{\star} = \arg\max_{t \in \mathcal{T}} |\rho(t)|, \qquad \rho^{\star} = \rho(t^{\star}),$$

Table 17: Pre-registered correlations between alignment-curve statistics and the two theory-motivated conventional statistics (adjusted homophily, label informativeness), Spearman ρ over $n = 22$ real datasets. q is Benjamini–Hochberg FDR computed over these 6 pre-registered tests only.

Curve stat.	Conventional stat.	ρ	p	q (FDR-6)	Survives $q<0.05$
t^*_{align}	label informativeness	0.617	0.0022	0.013	yes
t^*_{align}	adjusted homophily	0.576	0.0050	0.015	yes
$\text{role}_{\text{align}}$	label informativeness	−0.490	0.0206	0.041	yes
$\text{role}_{\text{align}}$	adjusted homophily	−0.392	0.0714	0.099	no
ρ^*_{align}	label informativeness	0.378	0.0827	0.099	no
ρ^*_{align}	adjusted homophily	0.266	0.2310	0.231	no

For contrast: the strongest hits from the full unrestricted 168-pair sweep (all 8 curve statistics × all 21 conventional statistics), with FDR computed across all 168 tests. None survive at $q < 0.05$ once every pairwise comparison is corrected for; the two exceptions are the pre-registered pairs above, which re-appear here at rank 2 and rank 3.

Curve stat.	Conventional stat.	ρ	p	q (FDR-168)
$\text{local}_{\text{align}}$	feat–edge distance	−0.644	0.0012	0.188
t^*_{align}	label informativeness†	0.617	0.0022	0.188
t^*_{align}	adjusted homophily†	0.576	0.0050	0.279
$\text{role}_{\text{align}}$	feat–edge distance	−0.545	0.0088	0.302
$\text{local}_{\text{task}}$	adjusted homophily	0.542	0.0091	0.302
ρ^*_{task}	density	0.531	0.0109	0.302
t^*_{align}	spectral gap (6th)	−0.515	0.0142	0.302
ρ^*_{align}	feat–edge distance	0.507	0.0160	0.302
$\text{role}_{\text{task}}$	density	0.503	0.0170	0.302

†Duplicates a pre-registered pair from the table above.

reported for the structural and task curves separately alongside the off-axis scalars ρ^{local} and ρ^{role} (Table 16). We additionally track the saturation scale t_{sat} at which a curve stops changing beyond its resampling spread; it is not itself a tabulated per-dataset number, but governs two things reported elsewhere: the decision to extend the diffusion-scale grid for the five datasets still rising at $t = 40$ (Appendix C.6), and the caution attached to any $t^\star$ falling in Table 20's $t > 20$ band. The three quantities answer different questions: $\rho^\star$ how strongly and in which direction, $t^\star$ at what relational range, t_{sat} where the estimate stops being informative.

Magnitudes do not transfer across graphs. Both curves are rank correlations over sampled pairs, and their scale depends on the pair population — in particular on edge density, which varies by orders of magnitude across benchmarks. Signs and shapes are comparable across datasets; raw magnitudes are not. We therefore report $\rho^\star$ in units of its own bootstrap standard deviation when comparing across datasets, and treat a curve whose entire range lies within its resampling band as *null* rather than as weakly shaped. This matters in practice: per-panel autoscaling makes a curve spanning 0.01 look like one spanning 0.3.

D.2 Reading the two curves jointly

The structural curve alone cannot recommend a computation: a graph may organize node attributes beautifully and still contribute nothing to the label. The prescription comes from the two curves together.

D.3 From diffusion time to propagation depth

Since $\exp(-tL) = e^{-t}\sum_k \frac{t^k}{k!}\hat{A}^k$, the heat kernel weights walk lengths by a Poisson(t) distribution, so t is nominally the expected number of hops. The distance, however, compares heat *profiles* rather than walks, and long walks on a small-diameter graph contribute nothing discriminative. Empirically the resolved neighborhood is roughly half the nominal one (Table 20).

Table 18: **A vocabulary of curve shapes:** Shapes a curve can take across the propagation schedule, and what each indicates about the range at which relational organization lives.

Shape	Name	Reading
flat, within band	*null*	that geometry carries no monotone signal
flat, outside band	*scale-free*	organization identical at every range; depth irrelevant
monotone rising	*coarse*	information lives at long range; favors diffusion or decoupled propagation
monotone falling	*fine*	information is short-range; depth dilutes it
interior maximum	*band-limited*	a single characteristic scale; match the receptive field to $t^\star$
crosses zero at t_0	*regime boundary*	opposite mechanisms below and above t_0; the strongest available signal
plateaus after t_{sat}	*saturated*	diffusion has mixed; points beyond t_{sat} are not informative

Table 19: Joint reading of the two curves at the scale where each is strongest. Rows 2–4 and 7 are the cases no single-scale, label-only descriptor can express.

ρ_{align}	ρ_{task}	Reading	Consistent with
$+$	$+$	classic homophily	low-pass propagation
$+$	≈ 0	features already encode the structure	feature-only model is competitive
$+$	$-$	structure organizes attributes but misleads on labels	ego–neighbor separation, high-pass
≈ 0	$+$	label signal is purely relational	propagation essential; largest expected gain
≈ 0	≈ 0	no monotone relational information	feature-only model
$-$	$-$	attribute and label opposition	signed or high-pass propagation
$-$	$+$	unlike attributes, shared labels	role-based or structural encoding

Two comparisons refine this, though — like the table itself — they are motivated hypotheses about how the two curves should combine, not claims we test below. When $t^\star_{\text{task}}$ falls well below $t^\star_{\text{align}}$, the labels need less range than the attributes support and stacked layers suffice. When ρ_{task} is still rising at $t^\star_{\text{align}}$, the task wants propagation range without representational depth, which favors decoupled architectures. And when the two curves disagree in *sign*, the task curve governs: attribute organization at a scale where the label signal is negative is not a reason to propagate there.

Empirical test: does reading depth off $t^\star$ help? We tested this on the three real datasets with the latest alignment peaks (CS $t^\star$=60, PubMed $t^\star$=40, Photo $t^\star$=90), comparing the fixed default (2-layer GCN, K=10 APPNP) against a depth read off $t^\star$ via the resolved-hops column above (GCN capped at 8 layers for training stability; APPNP hops uncapped, since decoupled propagation carries no over-smoothing risk). All three fall in Table 20's $t > 20$ band, which itself says to verify against t_{sat} first; we ran the naive reading anyway, to see what ignoring that caution costs.

The result does not support "read $t^\star$ off the curve and get a free accuracy gain," but the two architectures test different things. Stacking a plain GCN to the depth $t^\star$ nominally implies is actively harmful on all three datasets (up to -11.6 pp, with a training collapse on one PubMed seed). This is consistent with the caveat above, but it is not evidence *specific to* $t^\star$: a plain GCN degrades at 8 stacked layers on most graphs regardless of what any curve says, so this result mainly confirms that over-smoothing is real, not that $t^\star$ correctly locates it. Reading APPNP's hop count off $t^\star$ is directionally positive on all three datasets but by a trivial margin ($+0.01$ to $+0.21$ pp, well within the seed-to-seed standard deviation of the default configuration in Table 9), and APPNP is already known to be robust to large K regardless of dataset, so a small positive delta here is weak evidence that $t^\star$ *specifically* — rather than any sufficiently large K — is doing the work.

We report both directions plainly and do not currently claim $t^\star$ gives actionable depth guidance beyond model-family selection (decoupled vs. stacked). A test that would actually isolate $t^\star$'s contribution — left for future work — is a *mismatched*-depth design: take the $t^\star$-derived hop count from one dataset and apply it to a different dataset with a substantially different $t^\star$, within a single

Table 20: Nominal walk length under $\exp(-tL)$ and the propagation depth we recommend reading off $t^\star$, **for architectures whose propagation is decoupled from representational depth** (e.g. APPNP); the empirical test below finds this reading unreliable for stacked architectures. The resolved neighborhood is capped by the graph's mean shortest path $\bar{\ell}$; once $t \gtrsim 2\bar{\ell}$ the ranking is dominated by degree rather than proximity and $t^\star$ should be treated as an artifact.

t	mean hops	90% mass	$\approx$ resolved hops	recommended depth
≤ 1	≤ 1	0–3	0–1	one layer
2–4	2–4	0–8	1–2	two layers
6–10	6–10	2–16	3–5	3–4 layers, or decoupled propagation
12–20	12–20	7–29	6–10	decoupled propagation, not stacked layers
> 20	> 20	–	–	verify against t_{sat} before use

Table 21: Default vs. $t^\star$-derived depth, test accuracy (mean over 5 seeds). Δ is $t^\star$-derived minus default.

Dataset	$t^\star$	GCN default	GCN $t^\star$-derived	Δ_{GCN} (pp)
CS	60	94.00% (2L)	90.87% (8L)	-3.13
PubMed	40	87.46% (2L)	75.86% (8L)[†]	-11.60
Photo	90	93.32% (2L)	87.42% (8L)	-5.90

Dataset	$t^\star$	APPNP default	APPNP $t^\star$-derived	Δ_{APPNP} (pp)
CS	60	94.22% (K=10)	94.43% (K=30)	$+0.21$
PubMed	40	87.10% (K=10)	87.15% (K=20)	$+0.05$
Photo	90	93.60% (K=10)	93.62% (K=45)	$+0.01$

[†] Includes one seed that collapsed to 40.6% accuracy at 8 stacked layers; the other four seeds average 84.7%, still 2.8 pp below the 2-layer default.

decoupled architecture, across a range of datasets spanning both short and long $t^\star$. If $t^\star$ carries real information, matched depth should beat mismatched depth; if APPNP is simply insensitive to K at this scale, matched and mismatched depth should perform about the same — a distinction the present three-dataset, single-direction test cannot make.

D.4 Scope

Three limits are worth stating plainly, since they bound every reading above.

Both curves are *monotone and pairwise*. A task whose labels depend non-monotonically on the neighborhood — a count, a parity, a threshold — can produce $\rho_{\text{task}} \approx 0$ while a message-passing model performs well. A null reading is evidence of no monotone pairwise signal, not evidence that the graph is useless.

Both curves are *global averages*. A graph whose regions carry opposite mechanisms averages toward zero; this is the spatial-mixture case, and it is why the local coordinates of Appendix B.8 are estimated separately rather than read off the global curve.

Finally, ρ_{task} *consumes labels* and inherits the variance of the endpoint-disjoint cross-fitting of Appendix B.4. On small or sparsely labelled graphs its sign is far better determined than its peak location, and we report the resampling distribution of $t^\star$ rather than a point estimate wherever the modal scale is not stable (Appendix B.7).