Testing Statistical Dependence in Labeled Graphs under Mismatches

Nikolaos Papagiannis, Vasam Manjveekar Prabantu, Ananth Grama, Wojciech Szpankowski
Purdue University
West Lafayette, Indiana, USA
npapagia@purdue.edu,vmanjvee@purdue.edu,ayg@purdue.edu,szpan@purdue.edu

Abstract

Many real-world systems—ranging from protein structures to financial networks—are naturally represented as labeled graphs, where both topology and node attributes carry critical information. A fundamental question in analyzing such data is whether two graphs (or subgraphs) exhibit statistical dependence, which may indicate shared generative mechanisms or latent interactions. Unlike classical dependence testing, the graph setting introduces unique challenges: dependence can manifest through structural similarity, label correlation, or their interplay, potentially reinforcing or obscuring each other. We propose a novel and practical framework for dependence testing in labeled graphs via mutual information over a structure-weighted joint label distribution. This approach jointly captures topological and attribute-based signals while remaining robust to imperfect or noisy node alignments. We provide theoretical guarantees with explicit error bounds and validate our method on both synthetic and real-world datasets, including protein structures from the lipocalin family, and recurring motifs in the Cora citation network. Our results demonstrate that the proposed test is a statistically sound and an effective tool for uncovering nontrivial dependencies in graph data.

1 Introduction

Labeled graphs serve as models for various systems, including computer networks, financial transactions, protein interaction networks, and social interactions, among others. A central question in these domains is whether two graphs, or two subgraphs within a larger network, exhibit statistical dependence, that is, similarity unlikely to arise by chance and potentially indicative of interaction or recurring patterns. Formally, we represent a labeled graph as $G = (V, E, \ell)$, where V is the set of nodes, $E \subseteq V \times V$ is the set of edges connecting pairs of nodes, and $\ell : V \to \mathcal{A}$ assigns to each node a label drawn from an alphabet $\mathcal{A}$.

Statistical dependence between graphs can arise from two distinct but complementary sources of information. The first is *structure*, where the goal is to determine whether the connectivity patterns of two graphs were generated independently or in a correlated manner. A central challenge here is to identify the latent correspondence between nodes, known as the (sub)graph matching problem. Computing an optimal correspondence is equivalent to solving the Quadratic Assignment Problem, which is NP-hard [6].

The second source of dependence arises from the *labels*, where each node is assigned a value from the alphabet $\mathcal{A}$. By considering the collection of labels independently from the graph structure, one obtains strings over $\mathcal{A}$. As with structure, statistical dependence in this setting requires accounting for a latent correspondence between nodes to ensure a meaningful alignment of labels. Dependence in this case manifests through nontrivial correlations or repeated patterns in the distribution of labels.

Structure and labels therefore provide distinct signals of dependence, yet they are intrinsically linked in the generative process of graphs. Assessing them separately fails to capture their complementary nature. The objective, then, is to develop a unified framework that evaluates dependence jointly, requiring evidence of both structural similarity and label agreement in order to assess statistical dependence.

Contributions. In this paper, we study the problem of quantifying statistical dependence in labeled graphs, where both connectivity and node labels may provide complementary signals. Our contributions are twofold:

1. We introduce a new *statistical measure* of correlation in labeled graphs, defined as a structure–weighted mutual information that integrates both structural and label information under approximate alignments. This measure remains robust even in the presence of structural mismatches.

2. Building on this measure, we develop a *statistical test* that provides rigorous guarantees under independence. We further demonstrate the effectiveness of this test through experiments on synthetic graphs and real datasets, including protein networks and the Cora citation graph.

Beyond its theoretical contribution, our framework is intended as a statistically calibrated validation tool for graph mining pipelines. It is complementary to graph matching, graph kernels, and graph neural network methods: rather than learning an alignment or optimizing a predictive objective, it takes a candidate alignment as input and tests whether the aligned graphs exhibit significant joint structure–label dependence. This makes the proposed test useful as a post-hoc validation step in applications such as motif discovery, anomaly detection, biological network analysis, and knowledge graph reasoning, where candidate matches must be assessed with explicit false-positive control.

Prior Literature. Most prior work on graph dependence has focused on the *structural* aspect of the problem [10]. A large body of work has analyzed the graph matching problem in the context of Erdős–Rényi graphs, including [2, 11–13, 17], [14–16, 19] as well as [36]. Despite the simplicity of this model, where edges appear independently with a fixed probability, graph matching remains challenging, and significant effort has been devoted to determining conditions under which exact or partial recovery of the latent alignment is possible.

Beyond Erdős–Rényi graphs, extensions have been developed for more structured settings. For example, Shen et al. [29] propose correlation tests for binary graphs with community structure. More

generally, nonparametric approaches based on distance correlation have been explored. Shen et al. [30] introduced *multiscale graph correlation* (MGC), which aggregates local correlations across multiple scales to detect complex nonlinear dependencies. Building on this, Chung et al. [8] combined MGC with optimal transport Procrustes alignment to design valid two-sample tests for graphs.

A complementary line of work arises from the classical graph isomorphism problem, which studies structural equivalence in the absence of randomness. Foundational contributions include the Weisfeiler–Leman color-refinement procedure [35] and its limitations established by Cai, Fürer and Immerman [7]. Algorithmically, Babai [4] achieved a landmark quasipolynomial-time GI algorithm, while McKay and Piperno [26] developed highly effective canonical-labeling methods for large real-world graphs. Despite this progress, existing literature underscores the inherent difficulty of inferring statistical dependence purely from graph topology.

On the label side, related work arises from string similarity and sequence dependence. Cohen et al. [9] provide a comparative study of string distance metrics, widely used in record linkage though not designed specifically for correlation testing. Marteau [24] introduced *sequence covering similarity*, a measure of string similarity based on covering relations that can be computed more efficiently than classical edit distances. In bioinformatics, correlation metrics for genomic sequences have long been employed, for example in [1], where mutual information–based methods are used to detect dependence between DNA or protein sequences.

These contributions highlight the substantial progress made on structure-only or label-only dependence. A related direction concerns attributed alignment methods [37, 39, 40], which incorporate both structural and attribute information to improve correspondence. What remains lacking is a unified statistical framework for assessing dependence in labeled graphs under imperfect alignments.

2 Problem Formulation

Let model $\mathcal{F}$ denote the generative mechanism for the structure of an unweighted undirected graph $G = (V_1, E_G)$, where $E_G \subseteq V_1 \times V_1$. The same model produces a second graph $H = (V_2, E_H)$ with $E_H \subseteq V_2 \times V_2$, under the constraint $|V_1| = |V_2| = n$. Formally, we write $(G, H) \sim \mathcal{F}$. We distinguish two structural scenarios: (a) *structural independence*, where H is generated independently of G, and (b) *structural dependence*, where H is generated as a correlated or perturbed version of G.

The adjacency matrix of G is defined by

$$\mathbf{A}^G_{uv} = \begin{cases} 1 & \text{if } (u,v) \in E_G, \\ 0 & \text{otherwise,} \end{cases} \qquad u, v \in V_1,$$

with an analogous definition for $\mathbf{A}^H$ over V_2.

Each node of G is assigned a label from a finite alphabet $\mathcal{A}$ via a function $\ell_G : V_1 \to \mathcal{A}$. For H, we consider two labeling regimes: (a) *label independence*, where $\ell_H : V_2 \to \mathcal{A}$ assigns labels independently of those in G, and (b) *label dependence*, where ℓ_H produces labels that are correlated with the labels in G.

In our framework, statistical dependence requires both *structural dependence* and *label dependence* to be present. The test is explicitly designed to detect concurrent rather than marginal dependence, reflecting the principle that structural and label signals complement one another in establishing genuine correlation between graphs.

We assume access to an alignment $\pi : V_1 \to V_2$, which specifies a one-to-one correspondence between nodes of G and H. This allows comparison between node $u \in V_1$ in G and its aligned counterpart $\pi(u) \in V_2$ in H. Crucially, we impose no optimality requirement on π: it need not minimize structural discrepancies (as in Erdős–Rényi graph matching literature) or maximize label agreement. Our framework accommodates arbitrary, potentially imperfect alignments, reflecting realistic scenarios where the ground-truth correspondence is noisy or partially known.

We emphasize that our objective is not to solve the graph alignment problem. The alignment π is treated as an input, produced for example by a graph matching heuristic, or a learned model. Our test then evaluates whether the dependence induced by this candidate alignment is statistically significant. Poor alignments increase structural mismatch and therefore reduce the resulting statistic $I(\hat{Q})$, making the method a suitable post-hoc validation tool for candidate alignments.

2.1 Basic Results

Let $\mathbf{A}^G_u$ denote the u-th row of $\mathbf{A}^G$ and $\mathbf{A}^H_{\pi(u)}$ the corresponding row of $\mathbf{A}^H$ under the vertex permutation π. For each $u \in V_1$, define

$$w_u := 1 - \frac{1}{n}\left\|\mathbf{A}^G_u - \mathbf{A}^H_{\pi(u)}\right\|_1. \tag{1}$$

Thus $w_u \in (0, 1]$ quantifies how well the neighborhood of u in G matches the neighborhood of $\pi(u)$ in H, with higher values indicating stronger structural agreement.

We define the *empirical joint label distribution* over $\mathcal{A} \times \mathcal{A}$ as

$$\hat{P}(a, b) := \frac{1}{n} \sum_{u \in V_1} \mathbf{1}_{\ell_G(u)=a} \cdot \mathbf{1}_{\ell_H(\pi(u))=b}, \tag{2}$$

with its *population counterpart* $P(a, b) := \mathbb{E}_{\mathcal{F}}[\hat{P}(a, b)]$, where the expectation is taken with respect to the randomness of model $\mathcal{F}$.

To incorporate structural information, we extend the alphabet to $\mathcal{A}' := \mathcal{A} \cup \{\star\}$. The auxiliary symbol $\star$ serves as a sink state that absorbs probability mass associated with structurally inconsistent alignments, ensuring that mismatched neighborhoods are explicitly represented in the distribution. The *structure–weighted empirical joint label distribution* is then

$$\hat{Q}(a, b) = \frac{1}{n} \sum_{u \in V_1} \Big[f(w_u) \cdot \mathbf{1}_{\ell_G(u)=a} \cdot \mathbf{1}_{\ell_H(\pi(u))=b} + \big(1 - f(w_u)\big) \cdot \mathbf{1}_{a=\star} \cdot \mathbf{1}_{b=\star} \Big],$$

where $f : \{j/n : j = 1, \ldots, n\} \to (0, 1]$ is a monotone increasing function with $f(1) = 1$ that controls the trade-off between structural consistency and label agreement. Larger values of $f(w_u)$ assign greater weight to labels, thereby reducing the penalty for structural mismatches. Its population counterpart is: $Q(a, b) := \mathbb{E}_{\mathcal{F}}[\hat{Q}(a, b)]$.

For the remainder of the paper, we adopt the linear choice $f(w) = w$, which yields

$$\hat{Q}(a,b) = \frac{1}{n} \sum_{u \in V_1} \Big[w_u \cdot \mathbf{1}_{\ell_G(u)=a} \cdot \mathbf{1}_{\ell_H(\pi(u))=b} + (1-w_u) \cdot \mathbf{1}_{a=\star} \cdot \mathbf{1}_{b=\star} \Big]. \tag{3}$$

Next, for $u, v \in V_1$ we define the binary edge–mismatch indicator under the alignment π by:

$$\hat{m}_{uv} := \left| \mathbf{A}^G_{uv} - \mathbf{A}^H_{\pi(u)\,\pi(v)} \right|.$$

The corresponding normalized mismatch total is

$$\hat{M} := \frac{1}{2n^2} \sum_{u \in V_1} \sum_{v \in V_1} \hat{m}_{uv} = \frac{1}{2n} \sum_{u \in V_1} \frac{1}{n} \left\| \mathbf{A}^G_u - \mathbf{A}^H_{\pi(u)} \right\|_1, \tag{4}$$

where the factor $1/2$ compensates for double-counting edge disagreements. For probabilistic analysis, we also introduce the random variables: $m_{uv} := \mathbb{E}_{\mathcal{F}}[\hat{m}_{uv}]$ and $M := \mathbb{E}_{\mathcal{F}}[\hat{M}]$, where the expectation is taken with respect to the randomness of the graph-generating model $\mathcal{F}$. We also write $M_{\min} := \min_{u \in V_1} \left\| \mathbf{A}^G_u - \mathbf{A}^H_{\pi(u)} \right\|_1$, $M_{\max} := \max_{u \in V_1} \left\| \mathbf{A}^G_u - \mathbf{A}^H_{\pi(u)} \right\|_1$.

The following lemma makes explicit the connection between normalized structural mismatches $\hat{M}$ and the gap between the unweighted and structure–weighted empirical joint label distributions. The proof is in the Appendix.

Lemma 2.1. *Let $\hat{P}$ and $\hat{Q}$ be the unweighted and structure–weighted empirical joint label distributions, and let $\hat{M}$ be the normalized mismatch total defined in* (4)*. Then*

$$\hat{M} = \frac{1}{2} \sum_{(a,b) \in \mathcal{A} \times \mathcal{A}} \Big(\hat{P}(a,b) - \hat{Q}(a,b) \Big).$$

2.2 Mutual Information

We recall the general definition of mutual information. Let X and Y be random variables taking values from finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, with joint distribution $R(x,y)$ and marginals $R_X(x) = \sum_{y \in \mathcal{Y}} R(x,y)$ and $R_Y(y) = \sum_{x \in \mathcal{X}} R(x,y)$. The mutual information between X and Y is defined as

$$I(X;Y) := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} R(x,y) \log \frac{R(x,y)}{R_X(x)\, R_Y(y)}.$$

By construction, $I(X;Y) \geq 0$, with equality if and only if X and Y are independent. Throughout, logarithms are taken in base 2. $I(X;Y)$ quantifies the amount of information that one variable provides about the other.

Specializing to our setting, we consider mutual information under both the unweighted and structure–weighted empirical distributions. For $\hat{P}$ we define

$$I(\hat{P}) := I(\hat{P}_G; \hat{P}_H) = \sum_{(a,b) \in \mathcal{A} \times \mathcal{A}} \hat{P}(a,b) \log \frac{\hat{P}(a,b)}{\hat{P}_G(a)\, \hat{P}_H(b)}, \tag{5}$$

and for $\hat{Q}$ we define

$$I(\hat{Q}) := I(\hat{Q}_G; \hat{Q}_H) = \sum_{(a,b) \in \mathcal{A}' \times \mathcal{A}'} \hat{Q}(a,b) \log \frac{\hat{Q}(a,b)}{\hat{Q}_G(a)\, \hat{Q}_H(b)}, \tag{6}$$

where $\hat{P}_G(a) = \sum_b \hat{P}(a,b)$, $\hat{P}_H(b) = \sum_a \hat{P}(a,b)$, and analogously for $\hat{Q}_G, \hat{Q}_H$.

We now illustrate the above preliminary discussion with the following example:

Example 2.2. Consider the following graph pair:

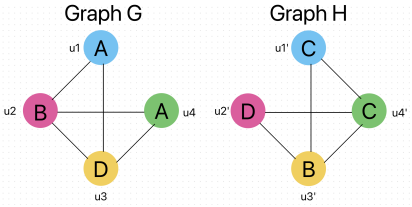


Figure 1: Graph pair example.

(1) Node mapping:

$$\pi = \{u_1 \mapsto u'_1,\ u_2 \mapsto u'_2,\ u_3 \mapsto u'_3,\ u_4 \mapsto u'_4\}.$$

(2) Node weights:

$$w_{u_1} = \tfrac{2}{4}, \quad w_{u_2} = w_{u_4} = \tfrac{3}{4}, \quad w_{u_3} = 1.$$

(3) Normalized mismatch count:

$$\hat{M} = \tfrac{1}{8}.$$

(4) Unweighted mutual information:

$$I(\hat{P}) = \sum_{(a,b) \in \mathcal{A} \times \mathcal{A}} \hat{P}(a,b) \log \frac{\hat{P}(a,b)}{\hat{P}_G(a)\hat{P}_H(b)} = 1.5.$$

(5) Weighted mutual information:

$$I(\hat{Q}) = \sum_{(a,b) \in \mathcal{A}' \times \mathcal{A}'} \hat{Q}(a,b) \log \frac{\hat{Q}(a,b)}{\hat{Q}_G(a)\hat{Q}_H(b)} = 1.98.$$

3 Main Results

In this section, we present our main theoretical findings, resulting in a computable threshold that assures, with high probability, that two graphs are correlated. We start with a proposition that bounds the difference between mutual information of $\hat{P}$ and $\hat{Q}$ via the mismatch function $\hat{M}$. The proof is in the Appendix.

Proposition 3.1. *Let $\hat{P}$ and $\hat{Q}$ be the unweighted and structure-weighted empirical joint label distributions, let $\hat{M}$ be the normalized mismatch total and let $M_{\min}$ and $M_{\max}$ denote the minimum and maximum per-node mismatch counts. Then*

$$I(\hat{Q}) - I(\hat{P}) \leq -\frac{M_{\min}}{n} I(\hat{P}) + \left(1 - \frac{M_{\min}}{n}\right) \cdot \left[-\frac{M_{\min}}{n} - 2\log\left(1 - \frac{M_{\max}}{n}\right) \right] + 2\hat{M} \log\left(\tfrac{1}{2\hat{M}}\right).$$

3.1 Structural Independence

In our first main result, we provide a high probability guarantee that structural dependence holds whenever empirical mutual information exceeds a computable threshold θ. More precisely, we

estimate

$$P\big(\text{declare dependence} \,\big|\, \text{structural independence}\big)$$
$$= P\big(I(\hat{Q}) > \theta \,\big|\, \text{structural independence}\big),$$

which is the error of declaring dependence where there is none. In other words, under structural independence, mutual information $I(\hat{Q})$ is expected to be zero, but random fluctuations may cause it to be positive. We want to make this error probability as small as possible.

THEOREM 3.2. *Let graphs $(G, H) \sim \mathcal{F}$ be sampled from a generative model $\mathcal{F}$ for which the normalized mismatch statistic $\hat{M}$ satisfies suitable concentration conditions under structural independence, including Erdős–Rényi graphs, block graph models, and their variants (see Remark 1 below for the precise condition, and [25, 38] for the concentration tools). Let $\hat{Q}$ denote the structure–weighted empirical joint label distribution. Let $M_{indep} := \mathbb{E}_{\mathcal{F}}[\hat{M} \mid \textit{independent structure}]$, where $\hat{M}$ is the normalized mismatch total. For $\theta > 0$, define the surrogate function:*

$$T(\hat{M}) = (1 - 2\hat{M}) \log\left(\frac{|\mathcal{A}|}{1 - 2\hat{M}}\right) - 2\hat{M}\log(2\hat{M}) \tag{7}$$

and define $\delta(\theta, |\mathcal{A}|)$ as the unique solution of $\delta(\theta, |\mathcal{A}|) = T^{-1}(\theta)$ for $0 < \hat{M} < \frac{1}{2(|\mathcal{A}|+1)}$ Then, for every $\theta > 0$ such that $0 \le \delta(\theta, |\mathcal{A}|) < M_{indep}$,

$$\mathbb{P}\Big(I(\hat{Q}) > \theta \,\Big|\, \textit{independent structure}\Big) \le \tag{8}$$
$$\exp\Big\{-4\big(M_{\text{indep}} - \delta(\theta, |\mathcal{A}|)\big)^2 n^2\Big\}. \tag{9}$$

PROOF. By (6) and the identity $\hat{Q}(\star, \star) = 2\hat{M}$ which follows from (3),

$$I(\hat{Q}) \;=\; \sum_{(a,b)\in\mathcal{A}\times\mathcal{A}} \hat{Q}(a,b) \log \frac{\hat{Q}(a,b)}{\hat{Q}_G(a)\hat{Q}_H(b)} + 2\hat{M}\log\Big(\frac{1}{2\hat{M}}\Big).$$

Thus, $I(\hat{Q}) > \theta$ implies

$$\sum_{(a,b)\in\mathcal{A}\times\mathcal{A}} \hat{Q}(a,b) \log \frac{\hat{Q}(a,b)}{\hat{Q}_G(a)\hat{Q}_H(b)} \;>\; \theta + 2\hat{M}\log(2\hat{M}). \tag{10}$$

Using $\sum_a \hat{Q}_G(a) = 1 - 2\hat{M}$, and $H(\hat{Q}_G) \le (1 - 2\hat{M}) \log\big(\frac{|\mathcal{A}|}{1-2\hat{M}}\big)$, as well as $I(\hat{Q}) \le H(\hat{Q}) - 2\hat{M}\log(2\hat{M})$, a necessary condition for (10) is:

$$T(\hat{M}) := (1 - 2\hat{M}) \log|\mathcal{A}| \;+\; (2\hat{M} - 1)\log(1 - 2\hat{M}) \;-\; 2\hat{M}\log(2\hat{M}) \;-\; \theta \;>\; 0. \tag{11}$$

from which we define $\delta(\theta, |\mathcal{A}|) = T^{-1}(\theta)$ and hence:

$$\{I(\hat{Q}) > \theta\} \;\subseteq\; \{\hat{M} \le \delta(\theta, |\mathcal{A}|)\}. \tag{12}$$

We observe that large deviations (LD) applies to the normalized mismatch total in (4). Since $|\hat{m}_{u,v}| \le 1$, the bounded difference condition or the Lipschitz condition holds. The only remaining issue is what conditions we need on $\hat{m}_{uv}$ to apply for LD. This holds for example for Erdős–Rényi graphs, block model graphs and their variants (see also Remark 1 below). Thus, since $\hat{M}$ is a $1/n^2$–Lipschitz of $\binom{n}{2}$ edge-mismatch variables, by McDiarmid's inequality [25] for any $t > 0$, we have

$$\mathbb{P}\Big(\hat{M} \le M - t \,\Big|\, \text{independent structure}\Big) =$$
$$\mathbb{P}\Big(\hat{M} \le M_{indep} - t\Big) \;\le\; \exp\{-4t^2n^2\}.$$

Since we are in the lower-tail regime $\delta(\theta, |\mathcal{A}|) < M_{indep}$, taking $t := M_{indep} - \delta(\theta, |\mathcal{A}|)$ yields

$$\mathbb{P}\Big(\hat{M} \le \delta(\theta, |\mathcal{A}|) \,\Big|\, \text{independent structure}\Big)$$
$$\le\; \exp\Big\{-4\big(M_{indep} - \delta(\theta, |\mathcal{A}|)\big)^2 n^2\Big\}.$$

Combining with (12) completes the proof. □

Remark 1. We address now, in some depth, the assumptions under which we can apply the concentration inequality on graphs. First, in settings where edge mismatches are fully independent, the conditions of the classical McDiarmid inequality hold directly. This includes Erdős–Rényi graphs, *noise-injected* models in which H is obtained from G by independently adding, deleting, or flipping edges, as well as inhomogeneous random graph families such as Chung–Lu and latent-space models with independent edges. In all these cases, each mismatch variable $\hat{m}_{uv}$ is independent of all others, so the standard bounded-difference argument applies without modification.

More generally, as shown by Zhang et al. (2019), a McDiarmid-type concentration bound continues to hold under *mild dependence*, provided that each mismatch variable depends on only a small number of other mismatches. This setting includes, for example, stochastic block models in which dependence is restricted within blocks, block-structured or "partly dependent" models of the type studied by Janson, and models whose edges satisfy a bounded-neighborhood dependency graph. In such cases, the mismatch variables admit a dependency graph of bounded maximum degree, and the resulting concentration bound retains the same functional form as the classical McDiarmid inequality, up to a constant factor depending on this degree.

When the dependence neighborhood is larger, the bound remains valid but acquires an explicit penalty factor. Specifically, if the dependency graph of the variables has maximum degree d, then

$$\mathbb{P}\Big(\hat{M} \le M - t \,\Big|\, \text{dependent structure}\Big) \le\; \exp\Big\{-\frac{2t^2n^2}{d+1}\Big\}.$$

Thus, whenever we have an estimate or upper bound on the size of the dependence neighborhood for each mismatch variable, the corresponding generalized McDiarmid-type inequality applies.

Remark 2. Let graphs $(G, H) \sim \mathcal{F}$ have adjacency matrices A^G and A^H. Under structural independence, M_{indep} can be expressed in terms of marginal edge probabilities as:

$$M_{\text{indep}} \;=\; \frac{1}{2n^2} \sum_{u\in V_1} \sum_{v\in V_1} \mathbb{P}\big(A^G_{uv} \neq A^H_{\pi(u)\pi(v)} \,\big|\, \text{independent}\big).$$

Independence implies factorization, so this expands to

$$M_{\text{indep}} \;=\; \frac{1}{2n^2} \sum_{u\in V_1} \sum_{v\in V_1} \Big[\mathbb{P}\big(A^G_{uv} = 1\big)\, \mathbb{P}\big(A^H_{\pi(u)\pi(v)} = 0\big)$$

$$+\mathbb{P}(A^G_{uv}=0)\,\mathbb{P}(A^H_{\pi(u)\pi(v)}=1)\Big].$$

In practice, these marginal probabilities can be replaced by empirical estimates obtained from the observed adjacency rows. For each $u \in V_1$, define:

$$\hat{p}^G_u := \frac{1}{n}\sum_{v\in V_1} A^G_{uv}, \qquad \hat{p}^H_{\pi(u)} := \frac{1}{n}\sum_{v\in V_2} A^H_{\pi(u)v},$$

that is, the normalized row sums of u in A^G and of $\pi(u)$ in A^H. This yields the empirical row-wise approximation

$$\tilde{M}_{\text{indep}} = \frac{1}{2n}\sum_{u\in V_1}\left(\hat{p}^G_u + \hat{p}^H_{\pi(u)} - 2\,\hat{p}^G_u\hat{p}^H_{\pi(u)}\right) + O(1/n^2).$$

3.2 Label Independence

We now turn our attention to label independence and establish a threshold on $I(\hat{Q})$ that guarantees label independence with high probability. The following result is our second main theoretical contribution. As before, our objective is to identify a threshold θ that minimizes the probability of erroneously declaring dependence when, in fact, no label dependence exists.

Theorem 3.3. *Let $\hat{Q}$ be the structure–weighted empirical joint label distribution, let $\hat{M}$ be the normalized mismatch total and let $M_{\min}$ and $M_{\max}$ denote the minimum and maximum per-node mismatch counts. Then for graph models sufficiently large to satisfy standard large deviation assumptions:*

$$\mathbb{P}\left(I(\hat{Q}) > \theta \,\middle|\, \textit{independent labels}\right) \le \exp\left(-\frac{(\ln 2)\,n}{1-\frac{M_{\min}}{n}}\left[\theta - \left(1-\frac{M_{\min}}{n}\right)\left(-\frac{M_{\min}}{n} - 2\log\left(1-\frac{M_{\max}}{n}\right)\right) - 2\hat{M}\log\left(\frac{1}{2\hat{M}}\right)\right]\right).$$

Proof. We follow the testing framework of [1] for independence on strings to bound $I(\hat{P})$. Note that $I(\hat{P})$ is—up to a negligible remainder—Pearson's χ^2 statistic scaled by $\frac{1}{2(\ln 2)\,n}$, with $(|\mathcal{A}|-1)^2$ degrees of freedom. This follows from a standard application of the multivariate central limit theorem for the joint empirical distribution showing that mutual information converges in distribution to a (scaled) χ^2 random variable. Let

$$Z := 2(\ln 2)\,n\,I(\hat{P}) \overset{d}{\Rightarrow} \chi^2\big((|\mathcal{A}|-1)^2\big),$$

and thus, for large n,

$$\mathbb{P}\Big(I(\hat{P}) > \theta \,\Big|\, \text{independent labels}\Big) = \mathbb{P}(Z > 2(\ln 2)\,\theta\,n) \le \exp\big(-(\ln 2)\,\theta\,n\big), \tag{13}$$

using the asymptotic fact that the χ^2 tail decays like $e^{-x/2}$.

By Proposition 3.1, this bound is then upgraded to $I(\hat{Q})$. Specifically, set

$$O_{a,b} = n\,\hat{P}(a,b), E_{a,b} = n\,\hat{P}_G(a)\hat{P}_H(b), \Delta_{a,b} = O_{a,b} - E_{a,b}$$

for $(a,b) \in \mathcal{A}\times\mathcal{A}$. By (5), we can rewrite the statistic as:

$$2(\ln 2)\,n\,I(\hat{P}) = 2\sum_{(a,b)\in\mathcal{A}\times\mathcal{A}} O_{a,b}\,\ln\frac{O_{a,b}}{E_{a,b}}.$$

Using the second–order Taylor expansion $\ln(1+u) = u - \frac{1}{2}u^2 + O(u^3)$ with $u = \Delta_{a,b}/E_{a,b}$ gives

$$2(\ln 2)\,n\,I(\hat{P}) = \sum_{(a,b)\in\mathcal{A}\times\mathcal{A}}\left(2\Delta_{a,b} + \frac{\Delta^2_{a,b}}{E_{a,b}}\right) + O\Big(\sum_{(a,b)\in\mathcal{A}\times\mathcal{A}}\frac{|\Delta_{a,b}|^3}{E^2_{a,b}}\Big).$$

The linear term vanishes because $\sum_{a,b}\Delta_{a,b} = 0$. Therefore,

$$2(\ln 2)\,n\,I(\hat{P}) = \sum_{(a,b)\in\mathcal{A}\times\mathcal{A}}\frac{\Delta^2_{a,b}}{E_{a,b}} + O\Big(\sum_{(a,b)\in\mathcal{A}\times\mathcal{A}}\frac{|\Delta_{a,b}|^3}{E^2_{a,b}}\Big).$$

Finally, Proposition 3.1 gives

$$I(\hat{Q}) \le I(\hat{P}) - \frac{M_{\min}}{n}\,I(\hat{P}) + \left(1-\frac{M_{\min}}{n}\right)\cdot\left[-\frac{M_{\min}}{n} - 2\log\left(1-\frac{M_{\max}}{n}\right)\right] + 2\hat{M}\log\left(\frac{1}{2\hat{M}}\right).$$

Hence

$$\big\{I(\hat{Q}) > \theta\big\} \subseteq \left\{I(\hat{P}) > \left(\theta - \left(1-\frac{M_{\min}}{n}\right)\cdot\left[-\frac{M_{\min}}{n} - 2\log\left(1-\frac{M_{\max}}{n}\right)\right] - 2\hat{M}\log\left(\frac{1}{2\hat{M}}\right)\right) \Big/ \left(1-\frac{M_{\min}}{n}\right)\right\}.$$

Plugging this threshold into the χ^2 tail bound from (13) yields the claimed bound. □

Finally, we combine structural and label independence to design a joint test for structure and labels. Fix $\alpha \in (0,1)$. Let $\theta_{\text{str}}(\alpha/2)$ be any threshold provided by Theorem 3.2 such that

$$\mathbb{P}(I(\hat{Q}) > \theta_{\text{str}}(\alpha/2) \mid \text{independent structure}) \le \alpha/2.$$

Next, let $\theta_{\text{lab}}(\alpha/2)$ be any threshold provided by Theorem 3.3 such that

$$\mathbb{P}(I(\hat{Q}) > \theta_{\text{lab}}(\alpha/2) \mid \text{independent labels}) \le \alpha/2.$$

Define

$$\theta^* := \max\{\theta_{\text{lab}}(\alpha/2),\, \theta_{\text{str}}(\alpha/2)\}. \tag{14}$$

Then, under the joint null of independent structure or independent labels,

$$\mathbb{P}\Big(I(\hat{Q}) > \theta^* \,\Big|\, \text{indep. structure} \vee \text{indep. labels}\Big) \le \alpha. \tag{15}$$

This, together with (14), yields the final theoretical threshold ensuring that two graphs can be considered potentially statistically correlated whenever the corresponding mutual information exceeds the threshold θ^*. We emphasize, however, that this is rigorously established only for certain graph models (e.g., Erdős–Rényi graphs). Nevertheless, as we demonstrate in the next section, our results extend effectively to more practical graph settings.

Remark 3. In this paper, we have only considered one type of error: when the structure or labels are independent but our test incorrectly declares them to be dependent. However, there is also a second type of error, namely when the structure or labels are dependent but our test incorrectly declares them to be independent. We briefly discuss this case for labels only. More precisely, the second type of error is defined as:

$$P\big(\text{declare independence} \,\big|\, \text{label dependence}\big) = P\big(I(\hat{Q}) < \theta \,\big|\, \text{label dependence}\big).$$

In this case, the probability no longer follows the χ^2 distribution but instead the normal distribution, as discussed in [1]. We combine this fact with the mismatch inequality analogous to Proposition 3.1

$$I(\hat{P}) - I(\hat{Q}) \;\le\; \frac{M_{\max}}{n}\, I(\hat{P}) - \left(1 - \frac{M_{\max}}{n}\right) \cdot \left(\log\left(1 - \frac{M_{\max}}{n}\right) + \frac{2M_{\min}}{n}\right) - 2\hat{M}\log\left(\frac{1}{2\hat{M}}\right).$$

Hence

$$\{\, I(\hat{Q}) \le \theta \,\} \;\subseteq\; \left\{ I(\hat{P}) \le \frac{\theta \;+\; F(M_{\min}, M_{\max}, \hat{M})}{1 - \frac{M_{\max}}{n}} \right\},$$

with

$$F(M_{\min}, M_{\max}, \hat{M}) := 2\hat{M}\, \log\left(\tfrac{1}{2\hat{M}}\right) + \left(1 - \frac{M_{\max}}{n}\right)\left(\log\left(1 - \frac{M_{\max}}{n}\right) + \frac{2M_{\min}}{n}\right)$$

Let I denote the true mutual information between aligned labels and let σ^2 be the variance of the log-likelihood ratio as in [1]. Then, under dependent labels,

$$\mathbb{P}\Big(I(\hat{Q}) \le \theta \,\Big|\, \text{dependent labels}\Big) \;\le\; \exp\left\{ - \frac{\left(I - \frac{\theta + F(M_{\min}, M_{\max}, \hat{M})}{1 - \frac{M_{\max}}{n}}\right)^2}{2\sigma^2}\; n \right\}.$$

This result shows that, under dependence, error probability decays exponentially in n, with additive terms accounting for structural mismatches. Since both I and σ^2 depend on the unknown generative mechanism, the bound should be viewed primarily as interpretive rather than directly computable. Consistent with [1], the choice of threshold θ reflects a tradeoff: smaller values increase the risk of spurious dependence under independence, while larger values increase the risk of overlooking true dependence. In practice, threshold selection is typically guided by controlling the error under independence, since the parameters governing the dependence case are not observable.

3.3 Practical Threshold Computation and Time Complexity

The above bounds yield a direct procedure for computing the decision threshold used by the test. First, the practitioner fixes a target significance level $\alpha \in (0, 1)$, for example $\alpha = 0.05$. We then invert the bounds in Theorems 3.2 and 3.3 with respect to the threshold parameter to obtain $\theta_{\text{str}}(\alpha/2)$ and $\theta_{\text{lab}}(\alpha/2)$ satisfying

$$\mathbb{P}\Big(I(\hat{Q}) > \theta_{\text{str}}(\alpha/2) \,\Big|\, \text{independent structure}\Big) \le \alpha/2$$

and

$$\mathbb{P}\Big(I(\hat{Q}) > \theta_{\text{lab}}(\alpha/2) \,\Big|\, \text{independent labels}\Big) \le \alpha/2.$$

The final threshold is

$$\theta^* = \max\{\theta_{\text{str}}(\alpha/2), \theta_{\text{lab}}(\alpha/2)\}.$$

Given an observed graph pair and candidate alignment π, the statistic $I(\hat{Q})$ is computed from the corresponding structure-weighted empirical distribution, and the test declares statistically significant dependence whenever

$$I(\hat{Q}) > \theta^*.$$

Given the alignment π, the dominant computational cost is computing the structural mismatch weights, which requires comparing aligned adjacency rows and costs $O(n^2)$ time for dense adjacency matrices. Once these weights are available, constructing $\hat{Q}$ requires one pass over the n aligned nodes, while computing $I(\hat{Q})$ requires summing over $\mathcal{A}' \times \mathcal{A}'$. Thus, these steps take $O(n + |\mathcal{A}|^2)$ time, and the overall runtime is dominated by the structural mismatch computation.

4 Experimental Results

Our objective in the experiments is not to outperform a competing method on a fixed benchmark, but to validate a statistical test for a setting where no directly comparable benchmark exists. To our knowledge, existing graph benchmarks and baselines address either structure-only similarity or attribute-based comparison, and do not provide a reference framework for joint structure–label dependence testing under a given alignment with a single computable significance threshold. Accordingly, we evaluate our method through controlled synthetic experiments with known ground truth and through real datasets where separation from null candidates and consistency with the theoretical threshold θ^* can be assessed.

To contextualize the proposed statistic, we also compare against representative partial baselines: Weisfeiler–Lehman subtree kernels [31, 35], which capture local structural similarity, and a chi-squared label-only test, which ignores graph topology. These baselines are not direct replacements for our method, since they do not provide calibrated Type I error control for joint structure–label dependence under a fixed alignment, but they help isolate the benefit of combining structural mismatch and label agreement.

In all experiments we set the overall significance level at $\alpha = 0.05$. While we present illustrative results in this section, a more comprehensive set of results is included in the Appendix.

4.1 Synthetic Erdős–Rényi Experiments and Representative Baselines

We first evaluate our framework on controlled synthetic data, where the ground-truth dependence structure is known. We use a correlated Erdős–Rényi model [23]: $G \sim \mathcal{G}(n, p_G)$ and $H \sim \mathcal{G}(n, p_H)$, with $n = 2000$ and $p_G, p_H \in [0.03, 0.07]$. In the correlated setting, each edge of H agrees with the corresponding edge of G with probability ρ_{str}, and is otherwise resampled from Bernoulli(p_H). Labels

are generated analogously: for each aligned pair $u, \pi(u)$, the label in H agrees with the label in G with probability ρ_{lab}, and is otherwise sampled uniformly from an alphabet of size $|\mathcal{A}| = 26$.

In the first experiment, we set $\rho_{\text{str}} = \rho_{\text{lab}} = 0.7$ and generate 12 graphs, including one correlated pair and ten mutually independent graphs. Figure 2 shows that the proposed statistic identifies the correlated pair as a clear spike above the null candidates, while Figure 3 visualizes representative subgraphs from this pair.

To contextualize the proposed statistic, we also compare against two representative partial baselines. First, we use the Weisfeiler–Lehman (WL) subtree kernel [31], which builds on the classical WL refinement procedure [35] and measures local structural similarity up to depth h. Since WL is not a hypothesis test, we calibrate it empirically by estimating its null distribution on independent Erdős–Rényi graph pairs and selecting the $(1-\alpha)$-quantile, with $\alpha = 0.05$, as the threshold. We report results for $h = 1$ and $h = 3$. Second, we include a chi-squared (χ^2) test for independence on aligned node labels, which captures label dependence but ignores graph topology.

Table 1 reports results as the planted agreement level ρ varies. Detection denotes the fraction of trials in which the corresponding statistic exceeds its threshold.

Table 1: Baseline comparison for synthetic Erdős–Rényi experiments with planted structure–label correlation. For each method, "Mean" denotes the average score over trials, θ denotes the corresponding decision threshold, and "Det." denotes the detection rate.

ρ	Our Method			WL ($h=1$)			WL ($h=3$)			Chi-squared (χ^2)		
	Mean	θ^*	Det.	Mean	θ	Det.	Mean	θ	Det.	Mean	θ	Det.
0.0	0.919	1.129	0.00	0.951	0.957	0.20	0.906	0.912	0.20	613.97	680.97	0.00
0.4	0.987	1.129	0.00	0.955	0.957	0.30	0.910	0.912	0.30	759.21	680.97	0.90
0.6	1.056	1.128	0.20	0.953	0.957	0.30	0.908	0.912	0.30	956.47	680.97	1.00
0.7	1.100	1.128	0.50	0.959	0.957	0.70	0.914	0.912	0.70	1093.01	680.97	1.00
0.8	1.158	1.127	1.00	0.954	0.957	0.40	0.909	0.912	0.30	1268.80	680.97	1.00
1.0	1.240	1.128	1.00	0.958	0.957	0.60	0.913	0.912	0.60	1579.00	680.97	1.00

Under independence ($\rho = 0$), our method produces no detections, while the WL baselines exhibit elevated false-positive rates despite empirical calibration. As ρ increases, the chi-squared test detects dependence early because it is sensitive to label correlation alone. In contrast, our method shows a sharper transition as joint structure–label agreement becomes strong enough to exceed the calibrated threshold. These results illustrate that structure-only and label-only baselines capture complementary but incomplete aspects of the dependence targeted by our test.

4.2 Protein Structures

4.2.1 Lipocalin Proteins. Complex biomolecules such as proteins perform functions determined not only by their amino acid sequence but also by higher-order structures and interactions. Sequence-based analyses, while indispensable, often obscure the fact that proteins with high sequence similarity should also be examined for conserved three-dimensional motifs to draw critical functional conclusions, with lipocalins providing a well-known example [32].

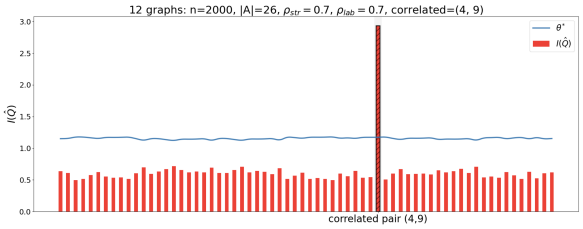


Figure 2: Mutual information across Erdős–Rényi graph pairs. Out of 12 graphs in total, a single correlated pair is included; the spike corresponds to this correlated pair with $\rho_{\text{str}} = \rho_{\text{lab}} = 0.7$. The blue line is the threshold θ^* computed according to (14).

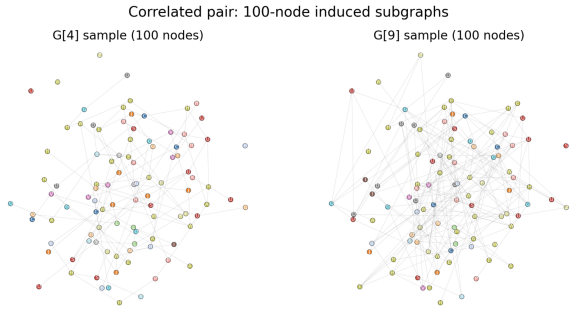


Figure 3: Visualization of the correlated Erdős–Rényi graph pair with $\rho_{\text{str}} = \rho_{\text{lab}} = 0.7$, showing representative subgraphs from each graph for clarity.

Representing proteins as sparse residue networks—with amino acids as labeled nodes and non-covalent interactions as edges—offers a principled abstraction for uncovering such motifs. Within this framework, conserved subgraph patterns capture recurring topologies of residue contacts [21] together with correlated amino acid labels, highlighting conservation that goes beyond either sequence or structure alone. Such motifs often mark functional regions such as active sites or binding pockets. For example, conserved motifs in these regions provide predictive markers of biochemical activity [22]. From this perspective, protein graphs exemplify sparse, heterogeneous networks with non-random recurrent patterns. Motif-centric analysis reveals hidden layers of conservation central to protein evolution, function, and design, while also connecting molecular biology with universal properties of networked systems [18].

We next demonstrate our theoretical framework on a dataset drawn from the lipocalin superfamily, each modeled as a labeled, unweighted graph. For a protein chain, we construct $G = (V_G, E_G, \ell_G)$ and $H = (V_H, E_H, \ell_H)$, where V_G and V_H are residues, E_G and E_H contain undirected edges between residues whose C_α atoms lie within 6 Å, and ℓ_G, ℓ_H assign amino acid identities from an alphabet $\mathcal{A}$ of size 20. To avoid trivial backbone contacts, edges are omitted whenever the sequence separation $|i - j| < 3$.

Figure 4: First protein in the most strongly correlated pair identified among 100 protein pairs. Visualization shows residue contacts.

Figure 5: Second protein in the most strongly correlated pair, aligned via TM-align.

To compare two graphs, we require an alignment $\pi : V_G \to V_H$ specifying residue correspondences. Biological graph alignment is an active area of research, with recent work, for example, studying learned approaches for identifying corresponding nodes across biological networks [27]. In practice, π is obtained using TM-align, a method based on three-dimensional structure. Our test then evaluates this alignment jointly with respect to both structure and labels. The *core* of aligned residues is defined as: $C = \{(u, v) : u \in V_G,\ v \in V_H,\ u \leftrightarrow v \text{ via TM-align}\}$. Restricting to the induced subgraphs on C yields $H_1 = G[V_G(C)]$ and $H_2 = H[V_H(C)]$, where testing is performed. In our dataset, the core size $|C|$ typically ranges between 100 and 200 residues.

For computational tractability, we select candidate pairs using simple global signatures and a relaxed length-ratio filter. Among the many pairs evaluated, a subset of 100 is presented for visualization, with several exceeding the significance threshold and one emerging as the strongest signal.

Figures 4 and 5 illustrate this most strongly correlated pair, while Figure 6 reports the corresponding mutual information spike compared with the null pairs.

4.2.2 ComPPI. Beyond individual protein structures, protein – protein interaction (PPI) networks provide a complementary view of cellular organization [5] by encoding functional relationships among proteins at the systems level. As in the structural setting, representing such systems as sparse, labeled graphs offers a principled abstraction for identifying conserved motifs. Here, nodes correspond to proteins and edges represent physical interactions, while node labels capture higher-level functional attributes such as subcellular localization and signaling pathway participation.

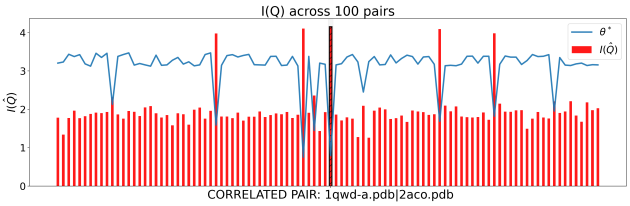


Figure 6: Mutual information values across 100 protein pairs. Several pairs exceed the threshold, with the spike corresponding to the most strongly correlated pair detected by the proposed test. The blue line is the threshold θ^* computed according to (14).

Within this framework, conserved subgraph patterns correspond to recurring interaction motifs whose structure is coupled with correlated functional annotations, revealing conservation that extends beyond either network topology or individual protein features alone. Such motifs often reflect biologically meaningful modules, including protein complexes and pathway-specific interaction patterns.

In this subsection, we apply our statistical framework to the ComPPI interactome [34] to assess the significance of structurally and label-wise correlated subgraphs. Starting from a seed subgraph that defines a reference motif, we identify candidate subgraphs across the network that are structurally similar and exhibit correlated node labels. To ensure connectivity, candidate subgraphs are generated using Breadth-First Search–based procedures.

We retain protein–protein interactions with confidence scores of at least 0.5, representing the likelihood of physical interaction. Each node is labeled with subcellular localization information provided by ComPPI, together with signaling cascade annotations obtained from [33] and Gene Ontology (GO) [3]. We evaluate multiple pairs consisting of a ComPPI seed subgraph and candidate subgraphs within the network. A representative subset of 100 such pairs is visualized below, with one pair exceeding the statistical significance threshold and yielding the strongest signal. In particular, Figure 7 illustrates the ComPPI seed subgraph together with its best-matched candidate subgraph identified by the proposed method. Figure 8 reports the corresponding structure–weighted mutual information values compared against the statistical significance threshold.

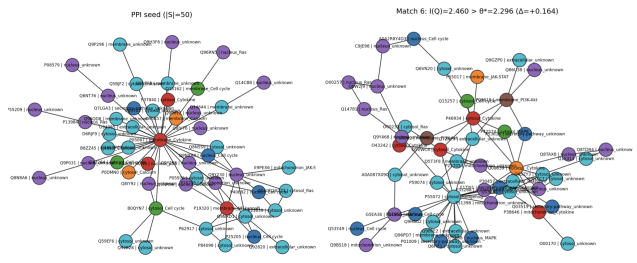


Figure 7: Seed subgraph in the ComPPI interactome (left) and its best-matched candidate subgraph identified (right).

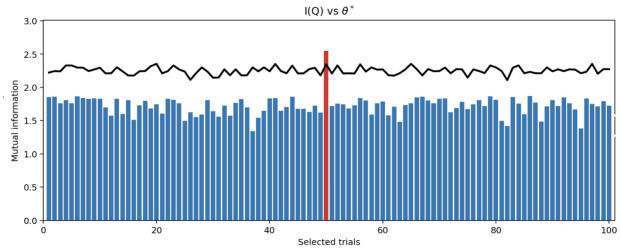
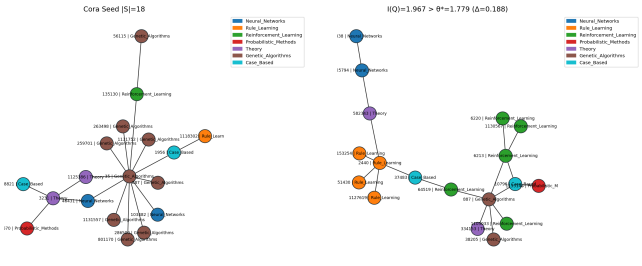


Figure 8: Mutual information values for a representative set of ComPPI subgraph pairs. The spike corresponds to the unique correlated pair. The black line is the threshold θ^* computed according to (14).

These experiments demonstrate that the proposed framework effectively detects structurally correlated subgraphs in PPI networks, capturing joint relationships among protein interactions, subcellular localization, and signaling pathway participation.



(a) Seed subgraph selected from the Cora citation network, chosen to ensure both connectivity and label diversity.

(b) Best correlated candidate subgraph identified in Cora, showing structural and thematic similarity to the seed.

Figure 9: Seed subgraph (left) and its most correlated candidate (right) in the Cora citation network.

4.3 Cora Dataset

Finally, we apply our framework to the Cora citation network [28], a benchmark dataset with 2,708 nodes and 5,429 edges. Each node corresponds to a scientific paper, and edges represent citation links. Every paper is labeled by one of seven research areas: case–based reasoning, genetic algorithms, neural networks, probabilistic methods, reinforcement learning, rule learning, and theory.

Our experimental setup is designed to identify *approximate copies of subgraphs* within the citation network. We begin by extracting a connected *seed* subgraph of prescribed size, chosen so that it contains papers from multiple subject areas. The search then proceeds by sampling connected candidate subgraphs of the same size elsewhere in the network and aligning them to the seed. By comparing candidates to the seed, we can detect recurring modules whose structural and thematic patterns are too consistent to arise by chance. To avoid trivial matches, the seed–construction process explicitly favors neighbors with different subject labels, ensuring that the starting point is both connected and label–diverse.

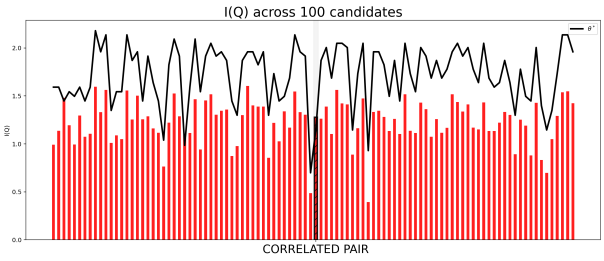


Figure 10: Mutual information across 100 candidate subgraphs in Cora. The spike corresponds to the unique correlated pair, while the remaining candidates serve as null comparisons. The blue line is the threshold θ^* computed according to (14).

Among thousands of sampled candidates, we highlight a subgraph that our test identifies as correlated with the seed. Figures 9a and 9b illustrate the seed together with this highlighted candidate, while Figure 10 shows the corresponding mutual information spike against the background of uncorrelated candidates.

4.4 Scalability Evaluation on ogbn-arxiv

To assess scalability beyond the smaller synthetic and real-data examples above, we also evaluate the method on the ogbn-arxiv citation network [20], which contains over 10^5 nodes. We extract aligned subgraphs of size $|V| = 1000$ and generate noisy aligned copies by randomly perturbing edges and node labels. The agreement level controls the probability with which structure and labels are preserved between the two aligned graphs.

Table 2: Scalability evaluation on ogbn-arxiv. Detection (Det.) denotes the fraction of trials where $I(\hat{Q}) > \theta^*$.

Agreement level	$I(\hat{Q})$	θ^*	Det.
0.30	0.238	0.440	0.00
0.50	0.325	0.424	0.05
0.70	0.466	0.427	0.60
0.80	0.492	0.433	0.55
0.90	0.592	0.433	0.70
0.95	0.846	0.429	0.85

As agreement increases, $I(\hat{Q})$ rises while θ^* remains stable, leading to higher detection rates. This suggests that the method remains stable on larger subgraphs and retains an interpretable decision rule in a heterogeneous real-world network.

5 Conclusion

We presented a framework for testing statistical dependence in labeled graphs using a structure-weighted joint label distribution and mutual information. The method yields computable thresholds that control false positives under structural and label independence. Experiments on synthetic graphs, biological and citation networks, and larger subgraphs show that the statistic detects joint structure–label dependence while remaining interpretable and practical. These results support its use as a calibrated post-hoc validation tool for graph mining with noisy or imperfect alignments.

References

[1] Hasan Metin Aktulga, Ioannis Kontoyiannis, L Alex Lyznik, Lukasz Szpankowski, Ananth Y Grama, and Wojciech Szpankowski. 2007. Identifying statistical dependence in genomic sequences via mutual information estimates. *EURASIP Journal on Bioinformatics and Systems Biology* 2007 (2007), 1–11.

[2] Taha Ameen and Bruce Hajek. 2024. Exact random graph matching with multiple graphs. (2024).

[3] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25–29.

[4] László Babai. 2016. Graph isomorphism in quasipolynomial time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 684–697.

[5] Albert-Laszlo Barabasi and Zoltan N Oltvai. 2004. Network biology: understanding the cell's functional organization. *Nature reviews genetics* 5, 2 (2004), 101–113.

[6] Tibério S Caetano, Julian J McAuley, Li Cheng, Quoc V Le, and Alex J Smola. 2009. Learning graph matching. *IEEE transactions on pattern analysis and machine intelligence* 31, 6 (2009), 1048–1058.

[7] Jin-Yi Cai, Martin Fürer, and Neil Immerman. 1992. An optimal lower bound on the number of variables for graph identification. *Combinatorica* 12, 4 (1992), 389–410.

[8] Jaewon Chung, Bijan Varjavand, Jesús Arroyo-Relión, Anton Alyakin, Joshua Agterberg, Minh Tang, Carey E Priebe, and Joshua T Vogelstein. 2022. Valid two-sample graph testing via optimal transport procrustes and multiscale graph correlation with applications in connectomics. *Stat* 11, 1 (2022), e429.

[9] William W Cohen, Pradeep Ravikumar, Stephen E Fienberg, et al. 2003. A comparison of string distance metrics for name-matching tasks.. In *IIWeb*, Vol. 3. 73–78.

[10] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. 2004. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence* 18, 03 (2004), 265–298.

[11] Daniel Cullina and Negar Kiyavash. 2016. Improved achievability and converse bounds for erdos-rényi graph matching. *ACM SIGMETRICS performance evaluation review* 44, 1 (2016), 63–72.

[12] Daniel Cullina and Negar Kiyavash. 2017. Exact alignment recovery for correlated Erd\H {o} sR\'enyi graphs. *arXiv preprint arXiv:1711.06783* (2017).

[13] Daniel Cullina, Negar Kiyavash, Prateek Mittal, and H Vincent Poor. 2020. Partial recovery of Erdős-Rényi graph alignment via k-core alignment. *ACM SIGMETRICS Performance Evaluation Review* 48, 1 (2020), 99–100.

[14] Jian Ding and Hang Du. 2023. Detection threshold for correlated erdős-rényi graphs via densest subgraph. *IEEE Transactions on Information Theory* 69, 8 (2023), 5289–5298.

[15] Jian Ding and Hang Du. 2023. Matching recovery threshold for correlated random graphs. *The Annals of Statistics* 51, 4 (2023), 1718–1743.

[16] Jian Ding, Hang Du, and Zhangsong Li. 2023. Low-Degree Hardness of Detection for Correlated Erd\H {o} sR\'enyi Graphs. *arXiv preprint arXiv:2311.15931* (2023).

[17] Luca Ganassali, Laurent Massoulié, and Marc Lelarge. 2021. Impossibility of partial recovery in the graph alignment problem. In *Conference on Learning Theory*. PMLR, 2080–2102.

[18] Lesley H Greene and Victoria A Higman. 2003. Uncovering network systems within protein structures. *Journal of molecular biology* 334, 4 (2003), 781–791.

[19] Georgina Hall and Laurent Massoulié. 2023. Partial recovery in the graph alignment problem. *Operations Research* 71, 1 (2023), 259–272.

[20] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems* 33 (2020), 22118–22133.

[21] Maria U Johansson, Vincent Zoete, and Nicolas Guex. 2013. Recurrent structural motifs in non-homologous protein structures. *International journal of molecular sciences* 14, 4 (2013), 7795–7814.

[22] Akira R Kinjo and Haruki Nakamura. 2012. Composite structural motifs of binding sites for delineating biological functions of proteins. *PloS one* 7, 2 (2012), e31437.

[23] Vince Lyzinski, Donniell E Fishkind, and Carey E Priebe. 2014. Seeded graph matching for correlated Erdös-Rényi graphs. *J. Mach. Learn. Res.* 15, 1 (2014), 3513–3540.

[24] Pierre-Francois Marteau. 2018. Sequence covering similarity for symbolic sequence comparison. *arXiv preprint arXiv:1801.07013* (2018).

[25] Colin McDiarmid et al. 1989. On the method of bounded differences. *Surveys in combinatorics* 141, 1 (1989), 148–188.

[26] Brendan D McKay and Adolfo Piperno. 2014. Practical graph isomorphism, II. *Journal of symbolic computation* 60 (2014), 94–112.

[27] Emre Sefer. 2026. Comparison of Biological Graph Alignment Algorithms with Hyperbolic Heterophilic Deep Graph Learning-based Approach. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications* (2026).

[28] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI Magazine* 29, 3 (2008), 93–93.

[29] Cencheng Shen, Jesüs Arroyo, Junhao Xiong, and Joshua T Vogelstein. 2019. Community correlations and testing independence between binary graphs. *arXiv preprint arXiv:1906.03661* (2019).

[30] Cencheng Shen, Carey E Priebe, and Joshua T Vogelstein. 2020. From distance correlation to multiscale graph correlation. *J. Amer. Statist. Assoc.* (2020).

[31] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. 2011. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research* 12, 9 (2011).

[32] Arne Skerra. 2008. Alternative binding proteins: anticalins–harnessing the structural plasticity of the lipocalin ligand pocket to engineer novel binding activities. *The FEBS journal* 275, 11 (2008), 2677–2683.

[33] Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E Peluso, Qingyao Huang, Tao Fang, et al. 2025. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research* 53, D1 (2025), D730–D737.

[34] Daniel V Veres, Dávid M Gyurkó, Benedek Thaler, Kristof Z Szalay, Dávid Fazekas, Tamás Korcsmáros, and Peter Csermely. 2015. ComPPI: a cellular compartment-specific database for protein–protein interaction network analysis. *Nucleic acids research* 43, D1 (2015), D485–D493.

[35] Boris Weisfeiler and Andrei Leman. 1968. The reduction of a graph to canonical form and the algebra which appears therein. *nti, Series* 2, 9 (1968), 12–16.

[36] Yihong Wu, Jiaming Xu, and H Yu Sophie. 2022. Settling the sharp reconstruction thresholds of random graph matching. *IEEE Transactions on Information Theory* 68, 8 (2022), 5391–5417.

[37] Ning Zhang, Ziao Wang, Weina Wang, and Lele Wang. 2024. Attributed graph alignment. *IEEE Transactions on Information Theory* 70, 8 (2024), 5910–5934.

[38] Rui Ray Zhang, Xingwu Liu, Yuyi Wang, and Liwei Wang. 2019. Mcdiarmid-type inequalities for graph-dependent variables and stability bounds. *Advances in Neural Information Processing Systems* 32 (2019).

[39] Si Zhang and Hanghang Tong. 2016. Final: Fast attributed network alignment. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1345–1354.

[40] Si Zhang and Hanghang Tong. 2018. Attributed network alignment: Problem definitions and fast solutions. *IEEE Transactions on Knowledge and Data Engineering* 31, 9 (2018), 1680–1692.

A Proof of Lemma 2.1

Proof. From (1), we have

$$1 - w_u \;=\; \frac{1}{n}\left\|\mathbf{A}_u^G - \mathbf{A}_{\pi(u)}^H\right\|_1.$$

Substituting this into (4) yields

$$\hat{M} \;=\; \frac{1}{2n}\sum_{u\in V_1}(1-w_u). \tag{16}$$

By the definitions of $\hat{P}$ and $\hat{Q}$ for some $(a,b)\in\mathcal{A}\times\mathcal{A}$, we obtain

$$\hat{P}(a,b) - \hat{Q}(a,b) = \frac{1}{n}\sum_{u\in V_1}(1-w_u)\,\mathbf{1}_{\ell_G(u)=a}\,\mathbf{1}_{\ell_H(\pi(u))=b}.$$

Summing over all $(a,b)\in\mathcal{A}\times\mathcal{A}$ gives

$$\sum_{(a,b)\in\mathcal{A}\times\mathcal{A}}\left(\hat{P}(a,b)-\hat{Q}(a,b)\right) = \frac{1}{n}\sum_{u\in V_1}(1-w_u). \tag{17}$$

Comparing with (16) establishes the claim. □

B Proof of Proposition 3.1

Proof. We start by bounding the mismatches for any individual node u as follows:

$$M_{\min} \le \left\|\mathbf{A}_u^G - \mathbf{A}_{\pi(u)}^H\right\|_1 \le M_{\max},$$

which is much more useful in practice. Thus:

$$\frac{M_{\min}}{n} \le 1 - w_u \le \frac{M_{\max}}{n}.$$

Therefore, we can estimate $\hat{P}(a,b)-\hat{Q}(a,b)$ as follows:

$$\hat{P}(a,b) - \hat{Q}(a,b) = \frac{1}{n}\sum_{u\in V_1}(1-w_u)\,\mathbf{1}_{\ell_G(u)=a}\,\mathbf{1}_{\ell_H(\pi(u))=b}$$

leading to

$$\frac{M_{\min}}{n}\hat{P}(a,b) \le \hat{P}(a,b)-\hat{Q}(a,b) \le \frac{M_{\max}}{n}\hat{P}(a,b) \Rightarrow$$

$$\left(1-\frac{M_{\max}}{n}\right)\hat{P}(a,b) \le \hat{Q}(a,b) \le \left(1-\frac{M_{\min}}{n}\right)\hat{P}(a,b).$$

Summing over all $b\in\mathcal{A}$ we obtain:

$$\left(1-\frac{M_{\max}}{n}\right)\hat{P}_G(a) \le \hat{Q}_G(a) \le \left(1-\frac{M_{\min}}{n}\right)\hat{P}_G(a)$$

and we can do the same for $a\in\mathcal{A}$. In conclusion, we find

$$\log\frac{\hat{Q}(a,b)}{\hat{Q}_G(a)\hat{Q}_H(b)} \le \log\left(\frac{\left(1-\frac{M_{\min}}{n}\right)\hat{P}(a,b)}{\left(1-\frac{M_{\max}}{n}\right)^2\hat{P}_G(a)\hat{P}_H(b)}\right) \Rightarrow$$

$$\log\frac{\hat{Q}(a,b)}{\hat{Q}_G(a)\hat{Q}_H(b)} \le \log\frac{\hat{P}(a,b)}{\hat{P}_G(a)\hat{P}_H(b)} + \log\left(\frac{1-\frac{M_{\min}}{n}}{\left(1-\frac{M_{\max}}{n}\right)^2}\right).$$

Now we are in position to estimate $I(\hat{Q})-I(\hat{P})$ as follows:

$$I(\hat{Q})-I(\hat{P}) = \sum_{(a,b)\in\mathcal{A}\times\mathcal{A}}\left(\hat{Q}(a,b)\log\frac{\hat{Q}(a,b)}{\hat{Q}_G(a)\hat{Q}_H(b)} - \hat{P}(a,b)\log\frac{\hat{P}(a,b)}{\hat{P}_G(a)\hat{P}_H(b)}\right) + \hat{Q}(\star,\star)\log\left(\frac{1}{\hat{Q}(\star,\star)}\right).$$

with

$$\hat{Q}(\star,\star)\log\left(\frac{1}{\hat{Q}(\star,\star)}\right) = 2\hat{M}\log\left(\tfrac{1}{2\hat{M}}\right).$$

Summing up, we obtain

$$\sum_{(a,b)\in\mathcal{A}\times\mathcal{A}}\left(\hat{Q}(a,b)\log\frac{\hat{Q}(a,b)}{\hat{Q}_G(a)\hat{Q}_H(b)} - \hat{P}(a,b)\log\frac{\hat{P}(a,b)}{\hat{P}_G(a)\hat{P}_H(b)}\right) \le$$

$$\sum_{(a,b)\in\mathcal{A}\times\mathcal{A}}\left(\left(1-\frac{M_{\min}}{n}\right)\hat{P}(a,b)\left(\log\frac{\hat{P}(a,b)}{\hat{P}_G(a)\hat{P}_H(b)} + \log\left(\frac{1-\frac{M_{\min}}{n}}{\left(1-\frac{M_{\max}}{n}\right)^2}\right)\right) - \hat{P}(a,b)\log\frac{\hat{P}(a,b)}{\hat{P}_G(a)\hat{P}_H(b)}\right) =$$

$$-\frac{M_{\min}}{n}I(\hat{P}) + \left(1-\frac{M_{\min}}{n}\right)\log\left(\frac{1-\frac{M_{\min}}{n}}{\left(1-\frac{M_{\max}}{n}\right)^2}\right)$$

To simplify the logarithmic term, since $M_{\min}/n = O(1/n)$, we may use the approximation

$$\log\left(1-\frac{M_{\min}}{n}\right) = -\frac{M_{\min}}{n} + O\left(\frac{1}{n^2}\right).$$

Substituting the expansion for the $M_{\min}$ term yields

$$I(\hat{Q})-I(\hat{P}) \;\le\; -\frac{M_{\min}}{n}I(\hat{P}) + \left(1-\frac{M_{\min}}{n}\right)$$

$$\left[-\frac{M_{\min}}{n} - 2\log\left(1-\frac{M_{\max}}{n}\right)\right] + 2\hat{M}\log\left(\tfrac{1}{2\hat{M}}\right).$$

□

C More Experimental Results

C.1 Additional Protein Structure Results

Here we present descriptive statistics from the protein pairwise testing experiments. A pair is counted as statistically significant when $I(\hat{Q}) > \theta^*$.

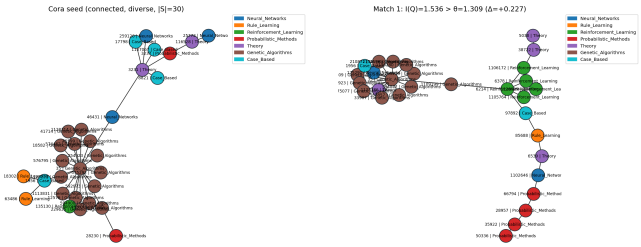
Table 3: Overall summary of protein pairwise testing results

Metric	Value
Total number of pairs	4,828
Pairs with $I(\hat{Q}) > \theta^*$	260
Fraction with $I(\hat{Q}) > \theta^*$	5.4%
Mean $I(\hat{Q})$	1.94
Median $I(\hat{Q})$	1.82
Min $I(\hat{Q})$	1.17
Max $I(\hat{Q})$	4.25

Table 4: Distributional statistics of $I(\hat{Q})$ and θ^* across protein pairs

Range of $I(\hat{Q})$	Count
1.167–1.681	417
1.681–2.194	4,003
2.194–2.707	136
2.707–3.220	12
3.220–3.733	64
3.733–4.247	196

Statistic	Value
Mean θ^*	3.31
Median θ^*	3.36
Min θ^*	2.81
Max θ^*	4.89



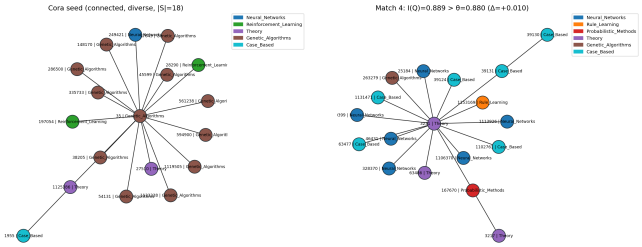
(a) Seed subgraph with 30 nodes. **(b) Best correlated match.**

Figure 12: Representative example with a 30-node seed (left) and its most correlated candidate (right) in the Cora citation network.

C.2 Additional Cora Dataset Results

To assess robustness across different seed sizes, we conducted experiments with multiple seeds in the Cora citation network. The figures below present two representative cases—an 18-node seed and a 30-node seed—illustrating that the proposed test consistently identifies correlated candidates when the seed size varies.

Figure 11 shows the 18-node seed (left) and its most correlated match (right). Despite the relatively small seed size, the method highlights a candidate subgraph with structural and thematic similarity.



(a) Seed subgraph with 18 nodes. **(b) Best correlated match.**

Figure 11: Representative example with an 18-node seed (left) and its most correlated candidate (right) in the Cora citation network.

As a complementary case, Figure 12 presents results for a 30-node seed. Again, the test isolates a nontrivial match, demonstrating that the approach remains effective for larger seeds.