Comparing Graph Transformers via Positional Encodings

Mitchell Black [*1] **Zhengchao Wan** [*2] **Gal Mishne** [2] **Amir Nayyeri** [1] **Yusu Wang** [2]

Abstract

The distinguishing power of graph transformers is tied to the choice of *positional encoding*: features used to augment the base transformer with information about the graph. There are two primary types of positional encoding: *absolute positional encodings (APEs)* and *relative positional encodings (RPEs)*. APEs assign features to each node and are given as input to the transformer. RPEs instead assign a feature to each *pair of nodes*, e.g., shortest-path distance, and are used to augment the attention block. A priori, it is unclear which method is better for maximizing the power of the resulting graph transformer. In this paper, we aim to understand the relationship between these different types of positional encodings. Interestingly, we show that graph transformers using APEs and RPEs are equivalent in their ability to distinguish non-isomorphic graphs. In particular, we demonstrate how to interchange APEs and RPEs while maintaining their distinguishing power in terms of graph transformers. However, in the case of graphs with node features, we show that RPEs may have an advantage over APEs. Based on our theoretical results, we provide a study of different APEs and RPEs—including the shortest-path and resistance distance and the recently introduced stable and expressive positional encoding (SPE)—and compare their distinguishing power in terms of transformers. We believe our work will help navigate the vast number of positional encoding choices and provide guidance on the future design of positional encodings for graph transformers.

[*]Equal contribution [1]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon, USA [2]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, California, USA. Correspondence to: Mitchell Black <blackmit@oregostate.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Graph transformers (GTs) (Dwivedi & Bresson, 2021; Ying et al., 2021) have recently emerged as a competitor to message passing neural networks (MPNNs) (Gilmer et al., 2017) for graph representation learning. While MPNNs only consider the immediate neighbors of a node when updating that node's feature, GTs consider all nodes in the graph. Accordingly, graph transformers can capture global information of graphs that can potentially be used to address the issue of oversquashing faced by MPNNs. Likewise, the attention mechanism of transformers can address the issue of oversmoothing faced by MPNNs. See (Müller et al., 2024).

However, one caveat in applying transformers to graph data is that transformers are ***equivariant***, meaning they treat nodes symmetrically and cannot distinguish two nodes at different positions in the graph. Accordingly, much of the research on graph transformers has been focused on the design of positional encodings that incorporate information about the graph structure into the input or architecture of transformers. There are currently two major types of positional encodings:

Absolute positional encodings (APEs) encode the graph structure as an embedding of the vertices $\phi : V \to \mathbb{R}^d$. APEs are either added to or concatenated with the initial vertex features that are the input to the transformer. Typical examples are the Laplacian eigenvectors (Dwivedi & Bresson, 2021) or information about random walks (Rampášek et al., 2022). Recent works have also proposed learnable APEs (Lim et al., 2022; Huang et al., 2024).

Relative positional encodings (RPEs) encode the graph structure as an embedding of pairs of vertices $\psi : V \times V \to \mathbb{R}^d$. RPEs are incorporated into a GT via a modified attention mechanism. Examples of RPEs include the shortest-path distance (Ying et al., 2021), resistance distance (Zhang et al., 2023), and heat kernels (Choromanski et al., 2022). The adjacency matrix can also serve as an RPE, which will result in an RPE GT with the same distinguishing power as MPNNs (Corollary 4.17).

Although different kinds of APEs and RPEs have been proposed to encode graph structure and enhance the performance of GTs, there is a lack of understanding of how different positional encodings compare—e.g., shortest-path

vs. resistance distance—for distinguishing non-isomorphic graphs. Only a few works (Zhang et al., 2023) have attempted to compare different RPEs, let alone conduct a systematic study comparing APEs and RPEs. This results in a lack of guidance for the design of positional encodings for GTs: which type of positional encoding—absolute or relative—should be used in practice? Furthermore, this could lead to inefficiency in constructing positional encodings. In some recent works on designing APEs like BasisNet (Lim et al., 2022) and the stable and expressive positional encoding (SPE) (Huang et al., 2024), the authors first construct well-defined RPEs and then use some extra machinery to artificially transform RPEs into APEs. This raises the natural question of whether the transition is worthwhile given the expensive extra computation.

Our contributions. We establish a theoretical framework for comparing positional encodings, not only for those in the same category, i.e., absolute or relative, but also across APEs and RPEs. Our main contributions are as follows:

(1) In Section 3, we establish that specific types of APE and RPE GTs (cf. Definition 2.2) are equivalent in their ability to distinguish non-isomorphic graphs. In particular, we show how to map APEs to RPEs and vice versa while maintaining their distinguishing power. We empirically validate our theoretical results on several graph classification datasets.

(2) We show that RPEs may have an advantage over APEs for distinguishing graphs with node features. In particular, we show that transforming an RPE to an APE results in a GT that is able to distinguish strictly fewer graphs with node features than a GT that uses the RPE directly.

(3) The techniques for proving the above results enable us to compare the distinguishing power of different positional encodings. In Section 4, we provide a case study on several popular APEs and RPEs by comparing their distinguishing power. For example, we prove that GTs using SPE (Huang et al., 2024) as an APE are stronger than the GTs using resistance distance as an RPE (Zhang et al., 2023).

(4) In the process of establishing our main results, we identify a new variant of the Weisfeiler-Lehman (WL) test (Weisfeiler & Leman, 1968) that we call the RPE-augWL test that allows us to connect the distinguishing power of RPEs and corresponding GTs. This RPE-augWL test is interesting in itself as a way to further the theoretical formulation of the representation power of graph learning frameworks. We generalize the equivalence between the WL and 2-WL tests to our RPE-augWL test and establish the equivalence between the resistance distance and the pseudoinverse of the Laplacian in terms of distinguishing power via the RPE-augWL test. Furthermore, we identify a class of RPEs resulting in a family of RPE GTs at least as strong as MPNNs.

2. Positional Encodings and Graph Transformers

In this section, we introduce some preliminaries on positional encodings and graph transformers, as well as some auxiliary results that will be used to prove our main results.

Let $G = (V, E)$ denote an unweighted graph, which can be either directed or undirected. A ***featured graph*** (G, X) includes node features $X \in \mathbb{R}^{|V| \times d}$, indicating a d-dimensional feature $X(v)$ for each node in $v \in V$. Let A denote the adjacency matrix, D the diagonal matrix of the degrees, and $L = D - A$ the graph Laplacian matrix. To prevent ambiguity when working with multiple graphs, we use G as a subscript in our notations, such as V_G for the vertex set or X_G for the node features of graph G.

For two graphs G and H, a graph isomorphism is a bijection $\sigma : V_G \to V_H$ such that $\{u, v\} \in E_G$ if and only if $\{\sigma(u), \sigma(v)\} \in E_H$. Two graphs G and H are isomorphic if there is a graph isomorphism $\sigma : V_G \to V_H$.

2.1. Positional Encodings

A positional encoding (PE) of a graph is a way of summarizing structural or positional information of the graph. In their most general form, there are two types of positional encodings we consider in this paper:

Definition 2.1. (Positional Encodings)
An ***absolute positional encoding (APE)*** ϕ assigns each graph G a map $\phi_G : V_G \to \mathbb{R}^l$ such that for any two isomorphic graphs G and H and graph isomorphism $\sigma : V_G \to V_H$, one has that $\phi_G = \phi_H \circ \sigma$.
A ***relative positional encoding (RPE)*** ψ assigns each graph G a map $\psi_G : V_G \times V_G \to \mathbb{R}^k$ such that for any two isomorphic graphs G and H and graph isomorphism $\sigma : V_G \to V_H$, one has that $\psi_G = \psi_H \circ (\sigma \times \sigma)$.

An APE is also naturally expressed as a matrix in $\mathbb{R}^{|V_G| \times l}$ and an RPE as a tensor $\mathbb{R}^{|V_G| \times |V_G| \times k}$.

Note: While node features and APEs both assign a vector to each node, we emphasize that the difference between node features and APEs is that APEs are dependent on the topology of a graph, while two isomorphic graphs can have different node features.

One of the simplest APEs is the degree map $\deg$: for any graph G, $\deg_G(u)$ is the degree of the vertex u. Any graph distance is an RPE, e.g., shortest-path distance (SPD) or resistance distance (RD).

Laplacian eigenvectors are commonly used as APEs (Dwivedi & Bresson, 2021; Kreuzer et al., 2021). However, there is no unique choice of eigenvector for a given eigenspace. Therefore, Laplacian eigenvectors are not a well-defined APE by our definition, as isomorphic graphs

may have different eigenvector encodings due to the choice of basis. However, the projection matrices onto the eigenspaces are a well-defined *RPE* (Lim et al., 2022).

2.2. Graph Transformers

A ***transformer*** is an architecture T composed of ***multi-headed attention layers*** $T = \mathrm{MHA}^{(L)} \circ \cdots \circ \mathrm{MHA}^{(1)}$, where a multiheaded attention layer with heads $1 \leq h \leq H$ is a map $\mathrm{MHA}^{(l)}$ that sends $X^{(l)} \in \mathbb{R}^{n \times d}$ to $\mathrm{MHA}^{(l)}(X^{(l)}) = X^{(l+1)} \in \mathbb{R}^{n \times d}$ as defined below:

$$A^{(l,h)}(X^{(l)}) = \mathrm{softmax}\left(\frac{X^{(l)} W_Q^{(l,h)} (X^{(l)} W_K^{(l,h)})^T}{\sqrt{d_h}}\right)$$

$$Y^{(l)} = X^{(l)} + \left(\sum_{h=1}^{H} A^{(l,h)}(X^{(l)}) X^{(l)} W_V^{(l,h)}\right) W_O^{(l)}$$

$$X^{(l+1)} = Y^{(l)} + \sigma\left(Y^{(l)} W_1^{(l)}\right) W_2^{(l)}$$

where $W_Q^{(l,h)}, W_K^{(l,h)}, W_V^{(l,h)}, W_O^{(l)} \in \mathbb{R}^{d \times d_h}$, $W_1 \in \mathbb{R}^{d_h \times d_r}$, $W_2^{(l)} \in \mathbb{R}^{d_r \times d}$ and σ is an activation function. Here Q, K, V refer to "query", "key" and "value", respectively, in the terminology of the attention literature (Vaswani et al., 2017), and d_h and d_r are the dimension of the hidden layer and the dimension of the residual layer, respectively. The activation function σ is usually the Gaussian error linear units (GELU) (Hendrycks & Gimpel, 2016).

Fact 1 (Transformers are Permutation Equivariant)**.** *Let T be a transformer. Let $X \in \mathbb{R}^{n \times d}$. Let $P \in \mathbb{R}^{n \times n}$ be a permutation matrix. Then $PT(X) = T(PX)$.*

When applying a transformer to a graph G, the input is a set of node features $X_G \in \mathbb{R}^{|V_G| \times d}$; however, initial node features often do not contain any information about the topology of a graph, so two graphs with the same multiset of features but different topologies will have the same output from a transformer. Graph transformers augment the transformer by incorporating information about the input graph in the form of absolute or relative positional encodings.

Definition 2.2 (Graph Transformers)**.** An ***APE graph transformer with APE ϕ (ϕ-APE-GT)*** is a transformer that either: concatenates the node features $X_G \in \mathbb{R}^{|V_G| \times d}$ with the APE $\phi_G \in \mathbb{R}^{|V_G| \times l}$, i.e. $\widehat{X_G} = [X_G | \phi_G] \in \mathbb{R}^{|V_G| \times (d+l)}$, or adds the node features and APE, i.e. $\widehat{X_G} = X_G + \phi_G$, before passing them through a transformer.

An ***RPE graph transformer with RPE ψ (ψ-RPE-GT)*** is a transformer whose attention layer is modified as follows for a graph G with node features $X \in \mathbb{R}^{|V_G| \times d}$:

$$A(X) = f_1(\psi_G) \odot \mathrm{softmax}\left(\frac{X W_Q (X W_K)^T}{\sqrt{d_h}} + f_2(\psi_G)\right)$$

where $f_1, f_2 : \mathbb{R}^k \to \mathbb{R}$ are functions applied entrywise to the tensor $\psi_G \in \mathbb{R}^{|V_G| \times |V_G| \times k}$.

The original transformer (Vaswani et al., 2017) used sines and cosines as APEs for sequences. Transformers for sequences with RPEs were subsequently proposed by Shaw et al. (2018). The first instance of an APE-GT for graphs was proposed by Dwivedi & Bresson (2021) using the Laplacian eigenvectors. The first instance of an RPE-GT was proposed by Ying et al. (2021) using the shortest-path distance, while the form of RPE-GT we consider was proposed by Zhang et al. (2023). Although other ways of incorporating an RPE into a transformer have been proposed (Mialon et al., 2021; Ma et al., 2023), in this paper, APE-GT and RPE-GT only refer to transformers of the forms in Definition 2.2.

2.3. Properties of RPEs

In this subsection, we describe two special types of RPEs that will be useful later.

Diagonally-Aware RPEs. One benefit of using graph distances as RPEs—compared to e.g., the adjacency matrix—is their distinct diagonal entries are all zeros, in contrast to the positive off-diagonal terms. We characterize this property as "diagonal-awareness". As we will see throughout the paper, this property is important as it allows various algorithms to distinguish the feature at a node from all other features.

Definition 2.3. An RPE ψ is ***diagonally-aware*** if for any two graphs G and H and vertices $v \in V_G$ and $x, y \in V_H$, then $\psi_G(v, v) = \psi_H(x, y)$ only if $x = y$.

Diagonal-awareness is a very mild condition as one can always augment an RPE so that it becomes diagonally-aware:

Definition 2.4. The **diagonal augmentation** D^ψ of an RPE ψ is defined, for a graph G, as the stacking $D_G^\psi := (I_G, \psi_G)$ where I_G is the identity matrix.

For any graph G, the diagonal augmentation $D_G^\psi(u, v)$ has a 1 in the first coordinate iff $u = v$ and 0 otherwise. Therefore, D^ψ is diagonally-aware:

Proposition 2.5. *Let ψ be an RPE. The diagonal augmentation D^ψ is diagonally-aware.*

Asymmetric RPEs. Most popular RPEs for undirected graphs are symmetric. However, for directed graphs, it is natural to consider asymmetric RPEs such as the adjacency matrix, the directed distance matrix, or the Laplacian matrix.

We identify the following special type of (possibly asymmetric) RPEs, which will be useful later:

Definition 2.6 (Pseudo-symmetric RPEs)**.** Let ψ be an RPE valued in $\mathbb{R}^k$ for either directed or undirected graphs. Let $f : \mathbb{R}^k \to \mathbb{R}^k$ be any injective function such that $f \circ f = \mathrm{Id}$. Then, ψ is ***(f-)pseudo-symmetric*** if $\psi_G(u, v) = f(\psi_G(v, u))$ for any graph G and any vertices $u, v \in V_G$.

Examples of f satisfying $f \circ f = \mathrm{Id}$ include the identify map, the reflection map, and any coordinate switching map. Obviously, any symmetric RPE ψ is Id-pseudo-symmetric. Any skew-symmetric matrix is pseudo-symmetric for $f(x) = -x$. If we consider complex-valued RPEs, then any Hermitian matrix is pseudo-symmetric for $f(x) = \overline{x}$, the complex conjugate map.

Although the adjacency or the Laplacian matrix of a **directed** graph can also serve as RPEs, they are in general not pseudo-symmetric. In order to guarantee pseudo-symmetry, given any RPE ψ, we can define an augmented RPE $S^{\psi} := (\psi, \psi^T)$, i.e., for any graph G, $S^{\psi}_G := (\psi_G, \psi^T_G)$.

Lemma 2.7 (Pseudo-symmetric augmentation)**.** *Let ψ be any RPE. Define $f : \mathbb{R}^{2k} \to \mathbb{R}^{2k}$ by letting $(x, y) \mapsto (y, x)$ for any $x, y \in \mathbb{R}^k$. Then, S^{ψ} is f-pseudo-symmetric.*

3. Comparison of APE-GT and RPE-GT

In this section, we establish the equivalence of the distinguishing power of APE and RPE GTs for distinguishing graphs without node features. This is done by showing that any RPE can be mapped to an APE without changing its distinguishing power and vice versa. Our strategy is to relate the distinguishing power of APE and RPEs, as well as the maps between the two, to variants of the Weisfeiler-Lehman tests. This approach generalizes the method of Zhang et al. (2023) for studying the expressive power of the Graphormer-GD. Our main results are summarized in Figure 1.

However, for the case of graphs with node features, we prove that our map from RPEs to APEs can result in a strict decrease in the distinguishing power of the corresponding GTs. In this way, our results suggest that (1) the two types of positional encodings are theoretically equivalent for solving graph isomorphism problems but (2) RPE GTs may be better suited for tasks where graph node features are involved.

Our theoretical results are tested empirically in Appendix C.

3.1. Distinguishing Power: PEs vs. Graph Transformers

Positional encodings provide a way of distinguishing different, non-isomorphic graphs. In their most basic form, two graphs can be distinguished by a positional encoding if it takes different values on the two graphs. For graphs with node features, a positional encoding can distinguish two graphs if it takes different values on nodes with different features. Specifically, two featured graphs (G, X_G) and (H, X_H) are ***indistinguishable*** by an APE ϕ if the concatenation of their features with their images by the APE ϕ are the same multiset, i.e., $\{\!\{(X_G(v), \phi_G(v)) : v \in V_G\}\!\} = \{\!\{(X_H(v), \phi_H(x)) : x \in V_H\}\!\}$, and G and H are indistinguishable by an RPE ψ if $\{\!\{(X_G(u), X_G(v), \psi_G(u, v)) : (u, v) \in V_G \times V_G\}\!\} = \{\!\{(X_H(x), X_H(y), \psi_H(x, y)) : (x, y) \in V_H \times V_H\}\!\}$. Likewise, two graphs are ***indistinguishable by a graph transformer*** if their outputs have the same multiset of node features.

Next, we compare the distinguishing power of positional encodings and GTs using these positional encodings.

APE vs. APE-GT. Interestingly, APE-GTs are only as strong as their initial APEs in distinguishing power. This was previously noted by Müller et al. (2024, Section 2.1), although we include a short proof (which follows from the permutation-equivariance of transformers) in Appendix A.4 for completeness.

Lemma 3.1 (Equivalence of APEs and APE-GT)**.** *Any two graphs (G, X_G) and (H, X_H) are indistinguishable by an APE ϕ iff (G, X_G) and (H, X_H) are indistinguishable by all ϕ-APE-GTs.*

RPE vs. RPE-GT. An analogous result does **not** hold for RPEs and RPE-GTs, as using an RPE in an RPE-GT can increase its distinguishing power; see Appendix A.1 for an example. In fact, one can obtain a stronger notion of distinguishing power of RPEs via a variation of the Weisfeiler-Lehman (WL) graph isomorphism test (Weisfeiler & Leman, 1968). This new notion of indistinguishability for RPEs will be equivalent to the indistinguishability of RPE-GTs.

We first recall the standard WL algorithm, which iteratively assigns colors to the vertices of a graph and utilizes the multisets of node colors to distinguish non-isomorphic graphs. The color updating process is defined explicitly as follows:

Definition 3.2. Let G be a graph. The ***Weisfeiler-Lehman (WL) algorithm*** assigns a coloring to the vertices $\chi^{(t)}_G : V_G \to C_t$ for each $t \geq 0$ defined by

$$\chi^{(0)}_G(v) = 1, \qquad \text{(1 denotes an arbitrary constant)}$$
$$\chi^{(t+1)}_G(v) = (\chi^{(t)}_G(v), \{\!\{\chi^{(t)}_G(u) : \{v, u\} \in E_G\}\!\}).$$

However, to understand the distinguishing power of RPE transformers, we need to consider the following augmented WL test for graphs with node features and RPEs.

Definition 3.3. Let ψ be an RPE. Let (G, X_G) be a featured graph. The ***Weisfeiler-Lehman algorithm augmented with RPE ψ (RPE-augWL or ψ-WL)*** assigns a coloring to the vertices $\chi^{(t)}_{\psi_G} : V_G \to C_t$ for each $t \geq 0$ defined by

$$\chi^{(0)}_{\psi_G}(v) = X_G(v),$$
$$\chi^{(t+1)}_{\psi_G}(v) = (\chi^{(t)}_{\psi_G}(v), \{\!\{(\chi^{(t)}_{\psi_G}(u), \psi_G(v, u)) : u \in V_G\}\!\}).$$

Note: For unfeatured graphs, ψ-WL is defined by setting $\chi^{(0)}_{\psi_G}(v) = 1$ (where 1 is an arbitrary constant.) We adopt this convention for all later constructions.

Note that the RPE-augWL algorithm has a different color updating mechanism from the WL algorithm: RPE-augWL

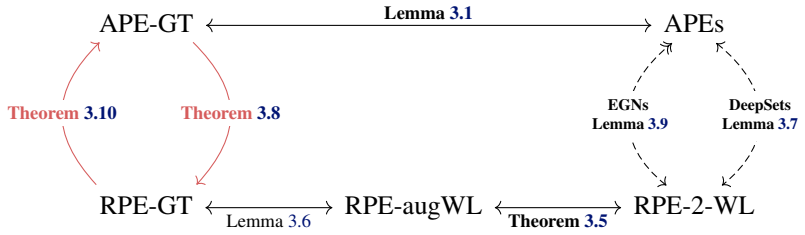


Figure 1. **Illustration of main results.** Arrows denote non-decreasing in distinguishing power. Our main results are the two red arrows on the left. The proofs of the two theorems are illustrated by other parts of the diagram. Our contributions are in **bold**. The two-head arrow from RPE-2-WL to APEs indicates that the non-decreasing property only holds for unfeatured graphs.

considers **all** nodes in the update stage, whereas WL considers only **neighbors**. However, when the RPE is the adjacency matrix A, A-WL is equivalent to WL (cf. Proposition B.16). Hence, we consider RPE-augWL to be a generalization of the WL algorithm.

The definition of RPE-augWL is inspired by the GD-WL test of Zhang et al. (2023) where the new feature is updated as $\widehat{\chi}^{(t+1)}_{\psi_G}(v) = \{\!\{(\widehat{\chi}^{(t)}_{\psi_G}(u),\, \psi_G(v,u)) : u \in V\}\!\}$, where ψ_G is a graph distance. We note that GD-WL is equivalent to our ψ-WL; in GD-WL, the previous color of the node can be deduced from the fact that $\psi_G(v,v) = 0$ since ψ_G is a graph distance.

The WL algorithm has a family of higher-order generalizations called the k-WL algorithms that assign colors to k-tuples of nodes. While Definition 3.3 augments the WL test by using RPEs within the feature updating step, we can also directly use RPEs as the initial colors for 2-WL.

Definition 3.4. Let ψ be an RPE. Let (G, X) be a featured graph. The ***2-Weisfeiler-Lehman algorithm with RPE ψ (RPE-2-WL or ψ-2-WL)*** assigns a color to each pair of vertices $\chi^{(t)}_2 : V \times V \to C_t$ for each $t \geq 0$ defined by[1]

$$\chi^{(0)}_{2,\psi_G}(u,v) = (X_G(u), X_G(v), \psi_G(u,v)),$$
$$\chi^{(t+1)}_{2,\psi_G}(u,v) = \big(\chi^{(t)}_{2,\psi_G}(u,v), \{\!\{\chi^{(t)}_{2,\psi_G}(u,w) : w \in V\}\!\},$$
$$\{\!\{\chi^{(t)}_{2,\psi_G}(w,v) : w \in V\}\!\}\big).$$

Because RPEs are dependent on the graph structure and will be the same (up to permutation) for two isomorphic graphs, we can use the RPE-augWL or RPE-2-WL as heuristics to distinguish non-isomorphic graphs. We say two featured graphs (G, X_G) and (H, X_H) are ***ψ-WL indistinguishable*** if $\{\!\{\chi^{(t)}_{\psi_G}(v) : v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{\psi_H}(v) : v \in V_H\}\!\}$ for all $t \geq 0$. We denote this as $\chi_\psi(G) = \chi_\psi(H)$. Likewise, they are ***$\psi$-2-WL indistinguishable*** if $\{\!\{\chi^{(t)}_{2,\psi_G}(u,v) : u, v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(u,v) : u, v \in V_H\}\!\}$ for all $t \geq 0$. We denote this as $\chi_{2,\psi}(G) = \chi_{2,\psi}(H)$.

ψ-WL and ψ-2-WL indistinguishability can be used as a heuristic test for a special type of graph isomorphism. Two featured graphs (G, X_G) and (H, X_H) are ***feature isomorphic*** if there is a graph isomorphism $\sigma : V_G \to V_H$ such that $X_G(v) = X_H(\sigma(v))$ for all $v \in V_G$. Two feature isomorphic graphs will be ψ-WL and ψ-2-WL indistinguishable.

Fact 2. *If (G, X_G) and (H, X_H) are feature isomorphic, then (G, X_G) and (H, X_H) are ψ-WL indistinguishable and ψ-2-WL indistinguishable for any RPE ψ.*

Interestingly, these two WL tests—RPE-augWL and RPE-2-WL—are equivalent for pseudo-symmetric RPEs. See Appendix A.2 for a proof.

Theorem 3.5 (Equivalence of RPE-augWL and RPE-2-WL.)**.** *Let ψ be a* pseudo-symmetric *RPE. Then two featured graphs (G, X_G) and (H, X_H) are ψ-WL indistinguishable iff they are ψ-2-WL indistinguishable.*

This result generalizes the well-known fact that the WL and 2-WL tests are equivalent; see (Huang & Villar, 2021).

Finally, in contrast to Lemma 3.1, we have the following slight generalization of (Zhang et al., 2023, Theorem 4) who only consider the case where the RPE is a distance function. See Appendix A.6 for a sketch of the proof.

Lemma 3.6 (Equivalence of RPE-augWL and RPE-GT)**.** *Let ψ be a* diagonally-aware *RPE. Let (G, X_G) and (H, X_H) be featured graphs. Then (G, X_G) and (H, X_H) are indistinguisable by the ψ-WL test iff (G, X_G) and (H, X_H) are indistinguisable by all ψ-RPE-GTs.*

3.2. Main Results: APE vs RPE Transformers

In this section, we show that any APE can be turned into an RPE without changing its distinguishing power in terms of transformers and vice versa.

Mapping APEs to RPEs. Let S be any set, and let $\mathrm{Mul}_2(S)$ denote the collection of all 2-element multi-

[1] In the literature (Chen et al., 2020), a pair (u, v) is colored differently from ours at the 0-th step by considering both its isomorphism type and its RPE: $\chi^{(0)}_{2,\psi}(u,v) = (1_{u=v}, A_{uv}, \psi(u,v))$. We ignore the first two terms as one can easily augment ψ with these terms to obtain a new RPE that reflects the isomorphism type if needed; see Section 4.5 for a further discussion.

subsets of S. For a map $f : S \to \mathbb{R}^d$, define the map $h_f : \mathrm{Mul}_2(S) \to \mathbb{R}^d$ as $h_f(\{\!\{x, y\}\!\}) = f(x) + f(y)$. The map h_f is a special case of the DeepSets network (Zaheer et al., 2017) as it operates on multisets. Accordingly, we refer to h_f as a DeepSet network in this paper.

We propose the following way of mapping APEs to RPEs: given a function f and an APE ϕ, we define the RPE ψ^f for a graph G as $\psi^f_G(u, v) := h_f(\{\!\{\phi_G(u), \phi_G(v)\}\!\})$. Our key observation is that given an APE ϕ, there exists a universal function f whose induced RPE ψ^f distinguishes the same pairs of graphs as ϕ (Lemma 3.7). The main Theorem 3.8 will then follow. Proofs of these results can be found in Appendix A.7.

Lemma 3.7. *Let ϕ be an APE. Then there is a function f such that any two featured graphs (G, X_G) and (H, X_H) are indistinguishable by ϕ iff (G, X_G) and (H, X_H) cannot be distinguished by the ψ^f-2-WL test.*

Theorem 3.8. *For any APE ϕ, there exists a function f such that any two graphs (G, X_G) and (H, X_H) are indistinguishable by all ϕ-APE-GTs iff they are indistinguishable by all ψ^f-RPE-GTs.*

Mapping RPEs to APEs. In the other direction, we can transform an RPE ψ to an APE ϕ^g by passing ψ through a 2-equivariant graph network (2-EGN) g. (See (Maron et al., 2019b) or Appendix A.3 for a definition of 2-EGN.) As 2-EGNs are permutation equivariant, the output ϕ^g will be a well-defined APE in the sense of Definition 2.1. This is the approach taken by Lim et al. (2022) to compute an APE from the eigenspaces of the graph Laplacian. It turns out that mapping RPEs to APEs using 2-EGNs can preserve their ability to distinguish non-featured graphs; however, mapping an RPE to an APE may decrease its ability to distinguish featured graphs. Proofs of the following theorems and example can be found in Appendix A.8.

Lemma 3.9. *Let ψ be a* diagonally-aware *RPE.*

For any 2-EGN g, if (G, X_G) and (H, X_H) are indistinguishable by the ψ-2-WL test then (G, X_G) and (H, X_H) are indistinguishable by ϕ^g.

Moreover, for any finite set of unfeatured graphs $\mathcal{G}$, there is a 2-EGN g such that if $G, H \in \mathcal{G}$ are indistinguishable by ϕ^g, then G and H are indistinguishable by the ψ-2-WL test.

Theorem 3.10. *Let ψ be a* diagonally-aware *RPE.*

For any 2-EGN g, if (G, X_G) and (H, X_H) are indistinguishable by all ψ-RPE-GTs, then (G, X_G) and (H, X_H) are indistinguishable by ϕ^g.

Moreover, for any finite set of unfeatured *graphs $\mathcal{G}$, there is a 2-EGN g such that if $G, H \in \mathcal{G}$ are indistinguishable by all ϕ^g-APE-GTs, then G and H are indistinguishable by all ψ-RPE-GTs.*

Example 1. There exists an RPE ψ and featured graphs (G, X_G) and (H, X_H) that are distinguishable by ψ-2-WL but are indistinguishable by ϕ^g for any 2-EGN g.

Remark 3.11. Even in the case of unfeatured graphs, there is still a slight asymmetry in the statements of the results for APEs (Theorem 3.8) and RPEs (Theorem 3.10), as the result for APEs does not have the restriction on "a finite set" of graphs. This restriction arises because there is no known universal 2-EGN equal in power to the RPE-2-WL test. This is in part due to the fact that known 2-EGNs would need an unbounded hidden dimension and number of layers to distinguish all pairs of graphs (Maron et al., 2019a).

More expressive RPE to APE maps. Theorem 3.10 shows that RPEs can be transformed into equally-powerful (for unfeatured graphs) APEs using 2-EGNs. However, it is worth noting that RPEs can be turned into APEs *that are even more powerful* than the original RPE using k-EGNs for $k > 2$ (Maron et al., 2019a). In the extreme case of $k \in \Omega(n)$, k-EGNs can distinguish all pairs of non-isomorphic graphs with n vertices. Such techniques for constructing APEs from RPEs have been previously explored in the literature (Huang et al., 2024); for example, Expressive BasisNet (Lim et al., 2022). However, the downside to these techniques is that their computational cost increases with their expressive power; k-EGNs take $\Theta(n^k)$ time as they operate on tensors of size $\Theta(n^k)$. In practice, techniques that map RPEs to APEs typically only use 2-EGNs because of the computational cost of higher k-EGNs.

Restrictions and Implications. Our main findings (Theorem 3.8 and Theorem 3.10) show that APE-GTs and RPE-GTs have comparable distinguishing capabilities for unfeatured graphs. However, in practice, learning the DeepSet (to serve as f) in Theorem 3.8 or the 2-EGN (to serve as g) in Theorem 3.10 may not always yield the desired maps. Hence, it remains an interesting open question to figure out whether it is theoretically easier to learn the DeepSet or the 2-EGN. For example, one would benefit in designing architectures from understanding which method requires a lower dimension for hidden layers.

Finally, we would like to point out that converting RPEs to APEs to fit a specific graph transformer architecture is not recommended, as shown by Example 1. See also discussion in Remark 4.5 and Appendix C.3 for some empirical validation of this point.

4. Comparing Graph Transformers with Different Positional Encodings

In this section, we compare the distinguishing power of graph transformers with different (absolute or relative) positional encodings. The main results are summarized in

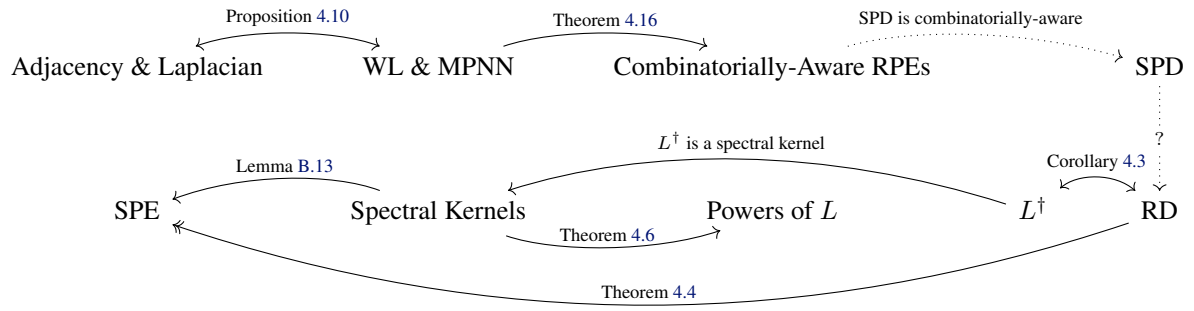


Figure 2. **Hierarchy of PEs.** The arrows indicate that the corresponding positional encoding is less strong than the one it points to in terms of distinguishing power. The two-head arrow indicates that the non-decreasing property only holds for unfeatured graphs. The dotted arrow between SPD and RD refers to some partial evidence (cf. Theorem B.18) that RD is stronger than SPD in some respects; however, it is an open question how the two compare as RPEs.

Figure 2. Intuitively, by Theorem 3.8, APE-GTs can be turned into RPE-GTs preserving the distinguishing power. This enables us to compare APE-GTs with RPE-GTs. When comparing two RPE-GTs, by Lemma 3.6 we simply need to compare the corresponding RPEs via the RPE-augWL test. For two RPEs ψ_1 and ψ_2, we say that ψ_1-WL is ***at least as strong as*** ψ_2-WL if for any two featured graphs (G, X_G) and (H, X_H) such that $\chi_{\psi_1}(G) = \chi_{\psi_1}(H)$, then we have $\chi_{\psi_2}(G) = \chi_{\psi_2}(H)$. Moreover, we say that ψ_1-WL and ψ_2-WL are ***equally strong*** if for any two featured graphs (G, X_G) and (H, X_H), we have that $\chi_{\psi_1}(G) = \chi_{\psi_1}(H)$ iff $\chi_{\psi_2}(G) = \chi_{\psi_2}(H)$.

Some highlights of this section include a comparison of SPE-APE-GTs with RD-RPE-GTs (Theorem 4.4), identification of a class of RPEs whose corresponding RPE-GTs are at least as strong as MPNNs (Proposition 4.10 and Corollary 4.17), and suggesting new RPEs for directed graphs (Proposition 4.12).

4.1. Resistance Distance, Spectral Kernels, and SPE

Many popular APEs (such as BasisNet (Lim et al., 2022) and SPE (Huang et al., 2024)) and RPEs (such as RD (Zhang et al., 2023)) are defined using the spectrum of the Laplacian L. While BasisNet and SPE have been shown to be equivalent (Huang et al., 2024, Proposition 3.2), it is unclear how either of these compares to RD in terms of transformers. More generally, it is unknown how much information RD-WL carries about the spectrum of L.

In this section, we establish that SPE-APE-GTs are stronger than resistance distance RPE-GTs (Theorem 4.4). As one step to prove this, we show that two graphs are indistinguishable by RD-WL iff they are indistinguishable by $L^\dagger$-WL where $L^\dagger$ is the pseudoinverse of L (Corollary 4.3). This result gives a partial answer to the question of how much information RD-WL carries about the Laplacian spectrum. This is weaker than saying that RD-WL indistinguishable graphs have the same $L^\dagger$ up to permutation, but it does show that $L^\dagger$ of these graphs are indistinguishable when combined with RPE-augWL (and thus RPE-GTs).

In fact, our result about RD-WL is a corollary to a stronger result about the equivalence of RPE-augWL tests with spectral distance and spectral kernel as RPEs.

Definition 4.1. Let $L = \sum_{i=2}^{n} \lambda_i z_i z_i^T$ be the spectral decomposition, where $\{\lambda_2, \dots, \lambda_n\}$ are the non-zero eigenvalues and $\{z_2, \dots, z_n\}$ is an orthonormal basis of eigenvectors. Let $f : \mathbb{R}^+ \to \mathbb{R}^+$ be a function. The ***spectral kernel*** (Hammond et al., 2011) corresponding to f is the matrix $K_G^f = \sum_{i=2}^{n} f(\lambda_i) z_i z_i^T$. The ***spectral distance*** corresponding to f is defined

$$d_G^f(u, v) := \sqrt{K_G^f(u, u) + K_G^f(v, v) - 2K_G^f(u, v)}.$$

Example 2. The ***diffusion distance*** (at time t) (Coifman & Lafon, 2006) is the spectral distance D^f for the function $f(x) = e^{-tx}$. The ***heat kernel*** is the corresponding spectral kernel denoted by $H^{(t)} = \sum_{i=2}^{n} e^{-\lambda_i t} z_i z_i^T$, which has been proposed as an RPE by Choromanski et al. (2022).

Note: Although eigenvectors are not unique due to the choice of bases for eigenspaces, the spectral kernel and spectral distance are unique up to graph isomorphism and so are well-defined RPEs.

Let D^f denote the RPE assigning the matrix of spectral distance d_G^f to a graph G and let K^f denote the RPE assigning the matrix of spectral kernel K_G^f.

Theorem 4.2. *Let $f : \mathbb{R}^+ \to \mathbb{R}^+$. D^f-WL is at least as strong as K^f-WL. K^f-WL with diagonal augmentation is at least as strong as D^f-WL.*

Theorem 4.2 is the result of the fact that the spectral distance and kernel are the distance and Gram matrix, respectively, of a point cloud in $\mathbb{R}^d$. A variant of the WL-algorithm for general distance matrices of point clouds was studied

by delle Rose et al. (2023) as a heuristic for determining if two point clouds are isometric, so our proof may have implications for these algorithms as well. A proof of this theorem can be found in Appendix B.2.

The ***resistance distance (RD)*** is the *squared* spectral distance of the function $f(x) = x^{-1}$, and the ***pseudoinverse*** of the Laplacian $L^\dagger$ is the corresponding spectral kernel, so Theorem 4.2 applies to RD and $L^\dagger$. Interestingly, RD has additional properties that allow us to drop the assumption of diagonal augmentation. A proof of the following theorem can be found in Appendix B.2.

Corollary 4.3. *RD-WL and $L^\dagger$-WL are equally strong.*

Finally, as an immediate consequence of Corollary 4.3 (combined with Theorem 3.10), we compare RD-RPE-GT, with existing APE-GPTs using the so-called *Stable and Expressive Positional Encodings (SPE)* (Huang et al., 2024). A proof of this theorem can be found in Appendix B.3.

Theorem 4.4. *For unfeatured graphs, diagonally augmented SPE-APE-GT is at least as strong as RD-RPE-GT.*

Remark 4.5. In Appendices C.1 and C.2, we empirically validate this theorem by showing that SPE-APE-GTs perform similarly to RD-RPE-GTs on some unfeatured graph classification/isomorphism tasks. However, when node features are involved, the theorem may not hold since the 2-EGN involved in the construction of SPE-APE-GT *may* have weaker distinguishing power than the RD-RPE-GT (cf. Example 1). In Appendix C.3, we empirically observe that for graph regression tasks when node features are involved, using RD-RPE-GT can indeed give us an edge in performance over SPE-APE-GT.

4.2. Resistance Distance and Cut Edges.

The first proposed RPE for graph transformers was the ***shortest-path distance (SPD)*** (Ying et al., 2021). Later, Zhang et al. (2023) proposed to use the stack of RD and SPD as an RPE. It is natural to wonder how RD-WL compares with SPD-WL. While this remains an open question, we prove that RD-WL is at least as strong as SPD-WL in at least one regard.

Zhang et al. (2023) showed that SPD-WL could distinguish cut edges but not cut vertices in unfeatured graphs. Furthermore, the authors showed that RD-WL could distinguish cut vertices. However, it was not known whether or not RD-WL could distinguish cut edges in a graph. We answer this open question in the affirmative: RD-WL can distinguish cut edges in a graph. See Theorem B.18 for a precise theorem statement and proof.

However, our experiments on the BREC dataset (Wang & Zhang, 2024) suggest (but do not prove) that RD-WL may not be strictly stronger than SPD-WL, as there are pairs of graphs in this dataset that an SPD-RPE-GT was able to learn to distinguish, while we were unsuccessful in training a RD-RPE-GT to distinguish these graphs. See Appendix C.2 for details.

4.3. Powers of Matrices and Spectral Kernels

Another common type of RPE is the stack of powers of some matrix associated with the graph; for example, GRIT (Ma et al., 2023) uses a stack of powers of the random-walk adjacency matrix of a graph called the ***relative random-walk positional encoding (RRWP)***. In this section, we show that using stacks of various matrices as an RPE is at least as strong as any spectral kernel RPEs for any $f : \mathbb{R}^+ \to \mathbb{R}^+$. Proofs for this section can be found in Appendix B.4.

We first prove that using sufficiently many powers of the Laplacian is at least as strong as any spectral kernel.

Theorem 4.6. *$(I, L, \ldots, L^{2n-1})$-WL is at least as strong as K^f-WL on graphs with at most n nodes.*

Remark 4.7. While it may seem excessive to consider $2n$ powers of L to match the power of a single spectral kernel, this is not computationally more expensive than computing the spectral kernel in the first place, which requires adding n matrices $f(\lambda_i)z_iz_i^T$ of size n^2, which takes $O(n^3)$ time.

A variant of RRWP that uses stacks powers of the symmetrically-normalized adjacency matrix $\widehat{A} = D^{-1/2}AD^{-1/2}$ is at least as strong as spectral kernels using the symmetrically-normalized Laplacian $\widehat{K}^f$ (see Appendix B.4 for definition).

Theorem 4.8. *$(I, \widehat{A}, \ldots, \widehat{A}^{2n-1})$-WL is at least as strong as $\widehat{K}^f$-WL on graphs with at most n nodes.*

Finally, we show that stacked heat kernels $H^{(t)}$ (cf. Example 2) are also at least as strong as any spectral kernel.

Theorem 4.9. *$(I, H^{(1)}, \ldots, H^{(2n-1)})$-WL is at least as strong as K^f-WL on graphs with at most n nodes.*

4.4. RPEs: Common Matrices for Graphs

There are multiple common matrices used to characterize graphs, such as the adjacency matrix, Laplacian matrix and their normalized versions. We will examine their corresponding RPEs and compare their distinguishing power.

Undirected graphs. For undirected graphs, we show that various common matrices lead to equivalent RPEs.

Let $\widetilde{A} = D^{-1}A$. Let $\widehat{L} = I - \widehat{A}$ and $\widetilde{L} = I - \widetilde{A}$ denote the symmetrically-normalized and the random-walk graph Laplacians, respectively. These, along with A, L and $\widehat{A}$, give rise to six RPEs and hence five corresponding RPE-augWL tests. The following result suggests that in practice, perhaps there is no need to use more than one such type of information with RPE-GTs and one can interchangeably use any of the matrix as RPE.

Proposition 4.10. *$A, \widehat{A}, \widetilde{A}, L, \widehat{L}$ and $\widetilde{L}$ induce equivalent RPE-augWL tests. In particular, all of these RPE-augWL tests are equally strong as the WL test.*

Directed graphs. Directed graphs are sometimes turned into a magnetic graph to obtain a Hermitian matrix called the ***magnetic Laplacian*** (Fanuel et al., 2018). Let G be a directed graph with adjacency matrix A. Let $A^* := (A + A^T)/2$ and let D^* denote the corresponding degree matrix. Then, given a parameter α, the magnetic Laplacian is defined as

$$L^\alpha := D^* - T^\alpha \odot A^*,$$

where $T^\alpha_{ij} = \exp(\iota \cdot 2\pi\alpha \cdot \mathrm{sgn}(A_{ji} - A_{ij}))$ and ι is the imaginary unit. As mentioned in Section 2, L^α can serve as a pseudo-symmetric RPE. In fact, L^α can be obtained via the pseudo-symmetric augmentation (A, A^T) of A (cf. Lemma 2.7) as well as an extra augmentation via D^*:

Lemma 4.11. *$L^\alpha = f^\alpha((D^*, A, A^T))$ where $f^\alpha : \mathbb{R}^3 \to \mathbb{R}$ is applied to each channel $n \times n \times \cdot$ such that*

$$f^\alpha(x, y, z) := x - \exp(\iota 2\pi\alpha \cdot \mathrm{sgn}(z - y)) \cdot \frac{y + z}{2}.$$

In this way, we obtain an RPE (D^*, A, A^T) more general than the magnetic Laplacian matrix. Hence, instead of using magnetic Laplacian as positional encoding, one should consider using (D^*, A, A^T)-RPE-GT.

Proposition 4.12. *The (D^*, A, A^T)-WL test is at least as strong as the L^α-WL test for any α.*

4.5. RPEs: Combinatorial-Awareness and WL

It is an interesting question what information about a graph a transformer can extract from a positional encoding. In Definition 2.2, the learnable functions in an RPE-GT f_i are applied elementwise to the RPEs. Moreover, the functions are also applied elementwise to the attention scores. In other words, given an RPE ψ, even though certain graph structure can be obtained indirectly through more complicated operations on ψ, such hidden information may not be accessible to the transformer because of the limited set of operations it can perform on the RPE. For example, while a graph can be reconstructed from its resistance distance matrix, an RPE-GT may not have access to the edges of a graph as it cannot perform the necessary operations to reconstruct the graph.

One basic piece of information a GT may want to access through an RPE is—of course—the adjacency matrix. We identify a desirable type of RPE where a GT will have this information.

Definition 4.13. An RPE ψ is ***combinatorially-aware*** if for any two graphs G and H and vertices $u, v \in V_G$ and $x, y \in V_H$, then $\psi_G(u, v) = \psi_H(x, y)$ iff one of the following conditions hold: (i) $\{u, v\} \in E_G$ and $\{x, y\} \in E_H$ (ii) $\{u, v\} \notin E_G$ and $\{x, y\} \notin E_H$.

As it turns out, elementwise operations are insufficient for reconstructing the graph structure from the resistance distance, as evidenced by the following example.

Example 3. The shortest-path distance (SPD) is combinatorially-aware as $\mathrm{SPD}(u, v) = 1$ iff $\{u, v\} \in E$. The resistance distance is, however, not combinatorially-aware. See Appendix B.7 for a counterexample.

As in the case of diagonal-awareness, given any RPE, one can always augment it to be combinatorially-aware.

Definition 4.14 (Combinatorial Augmentation)**.** Let ψ be any RPE. We define its **combinatorial augmentation** C^ψ as follows: for any graph G, $C^\psi_G(u, v) := (A_G(u, v), \psi_G(u, v))$ for any $u, v \in V$, where A_G denotes the adjacency matrix of G.

This in fact provides an equivalent characterization of combinatorially-aware RPEs.

Proposition 4.15. *An RPE ψ is combinatorially-aware if for any G, the following condition holds for all $u, v, x, y \in V_G$: $\psi_G(u, v) = \psi_G(x, y)$ iff $C^\psi_G(u, v) = C^\psi_G(x, y)$.*

The reason combinatorially-aware RPEs are important is because RPE-augWL with combinatorially-aware RPEs are at least as strong as the WL test (Weisfeiler & Leman, 1968). A proof of this theorem can be found in Appendix B.6.

Theorem 4.16. *Let ψ be a combinatorially-aware RPE. Then, ψ-WL is at least as strong as WL.*

Theorem 4.16 is of interest because the WL test is equally strong at distinguishing graphs as message passing graph neural networks (MPNNs) (Xu et al., 2018).

Corollary 4.17. *Let ψ be a combinatorially-aware RPE. Then ψ-RPE-GTs are at least as strong as MPNNs.*

5. Conclusion and Open Questions

We have established a framework for comparing different types of positional encodings. Future work can continue this line of work of comparing different specific positional encodings, e.g. shortest-path vs. resistance distance or resistance distance vs. SPE. Furthermore, our theoretical results provide some suggestions for designing new positional encodings, e.g., one may not want to convert RPEs into APEs as done in the literature. Finally, it would be interesting to shift the focus of research on positional encodings for graph transformers from coarse-grained distinguishing-power results (what pairs of graphs can GTs distinguish) to more fine-grained results on expressive power, for example, approximation results (Azizian & Lelarge, 2021) or results using RPE-augWL-inspired distances to study graph transformers (Chen et al., 2022; 2023; Böker et al., 2024).

Acknowledgements

This work was supported by the NSF (Grants CCF-2217058 for GM, ZW, and YW; CCF-2112665 and CCF-2310411 for YW; and CCF-2311180 and CCF-1941086 for MB and AN). MB would like to thank YW for supporting a visit to UCSD where this research was initiated.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of potentially more effective graph learning models due to our work, none of which we feel must be specifically highlighted here.

References

Azizian, W. and Lelarge, M. Expressive power of invariant and equivariant graph neural networks. In *ICLR 2021-International Conference on Learning Representations*, 2021.

Böker, J., Levie, R., Huang, N., Villar, S., and Morris, C. Fine-grained expressivity of graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Chen, S., Lim, S., Mémoli, F., Wan, Z., and Wang, Y. Weisfeiler-Lehman meets Gromov-Wasserstein. In *International Conference on Machine Learning*, pp. 3371–3416. PMLR, 2022.

Chen, S., Lim, S., Mémoli, F., Wan, Z., and Wang, Y. The Weisfeiler-Lehman distance: Reinterpretation and connection with GNNs. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 404–425. PMLR, 2023.

Chen, Z., Chen, L., Villar, S., and Bruna, J. Can graph neural networks count substructures? *Advances in neural information processing systems*, 33:10383–10395, 2020.

Choromanski, K., Lin, H., Chen, H., Zhang, T., Sehanobish, A., Likhosherstov, V., Parker-Holder, J., Sarlos, T., Weller, A., and Weingarten, T. From block-Toeplitz matrices to differential equations on graphs: towards a general theory for scalable masked transformers. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3962–3983. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/choromanski22a.html.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

delle Rose, V., Kozachinskiy, A., Rojas, C., Petrache, M., and Barcelo, P. Three iterations of (d − 1)-WL test distinguish non isometric clouds of d-dimensional points. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=9yhYcjsdab.

Dwivedi, V. P. and Bresson, X. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

Dwivedi, V. P., Joshi, C. K., Luu, A. T., Laurent, T., Bengio, Y., and Bresson, X. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24 (43):1–48, 2023.

Fanuel, M., Alaíz, C. M., Fernández, Á., and Suykens, J. A. Magnetic eigenmaps for the visualization of directed networks. *Applied and Computational Harmonic Analysis*, 44(1):189–199, 2018.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

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

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Huang, N. T. and Villar, S. A short tutorial on the Weisfeiler-Lehman test and its variants. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8533–8537. IEEE, 2021.

Huang, Y., Lu, W., Robinson, J., Yang, Y., Zhang, M., Jegelka, S., and Li, P. On the stability of expressive positional encodings for graph neural networks. In *The Twelfth International Conference on Learning Representations*, 2024.

Jin, B., Zhang, Y., Meng, Y., and Han, J. Edgeformers: Graph-empowered transformers for representation learning on textual-edge networks. *arXiv preprint arXiv:2302.11050*, 2023.

Kreuzer, D., Beaini, D., Hamilton, W., Létourneau, V., and Tossou, P. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34:21618–21629, 2021.

Lim, D., Robinson, J. D., Zhao, L., Smidt, T., Sra, S., Maron, H., and Jegelka, S. Sign and basis invariant networks for spectral graph representation learning. In *The Eleventh*

International Conference on Learning Representations, 2022.

Ma, L., Lin, C., Lim, D., Romero-Soriano, A., Dokania, P. K., Coates, M., Torr, P., and Lim, S.-N. Graph inductive biases in transformers without message passing. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23321–23337. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ma23c.html.

Maron, H., Ben-Hamu, H., Serviansky, H., and Lipman, Y. Provably powerful graph networks. *Advances in neural information processing systems*, 32, 2019a.

Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019b.

Mialon, G., Chen, D., Selosse, M., and Mairal, J. Graphit: Encoding graph structure in transformers, 2021.

Müller, L., Galkin, M., Morris, C., and Rampášek, L. Attending to graph transformers. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=HhbqHBBrfZ.

Rampášek, L., Galkin, M., Dwivedi, V. P., Luu, A. T., Wolf, G., and Beaini, D. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Shaw, P., Uszkoreit, J., and Vaswani, A. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, 2018.

Spielman, D. Spectral and algebraic graph theory. Available at http://cs-www.cs.yale.edu/homes/spielman/sagt/sagt.pdf (2021/12/01), 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, Y. and Zhang, M. An empirical study of realized GNN expressiveness. In *International Conference on Machine Learning*. PMLR, 2024.

Weisfeiler, B. and Leman, A. The reduction of a graph to canonical form and the algebra which appears therein. *nti, Series*, 2(9):12–16, 1968.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34:28877–28888, 2021.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Zhang, B., Luo, S., Wang, L., and He, D. Rethinking the expressive power of GNNs via graph biconnectivity. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhu, W., Wen, T., Song, G., Wang, L., and Zheng, B. On structural expressive power of graph transformers. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 3628–3637, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599451. URL https://doi.org/10.1145/3580305.3599451.

A. Details from Section 3

A.1. Example of RPE vs RPE-augWL

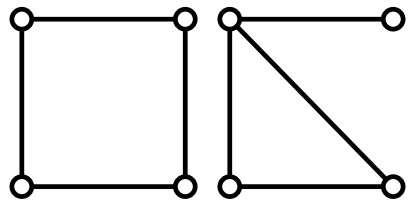


Figure 3. Left: G. Right: H

In this section, we show that RPE-augWL can distinguish graphs that are indistinguishable by just their RPE. By Lemma 3.6, this also means the graphs are distinguishable by some RPE transformer. Consider the graphs G and H in Figure 3 and the RPE A that assigns each graph its adjacency matrix. Both graphs have 4 vertices and 4 edges, so their adjacency matrices A_G and A_H are indistinguishable. However, after one iteration of A-WL, all vertices of G will have the same color, while the degree 1 vertex in H will have a different color from all vertices in G. Therefore, G and H are distinguishable by A-WL.

A.2. Equivalence of RPE-augWL and RPE-2-WL: Proof of Theorem 3.5

Our proof of Theorem 3.5 is based on the following lemma.

Lemma A.1. *Let ψ be a pseudo-symmetric RPE. Let (G, X_G) and (H, X_H) be featured graphs. Let $u, v \in V_G$ and $x, y \in V_H$. Then $\chi^{(t)}_{2,\psi_G}(u, v) = \chi^{(t)}_{2,\psi_H}(x, y)$ iff $\psi_G(u, v) = \psi_H(x, y)$, $\chi^{(t)}_{\psi_G}(u) = \chi^{(t)}_{\psi_H}(x)$, and $\chi^{(t)}_{\psi_G}(v) = \chi^{(t)}_{\psi_H}(y)$.*

Proof of Lemma A.1. We prove this by induction on t. For $t = 0$, this is follows directly from definition as $\chi^{(t)}_{\psi_G}(v) = X_G(v)$ and $\chi^{(0)}_{2,\psi_G}(u, v) = (X_G(u), X_G(v), \psi_G(u, v))$.

Now inductively suppose this is true for some value t; we will show this is true for $t + 1$.

First assume that $\chi^{(t+1)}_{2,\psi_G}(u, v) = \chi^{(t+1)}_{2,\psi_H}(x, y)$. Then, it follows from the definition that

1. $\chi^{(t)}_{2,\psi_G}(u, v) = \chi^{(t)}_{2,\psi_H}(x, y)$.
2. $\{\!\{\chi^{(t)}_{2,\psi_G}(u, w) : w \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(x, z) : z \in V_H\}\!\}$;
3. $\{\!\{\chi^{(t)}_{2,\psi_G}(w, v) : w \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(z, y) : z \in V_H\}\!\}$.

By induction, Item 1 implies that

4. $\psi_G(u, v) = \psi_H(x, y)$, and by pseudo-symmetry, $\psi_G(v, u) = \psi_H(y, x)$;
5. $\chi^{(t)}_{\psi_G}(u) = \chi^{(t)}_{\psi_H}(x)$ and $\chi^{(t)}_{\psi_G}(v) = \chi^{(t)}_{\psi_H}(y)$;

By item 2, there is a bijection $\sigma : V_G \to V_H$ such that $\chi^{(t)}_{2,\psi_G}(u, w) = \chi^{(t)}_{2,\psi_H}(x, \sigma(w))$. The inductive hypothesis implies that $\psi_G(u, w) = \psi_H(x, \sigma(w))$ and $\chi^{(t)}_{\psi_G}(w) = \chi^{(t)}_{\psi_H}(\sigma(w))$. Therefore, the ψ-WL colors of u and x at $t + 1$ are the same as

$$\begin{aligned}\chi^{(t+1)}_{\psi_G}(u) &= (\chi^{(t)}_{\psi_G}(u), \{\!\{ (\chi^{(t)}_{\psi_G}(w), \psi_G(u, w)) : w \in V_G \}\!\}) \\ &= (\chi^{(t)}_{\psi_H}(x), \{\!\{ (\chi^{(t)}_{\psi_H}(z), \psi_H(x, z)) : z \in V_H \}\!\}) = \chi^{(t+1)}_{\psi_H}(x).\end{aligned}$$

By item 3, there is a bijection $\sigma' : V_G \to V_H$ such that $\chi^{(t)}_{2,\psi_G}(w, v) = \chi^{(t)}_{2,\psi_H}(\sigma'(w), y)$. The inductive hypothesis implies that $\psi_G(w, v) = \psi_H(\sigma'(w), y)$ and $\chi^{(t)}_{\psi_G}(w) = \chi^{(t)}_{\psi_H}(\sigma'(w))$. By pseudo-symmetry, one also has that $\psi_G(v, w) =$

$\psi_H(y, \sigma'(w))$. Therefore, the ψ-WL colors of v and y at $t+1$ are the same as

$$\begin{aligned}\chi^{(t+1)}_{\psi_G}(v) &= (\chi^{(t)}_{\psi_G}(v), \{\!\{ (\chi^{(t)}_{\psi_G}(w), \psi_G(v,w)) : w \in V_G \}\!\}) \\ &= (\chi^{(t)}_{\psi_H}(y), \{\!\{ (\chi^{(t)}_{\psi_H}(y), \psi_H(y,z)) : z \in V_H \}\!\}) = \chi^{(t+1)}_{\psi_H}(y).\end{aligned}$$

Conversely, assume that $\psi_G(u,v) = \psi_H(x,y)$, $\chi^{(t+1)}_{\psi_G}(u) = \chi^{(t+1)}_{\psi_H}(x)$, and $\chi^{(t+1)}_{\psi_G}(v) = \chi^{(t+1)}_{\psi_H}(y)$. By definition of the ψ-WL, we have that $\chi^{(t)}_{\psi_G}(u) = \chi^{(t)}_{\psi_H}(x)$ and $\chi^{(t)}_{\psi_G}(v) = \chi^{(t)}_{\psi_H}(y)$. Hence, by induction, we have that $\chi^{(t)}_{2,\psi_G}(u,v) = \chi^{(t)}_{2,\psi_H}(x,y)$. Moreover, there exist bijections $\sigma_{ux} : V_G \to V_H$ and $\sigma_{vy} : V_G \to V_H$ such that

1. for any $w \in V_G$, $\chi^{(t)}_{\psi_G}(w) = \chi^{(t)}_{\psi_H}(\sigma_{ux}(w))$ and $\psi_G(u,w) = \psi_H(x, \sigma_{ux}(w))$;
2. for any $w \in V_G$, $\chi^{(t)}_{\psi_G}(w) = \chi^{(t)}_{\psi_H}(\sigma_{vy}(w))$ and $\psi_G(v,w) = \psi_H(y, \sigma_{vy}(w))$, and by pseudo-symmetry, we also have that $\psi_G(w,v) = \psi_H(\sigma_{vy}(w), y)$.

For any given $w \in V_G$, by induction, one has that $\chi^{(t)}_{2,\psi_G}(u,w) = \chi^{(t)}_{2,\psi_H}(x, \sigma_{ux}(w))$ and $\chi^{(t)}_{2,\psi_G}(w,v) = \chi^{(t)}_{2,\psi_H}(\sigma_{vy}(x), y)$. Therefore, $\{\!\{\chi^{(t)}_{2,\psi_G}(u,w) : w \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(x,z) : z \in V_H\}\!\}$ and $\{\!\{\chi^{(t)}_{2,\psi_G}(w,v) : w \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(z,y) : z \in V_H\}\!\}$. This implies that $\chi^{(t+1)}_{2,\psi_G}(u,v) = \chi^{(t+1)}_{2,\psi_H}(x,y)$ which concludes the proof. □

Now we are ready to prove our main theorem in this section.

Theorem 3.5 (Equivalence of RPE-augWL and RPE-2-WL.)**.** *Let ψ be a* pseudo-symmetric *RPE. Then two featured graphs (G, X_G) and (H, X_H) are ψ-WL indistinguishable iff they are ψ-2-WL indistinguishable.*

Proof of Theorem 3.5. For the forward direction, we will show that for all $t \geq 0$, if $\{\!\{\chi^{(t+1)}_{\psi_G}(u) : u \in V_G\}\!\} = \{\!\{\chi^{(t+1)}_{\psi_H}(v) : v \in V_H\}\!\}$, then $\{\!\{\chi^{(t)}_{2,\psi_G}(u,v) : u,v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(x,y) : x,y \in V_H\}\!\}$. This implies that if G and H are indistinguishable by the ψ-WL, they are indistinguishable by the ψ-2-WL.

We first prove the base case of $t = 0$. As $\{\!\{\chi^{(1)}_{\psi_G}(u) : u \in V_G\}\!\} = \{\!\{\chi^{(1)}_{\psi_H}(v) : v \in V_H\}\!\}$, then there is a bijection $\sigma : V_G \to V_H$ such that $\chi^{(1)}_{\psi_G}(u) = \chi^{(1)}_{\psi_G}(\sigma(u))$. Moreover, by the definition of $\chi^{(1)}_{\psi}$, for each $u \in V$, there is a bijection $\sigma_u : V_G \to V_H$ such that $\psi_G(u,v) = \psi_H(\sigma(u), \sigma_u(v))$. Therefore, we can define a bijection $\Phi : V_G \times V_G \to V_H \times V_H$ defined $\Phi(u,v) = (\sigma(u), \sigma_u(v))$ such that $\psi_G(u,v) = \psi_H(\Phi(u,v))$. Therefore, $\{\!\{\chi^{(0)}_{2,\psi_G}(u,v) : u,v \in V_G\}\!\} = \{\!\{\chi^{(0)}_{2,\psi_H}(x,y) : x,y \in V_H\}\!\}$.

We now prove that for each $t \geq 1$, if $\{\!\{\chi^{(t+1)}_{\psi_G}(u) : u \in V_G\}\!\} = \{\!\{\chi^{(t+1)}_{\psi_H}(v) : v \in V_H\}\!\}$, then $\{\!\{\chi^{(t)}_{2,\psi_G}(u,v) : u,v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(x,y) : x,y \in V_H\}\!\}$. Assume that both graphs are assigned $k^{(t+1)}$ colors by $\chi^{(t+1)}$. More specifically, assume there are partitions $V_G = V^{(t+1)}_{G,1} \cup \cdots \cup V^{(t+1)}_{G,k^{(t+1)}}$ and $V_H = V^{(t+1)}_{H,1} \cup \cdots \cup V^{(t+1)}_{H,k^{(t+1)}}$ such that

1. $|V^{(t+1)}_{G,i}| = |V^{(t+1)}_{H,i}|$;
2. for any $u \in V^{(t+1)}_{G,i}$ and $v \in V^{(t+1)}_{G,j}$, $\chi^{(t+1)}_{\psi_G}(u) = \chi^{(t+1)}_{\psi_G}(v)$ iff $i = j$;
3. for any $u \in V^{(t+1)}_{G,i}$ and $x \in V^{(t+1)}_{H,i}$, $\chi^{(t+1)}_{\psi_G}(u) = \chi^{(t+1)}_{\psi_H}(x)$.

We can define analogous partitions for $\chi^{(t)}_{\psi}$.

These partitions induce a partition on the set of pairs of vertices $V_G \times V_G = \cup_{1 \leq i \leq k^{(t+1)}} \cup_{1 \leq j \leq k^{(t)}} V^{(t+1)}_{G,i} \times V^{(t)}_{G,j}$ and $V_H \times V_H = \cup_{1 \leq i \leq k^{(t+1)}} \cup_{1 \leq j \leq k^{(t)}} V^{(t+1)}_{H,i} \times V^{(t)}_{H,j}$; see Figure 4 for an illustration of the decomposition.

We show that $\chi^{(t)}_{2,\psi}$ has the same multiset of colors on $V^{(t+1)}_{G,i} \times V^{(t)}_{G,j}$ and $V^{(t+1)}_{H,i} \times V^{(t)}_{H,j}$. By the definition of the partitions, there exists a bijection $\sigma^{(t+1)}_i : V^{(t+1)}_{G,i} \to V^{(t+1)}_{H,i}$ such that $\chi^{(t+1)}_{\psi_G}(u) = \chi^{(t+1)}_{\psi_H}(\sigma^{(t+1)}_i(u))$ for any $u \in V^{(t+1)}_{G,i}$. By definition of ψ-WL, we have that $\chi^{(t)}_{\psi_G}(u) = \chi^{(t)}_{\psi_H}(\sigma^{(t+1)}_i(u))$ for any $u \in V^{(t+1)}_{G,i}$. Furthermore, for any $u \in V^{(t+1)}_{G,i}$,

$$\{\!\{ (\chi^{(t)}_{\psi_G}(w), \psi_G(u,w)) : w \in V_G \}\!\} = \{\!\{ (\chi^{(t)}_{\psi_H}(z), \psi_H(\sigma^{(t+1)}_i(u), z)) : z \in V_H \}\!\}.$$

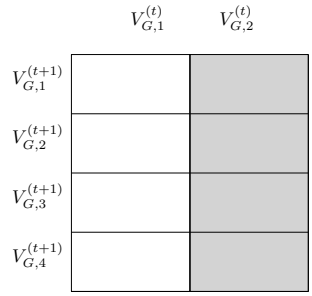


Figure 4. Decomposition of $V_G \times V_G$.

Now for each $1 \leq j \leq k^{(t)}$, consider the sub-multiset $\{\!\{ (\chi^{(t)}_{\psi_G}(w), \psi_G(u,w)) : w \in V^{(t)}_{G,j} \}\!\}$ of $\{\!\{ (\chi^{(t)}_{\psi_G}(w), \psi_G(u,w)) : w \in V_G \}\!\}$, then we have that

$$\{\!\{ (\chi^{(t)}_{\psi_G}(w), \psi_G(u,w)) : w \in V^{(t)}_{G,j} \}\!\} = \{\!\{ (\chi^{(t)}_{\psi_H}(z), \psi_H(\sigma_i^{(t+1)}(u), z)) : z \in V^{(t)}_{H,j} \}\!\}.$$

This is because when $j \neq b$, one has that $\chi^{(t)}_{\psi_G}(w) \neq \chi^{(t)}_{\psi_H}(z)$ for $w \in V^{(t)}_{G,j}$ and $z \in V^{(t)}_{H,b}$.

Therefore,

$$\begin{aligned}&\{\!\{(\psi_G(u,w), \chi^{(t)}_{\psi_G}(u), \chi^{(t)}_{\psi_G}(w)) : w \in V^{(t)}_{G,j}\}\!\}\\ =&\{\!\{(\psi_H(\sigma_i^{(t+1)}(u), z), \chi^{(t)}_{\psi_H}(\sigma_i^{(t+1)}(u)), \chi^{(t)}_{\psi_H}(z)) : z \in V^{(t)}_{H,j}\}\!\}.\end{aligned}$$

Therefore, by Lemma A.1, one has that

$$\{\!\{\chi^{(t)}_{2,\psi_G}(u,w) : w \in V^{(t)}_{G,j}\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(\sigma_i^{(t+1)}(u), z) : z \in V^{(t)}_{H,j}\}\!\}.$$

Since $\sigma_i^{(t+1)}$ is bijective, we conclude that

$$\{\!\{\chi^{(t)}_{2,\psi_G}(u,v) : u \in V^{(t+1)}_{G,i}, v \in V^{(t)}_{G,j}\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(x,y) : x \in V^{(t+1)}_{H,i}, y \in V^{(t)}_{H,j}\}\!\}.$$

Taking the multiset union over $1 \leq i \leq k^{(t+1)}$ and $1 \leq j \leq k^{(t)}$, this implies that $\{\!\{\chi^{(t)}_{2,\psi_G}(u,v) : u,v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(x,y) : x,y \in V_H\}\!\}$.

Conversely, we will show that for $t \geq 0$, if $\{\!\{\chi^{(t)}_{2,\psi_G}(u,v) : u,v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(x,y) : x,y \in V_H\}\!\}$, then $\{\!\{\chi^{(t)}_{\psi_G}(u) : u \in V_G\}\!\} = \{\!\{\chi^{(t)}_{\psi_H}(x) : x \in V_H\}\!\}$.

For the base case of $t = 0$, we have that $\{\!\{(X_G(u), X_G(v), \psi_G(u,v)) : u,v \in V_G\}\!\} = \{\!\{(X_H(x), X_H(y), \psi_H(x,y)) : x,y \in V_H\}\!\}$. Therefore, $\{\!\{(X_G(u), X_G(v)) : u,v \in V_G\}\!\} = \{\!\{(X_H(x), X_H(y)) : x,y \in V_H\}\!\}$ and hence $\{\!\{X_G(u) =: u \in V_G\}\!\}^2 = \{\!\{X_H(x) : x \in V_H\}\!\}^2$. Now we apply the following fact:

Fact 3. *Let A and B be two finite multisets. If $A^2 = B^2$, then $A = B$.*

Then, we have that $\{\!\{\chi^{(0)}_{\psi_G}(u) = X_G(u) : u \in V_G\}\!\} = \{\!\{\chi^{(0)}_{\psi_H}(x) = X_H(x) : x \in V_H\}\!\}$.

Now assume that for $k \geq 1$, we have that

$$\{\!\{\chi^{(t)}_{2,\psi_G}(u,v) : u,v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{2,\psi_H}(x,y) : x,y \in V_H\}\!\}.$$

Lemma A.1 implies that

$$\{\!\{(\psi_G(u,v), \chi^{(t)}_{\psi_G}(u), \chi^{(t)}_{\psi_G}(v)) : u,v \in V_G\}\!\} = \{\!\{(\psi_H(x,y), \chi^{(t)}_{\psi_H}(x), \chi^{(t)}_{\psi_H}(y)) : x,y \in V_H\}\!\}.$$

Dropping the ϕ term, one has that

$$\{\!\{(\chi^{(t)}_{\psi_G}(u), \chi^{(t)}_{\psi_G}(v)) : u,v \in V_G\}\!\} = \{\!\{(\chi^{(t)}_{\psi_H}(x), \chi^{(t)}_{\psi_H}(y)) : x,y \in V_H\}\!\}.$$

This implies that the following Cartesian products are equal:

$$\{\!\{\chi^{(t)}_{\psi_G}(u) : u \in V_G\}\!\}^2 = \{\!\{\chi^{(t)}_{\psi_H}(x) : x \in V_H\}\!\}^2.$$

Then, by Fact 3 again, we have that

$$\{\!\{\chi^{(t)}_{\psi_G}(u) : u \in V_G\}\!\} = \{\!\{\chi^{(t)}_{\psi_H}(x) : x \in V_H\}\!\}.$$

This means that G and H are indistinguishable by ψ-WL.

We now end this proof by proving Fact 3.

Proof of Fact 3. We let $C := A \cup B$ denote the set of all distinct elements in A and B. Let $C = \{c_1, \dots, c_n\}$ and assume that A contains r_i^A many c_i and B contains r_i^B many c_i for each $i = 1, \dots, n$. Then, A^2 contains $r_i^A r_j^A$ many (c_i, c_j) and B^2 contains $r_i^B r_j^B$ many (c_i, c_j). Therefore, $A^2 = B^2$ implies that $r_i^A r_j^A = r_i^B r_j^B$ for all $i, j = 1, \dots, n$. For each $i = 1, \dots, n$, we let $j = i$ and hence $(r_i^A)^2 = (r_i^B)^2$. Since we are dealing with nonnegative integers, we must have that $r_i^A = r_i^B$. Therefore, $A = B$. □

□

A.3. RPE-2-WL vs 2-EGN

Equivariant or invariant graph networks (EGNs and IGNs) (Maron et al., 2019b) are another important type of graph neural networks. These models will be important when trying to understand the distinguishing power of positional encodings later.

Definition A.2 (EGNs)**.** A kth **order Equivariant Graph Network (k-EGN)** is a function $F : \mathbb{R}^{n^k \times d_0} \to \mathbb{R}^{n^s \times d_o}$ for $s \leq k$ of the form

$$F = h \circ m \circ L^{(T)} \circ \sigma \circ \cdots \circ \sigma \circ L^{(1)}$$

where h and $L^{(t)} : \mathbb{R}^{n^{k_{t-1}} \times d_{t-1}} \to \mathbb{R}^{n^{k_t} \times d_t}$ are equivariant linear layers for each $t = 1, \dots, T$ such that $k_t \leq k$, $k_0 = k$ and $k_L = s$, σ is an activation function, and m is a pointwise MLP. When $s = 0$, the resulting k-EGN is also called a kth **order Invariant Graph Network (k-IGN)**.

Note: Some implementations of 2-IGNs and 2-EGNs (e.g. (Maron et al., 2019b)) stack the adjacency matrix to the input W. In this paper, unless otherwise stated, the adjacency matrix is **not** stacked to the input.

EGNs are permutation equivariant in the sense that permutation of the order of input tensors will result in a corresponding permutation of the order of output tensors. Likewise, IGNs are permutation invariant as any permutation of the input tensors results in the same output.

We are interested in 2-EGNs and 2-IGNs since they are closely related to the 2-WL graph isomorphism test. Let ψ be an RPE valued in $\mathbb{R}^k$ and let $g : \mathbb{R}^{n^2 \times k} \to \mathbb{R}^{n \times l}$ be any 2-EGN. Recall that for any graph G with n vertices, the RPE ψ_G can be represented as a tensor in $\mathbb{R}^{n^2 \times k}$. If $X_G \in \mathbb{R}^{n \times l}$ are node features for G, let $\mathrm{row}(X_G), \mathrm{col}(X_G) \in \mathbb{R}^{n^2 \times l}$ be the tensors where each row or column respectively are the node feature X_G, e.g. $\mathrm{row}(X_G)[i, :, :] = X_G$ and $\mathrm{col}(X_G)[:, j, :] = X_G$ for all $1 \leq i, j \leq n$.

We now show that 2-EGNs with the input $[\psi_G, \mathrm{row}(X_G), \mathrm{col}(X_G)] \in \mathbb{R}^{n^2 \times (k+2l)}$ have the same distinguishing powers as the ψ-2-WL test.

Proposition A.3 (Equivalence of RPE-2-WL and 2-EGN)**.** *Let ψ be a* diagonally-aware *RPE valued in $\mathbb{R}^k$. Let $\mathcal{G}$ be any finite set of featured graphs. Then there exists a 2-EGN $g : \mathbb{R}^{n^2 \times k} \to \mathbb{R}^{n \times l}$ such that any $(G, X_G), (H, X_H) \in \mathcal{G}$ can be distinguished by g with the input $[\psi, \mathrm{row}(X), \mathrm{col}(X)]$ iff G and H can be distinguished by the ψ-2-WL test.*

This result is a generalization of existing results showing that 2-IGNs of the form $g : \mathbb{R}^{n^2 \times k} \to \mathbb{R}^m$ are equivalent to the 2-WL test (Maron et al., 2019a; Chen et al., 2020).

Proof of Proposition A.3. We will prove the forward and backward direction of this statement separately in the following lemmas. However, we emphasize that only Lemma A.5 needs the assumption of diagonal-awareness. Moreover, we

emphasize that the "proofs" of both of these lemmas are just slightly adapting existing results connecting 2-IGNs to the 2-WL test.

Lemma A.4 (2-EGNs are stronger than RPE-2-WL)**.** *Let ψ be an RPE valued in $\mathbb{R}^k$. Let $\mathcal{G}$ be a finite set of graphs. Then, there exists a 2-EGN $g : \mathbb{R}^{n^2\times(k+2l)} \to \mathbb{R}^{n\times m}$ such that if any $(G, X_G), (H, X_H) \in \mathcal{G}$ can be distinguished by ψ-2-WL test then $\{\!\{g([\psi_G, \mathrm{row}(X_G), \mathrm{col}(X_G)])(u) : u \in V_G\}\!\} \neq \{\!\{g([\psi_H, \mathrm{row}(X_H), \mathrm{col}(X_H)])(v) : v \in V_H\}\!\}$.*

Proof. This follows from the proof by Maron et al. (2019a, Theorem 1) showing that there exists a 2-EGN that can distinguish any two graphs from a finite set of graphs well as the classical 2-WL algorithm. (While their statement of Theorem 1 is only for two graphs, their proof explains why this can be generalized to a finite set of graphs.) This lemma shows how this result can be generalized to the RPE-2-WL test.

Their proof is broken into two steps. First, the authors show that given a tensor $W \in \mathbb{R}^{n^2\times(e+1)}$ of a diagonal feature matrix stacked with the adjacency matrix, there is a 2-equivariant linear layer $f : \mathbb{R}^{n^2\times(e+1)} \to \mathbb{R}^{n^2\times 4\times(e+2)}$ such that, for each pair (u, v), the vector $f(W)[u, v, :]$ corresponds to the isomorphism type of (u, v), which is the initial color of (u, v) in the 2-WL test. Second, they show that there is a d layers 2-EGN $f = L^{(d)} \circ \sigma \circ \cdots \circ \sigma \circ L^{(1)} : \mathbb{R}^{n^2\times k} \to \mathbb{R}^{n^2\times l}$ that can implement d iterations of the 2-WL algorithm; that is to say, $f(A_G)(u, v) = f(A_H)(u, v)$ iff $\chi^{(d)}(u, v) = \chi^{(d)}(x, y)$. Finally, they show that there exists an MLP m, together with a final layer h of summing over all entries, such that $g = h \circ m \circ f$ implements the 2-WL test, i.e. $g(A_G) = g(A_H)$ iff $\chi(G) = \chi(H)$.

The difference between the 2-WL test and the RPE-2-WL test lies in the initialization step. In the classical 2-WL test, a pair (u, v) is colored with its isomorphism type, while in the RPE-2-WL test, the initial color is $(X_G(u), X_G(v), \psi_G(u, v))$. If the input to the 2-EGN is the RPE matrix concatenated with the node features (i.e. the initial coloring), we can skip the initialization step and just perform the update steps of the RPE-2-WL through a 2-EGN $f = L^{(d)} \circ \sigma \circ \cdots \circ \sigma \circ L^{(1)} : \mathbb{R}^{n^2\times k} \to \mathbb{R}^{n^2\times l}$. Finally, instead of taking the sum of all entries, we can apply the operation of row sum (which is equivariant) to the output of f to obtain our target 2-EGN $g : \mathbb{R}^{n^2\times k} \to \mathbb{R}^{n\times l}$. □

Lemma A.5 (RPE-2-WL is stronger than 2-EGNs)**.** *Let ψ be a* diagonally-aware *RPE. Let (G, X_G) and (H, X_H) be graphs. If G and H cannot be distinguished by ψ-2-WL, then (G, X_G) and (H, X_H) cannot be distinguished by any 2-EGN with input $[\psi, \mathrm{row}(X), \mathrm{col}(X)]$.*

Proof. This follows directly from the proof showing that if two graphs cannot be distinguished by the 2-WL, they cannot be distinguished by 2-EGNs (Chen et al., 2020, Lemma 4.2). The only reason we cannot apply this proof directly is that it relies on the fact that the initial colors to the 2-WL—and therefore all subsequent colors—are diagonally-aware (Chen et al., 2020, Lemma D.1), which is not necessarily the case for the initial colors of RPE-2-WL for general RPEs. However, if the RPE is diagonally-aware, then our lemma can be proven following the proof of Chen et al. (2020, Lemma 4.2). □

□

Importance of diagonal-awareness. In the proof of Proposition A.3 (specifically in Lemma A.5), we show that RPE-2-WL is at least as strong as 2-EGNs if the RPE is diagonally-aware. We emphasize that if we drop the assumption of diagonal-awareness, the lemma does not hold. Therefore, 2-EGN can be strictly stronger than th RPE-2-WL test for non-diagonally-aware RPEs, as shown by the following simple example.

Consider the two graphs on two vertices G and H where G has an edge and H does not. (In this example, the graphs have no node features.) Consider the (artificial) RPE ψ that assigns

$$\psi_G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \psi_H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Note that ψ is *not* diagonally-aware, as the values 0 and 1 appear on both the diagonal and off-diagonal elements.

The ψ-2-WL colorings for $t = 0$ are

$$\chi^{(0)}_{2,\psi}(G) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \chi^{(0)}_{2,\psi}(H) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

These are, in fact, already stable ψ-2-WL colorings, as another iteration of ψ-2-WL gives the same partition. This means further iterations of ψ-2-WL won't be able to distinguish these two graphs. For example,

$$\chi^{(1)}_{2,\psi}(G) = \begin{bmatrix} (1, (\{\!\{0,1\}\!\}, \{\!\{0,1\}\!\})) & (0, (\{\!\{0,1\}\!\}, \{\!\{0,1\}\!\})) \\ (0, (\{\!\{0,1\}\!\}, \{\!\{0,1\}\!\})) & (1, (\{\!\{0,1\}\!\}, \{\!\{0,1\}\!\})) \end{bmatrix}$$

$$\chi^{(1)}_{2,\psi}(H) = \begin{bmatrix} (0, (\{\!\{0,1\}\!\}, \{\!\{0,1\}\!\})) & (1, (\{\!\{0,1\}\!\}, \{\!\{0,1\}\!\})) \\ (1, (\{\!\{0,1\}\!\}, \{\!\{0,1\}\!\})) & (0, (\{\!\{0,1\}\!\}, \{\!\{0,1\}\!\})) \end{bmatrix},$$

As $\chi^{(0)}_{2,\psi}(G)$ and $\chi^{(0)}_{2,\psi}(H)$ have the same multiset of colors $\{\!\{0,0,1,1\}\!\}$, they are indistinguishable by ψ-2-WL. However, there is a 2-IGN that distinguishes these graphs. Specifically, one invariant linear matrix is $B : \mathbb{R}^{n\times n} \to \mathbb{R}$ that maps $B(W) = \sum_{v\in V} W_{vv}$. (In the language of the original IGN paper (Maron et al., 2019b), this corresponds to the partition $\gamma = \{\!\{1,2\}\!\}$.) Then $B(\psi_G) = 2 \neq B(\psi_H) = 0$, so this 2-IGN distinguishes ψ_G and ψ_H.

A.4. Equivalence of APEs and APE-GTs: Proof of Lemma 3.1

Lemma 3.1 (Equivalence of APEs and APE-GT)**.** *Any two graphs (G, X_G) and (H, X_H) are indistinguishable by an APE ϕ iff (G, X_G) and (H, X_H) are indistinguishable by all ϕ-APE-GTs.*

Proof. Let T be any transformer, and let (G, X_G) and (H, X_H) indistinguishable by ϕ. The initial input to T are $\hat{X}_G = [X_G|\phi_G]$ and $\hat{X}_H = [X_H|\phi_H]$. As G and H are indistinguishable by ϕ, then there is a permutation P such that $P\hat{X}_G = \hat{X}_H$; however, as transformers are permutation-equivariant (Fact 1), then $T(\hat{X}_H) = T(P\hat{X}_G) = PT(\hat{X}_G)$, so the theorem follows.

Now we prove the inverse. Consider the transformer T with all weight matrices set to the zero matrix. The transformer T is exactly the identity map. Thus, if G and H are distinguishable by ϕ, then they are distinguishable by T. □

A.5. Isomorphic implies WL Indistinguishable: Proof of Fact 2

Fact 2. *If (G, X_G) and (H, X_H) are feature isomorphic, then (G, X_G) and (H, X_H) are ψ-WL indistinguishable and ψ-2-WL indistinguishable for any RPE ψ.*

Proof. Let (G, X_G) and (H, X_H) be feature isomorphic graphs connected by isomorphism $\sigma : V_G \to V_H$. Let ψ be an RPE. We will prove the stronger statement that $\chi^{(t)}_{\psi_G}(u) = \chi^{(t)}_{\psi_H}(\sigma(u))$ for all $v \in V_G$ and $t \geq 0$. For the base case of $t = 0$, this is trivial as $\chi^{(0)}_{\psi_G}(u) = X_G(u) = X_H(\sigma(u)) = \chi^{(0)}_{\psi_H}(\sigma(u))$ for all $v \in V_G$ by the definition of feature isomorphism.

Now suppose the statement is true for t, and we will show it is true for $t+1$. Let $u \in V_G$. Consider the colors $\chi^{(t+1)}_{\psi_G}(u) = (\chi^{(t)}_{\psi_G}(u), \{\!\{(\chi^{(t)}_{\psi_G}(v), \psi_G(u,v)) : v \in V_G\}\!\})$ and $\chi^{(t+1)}_{\psi_H}(\sigma(u)) = (\chi^{(t)}_{\psi_H}(\sigma(u)), \{\!\{(\chi^{(t)}_{\psi_H}(v), \psi_H(\sigma(u), v)) : v \in V_H\}\!\})$. By induction, we know that the first coordinates $\chi^{(t)}_{\psi_G}(u) = \chi^{(t)}_{\psi_H}(\sigma(u))$. Moreover, by the inductive hypothesis and the definition of RPE, we know that $(\chi^{(t)}_{\psi_G}(v), \psi_G(u,v)) = (\chi^{(t)}_{\psi_H}(\sigma(v)), \psi_H(\sigma(u), \sigma(v)))$, which implies the second coordinates are equals. Therefore, the theorem holds.

A similar argument shows that $\chi^{(t)}_{2,\psi_G}(u,v) = \chi^{(t)}_{2,\psi_H}(\sigma(u), \sigma(v))$ for all $u, v \in V_G$ and $t \geq 0$. □

A.6. Equivalence of RPE-augWL and RPE-GTs: Proof of Lemma 3.6

Lemma 3.6 (Equivalence of RPE-augWL and RPE-GT)**.** *Let ψ be a* diagonally-aware *RPE. Let (G, X_G) and (H, X_H) be featured graphs. Then (G, X_G) and (H, X_H) are indistinguisable by the ψ-WL test iff (G, X_G) and (H, X_H) are indistinguisable by all ψ-RPE-GTs.*

Proof. We will denote the input at the l-th layer of the RPE-GT for graph G as $X^{(l)}_G \in \mathbb{R}^{|V_G|\times\ell}$, and for each node $u \in V_G$, its feature is denoted by $X^{(l)}_G(u)$. Note $X^{(0)}_G(u) = X_G(u)$. We adopt similar notations for H.

For the forward direction, it suffices to show that if G and H are indistinguishable by the ψ-WL test, they are indistinguishable by any ψ-RPE-GT. We will prove the stronger statement that if the multiset of colors $\{\!\{\chi^{(t)}_{\psi_G}(u) : v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{\psi_H}(x) :$

$x \in V_H\}\!\}$, then any bijection $\sigma^{(t)} : V_G \to V_H$ such that $\chi^{(t)}_{\psi_G}(u) = \chi^{(t)}_{\psi_H}(\sigma^{(t)}(u))$ will satisfy the following property: $X^{(t)}_G(u) = X^{(t)}_H(\sigma^{(t)}(u))$ for all $u \in V_G$, and hence $\{\!\{X^{(t)}_G(u) : u \in V_G\}\!\} = \{\!\{X^{(t)}_H(x) : x \in V_H\}\!\}$. We will prove this by induction on t.

The key observation is that the only part of the transformer architecture that is not applied separately to the node features $X^{(t)}(u)$ is the multiplication by the attention matrix $A^{(l,h)}(X^{(l)})X^{(l)}$, meaning for all other parts of the transformer, if the inputs $X^{(t)}_G(u) = X^{(t)}_H(x)$ are equal, then the outputs at u and x are equal.

For the base case of $t = 0$, the statement holds as the initial node features equals the ψ-WL colors for $t = 0$, i.e. $\{\!\{\chi^{(0)}_{\psi_G}(v) : v \in V_G\}\!\} = \{\!\{X^{(0)}_G(v) : v \in V_G\}\!\}$.

Now suppose the theorem inductively holds for some t. Assume that $\{\!\{\chi^{(t+1)}_{\psi_G}(u) : v \in V_G\}\!\} = \{\!\{\chi^{(t+1)}_{\psi_H}(x) : x \in V_H\}\!\}$ and let $\sigma^{(t+1)} : V_G \to V_H$ denote any bijection such that $\chi^{(t+1)}_{\psi_G}(u) = \chi^{(t+1)}_{\psi_H}(\sigma^{(t+1)}(u))$ for any $u \in V_G$. By the definition of χ_ψ and the inductive hypothesis, we have that $\{\!\{\chi^{(t)}_{\psi_G}(u) : v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{\psi_H}(x) : x \in V_H\}\!\}$, $\chi^{(t)}_{\psi_G}(u) = \chi^{(t+1)}_{\psi_H}(\sigma^{(t)}(u))$ and $X^{(t)}_G(u) = X^{(t)}_H(\sigma^{(t+1)}(u))$ for all $u \in V_G$.

Now for any $v \in V_G$ and $y \in V_H$ such that $X^{(t)}_G(v) = X^{(t)}_H(y)$, we have that

$$X^{(t)}_G(u)W_Q(X^{(t)}_G(v)W_K)^T = X^{(t)}_H(\sigma^{(t+1)}(u))W_Q(X^{(t)}_H(y)W_K)^T \tag{1}$$

for any matrices W_Q and W_K.

Now, for any $u \in V_G$, $\chi^{(t+1)}_{\psi_G}(u) = \chi^{(t+1)}_{\psi_H}(\sigma^{(t+1)}(u))$ implies that

$$\{\!\{(\chi^{(t)}_{\psi_G}(v), \psi_G(u, v)) : v \in V_G\}\!\} = \{\!\{(\chi^{(t)}_{\psi_H}(y), \psi_H(\sigma^{(t+1)}(u), y)) : y \in V_H\}\!\}.$$

Hence,

$$\{\!\{(X^{(t)}_G(v), \psi_G(u, v)) : v \in V_G\}\!\} = \{\!\{(X^{(t)}_H(y), \psi_H(\sigma^{(t+1)}(u), y)) : y \in V_H\}\!\}.$$

Note that for any given value a,

$$\{\!\{\psi_G(u, v) : v \in V_G, X^{(t)}_G(v) = a\}\!\} = \{\!\{\psi_H(\sigma^{(t+1)}(u), y) : y \in V_H, X^{(t)}_H(y) = a\}\!\}. \tag{2}$$

Then, we have that

$$\begin{aligned}
&[A^{(t,h)}(X^{(t)}_G)X^{(t)}_G](u) = \sum_v A^{(t,h)}(X^{(t)}_G)(u, v)X^{(t)}_G(v) \\
&= \sum_v f_1(\psi_G(u, v))\mathrm{softmax}(X^{(t)}_G(u)W_Q(X^{(t)}_G(\cdot)W_K)^T/\sqrt{d_h} + f_2(\psi_G(u, \cdot)))(v)X^{(t)}_G(v) \\
&= \sum_y f_1(\psi_H(\sigma^{(t+1)}(u), y))\mathrm{softmax}(X^{(t)}_H(\sigma^{(t+1)}(u))W_Q(X^{(t)}_H(\cdot)W_K)^T/\sqrt{d_h} + f_2(\psi_H(\sigma^{(t+1)}(u), \cdot)))(y)X^{(t)}_H(y) \\
&= [A^{(t,h)}(X^{(t)}_H)X^{(t)}_H](\sigma^{(t+1)}(u)).
\end{aligned}$$

Here the third equation follows from Equation (1) and Equation (2).

It is then easy to check that $X^{(t+1)}_G(u) = X^{(t+1)}_H(\sigma^{(t+1)}(u))$ for any $u \in V_G$ and this concludes the proof.

For the other direction, the proof follows from the one of Zhang et al. (2023, Theorem 4); More precisely, Zhang et al. (2023) are dealing with the case when the RPE ψ is taking value from a finite set D for graphs with bounded sizes. Furthermore, their RPE ψ is a distance (or a stack of distance functions) so that $\psi_G(u, v) = 0$ iff $u = v \in V_G$ for any graph G. In this way, elements in D can be written as $0 = r_0 < r_1 < \cdots r_H$. Note that, for one step of ψ-WL, the second argument can be decomposed by

$$\{\!\{(\chi^{(t)}_{\psi_G}(u),\ \psi_G(v, u)) : u \in V_G\}\!\} = \bigcup_{r_i} \{\!\{(\chi^{(t)}_{\psi_G}(u), r_i) :\ \psi_G(v, u) = r_i, u \in V_G\}\!\}.$$

Then, Zhang et al. (2023) explicitly constructed $H+1$ different RPE attention heads A^h_{RPE} which can simulate $\{\!\{(\chi^{(t)}_{\psi_G}(u), r_i) : \psi_G(v,u) = r_h, u \in V_G\}\!\}$ in an injective manner for each $h = 0, \ldots, H$. In this way, they argue that by concatenating these heads, one can simulate their GD-WL test where each step is updated by $\widehat{\chi}^{(t+1)}_{\psi_G}(v) = \{\!\{(\widehat{\chi}^{(t)}_{\psi_G}(u),\, \psi_G(v,u)) : u \in V_G\}\!\}$. But as mentioned already in Section 3.1, this GD-WL test is equivalent to the ψ-WL test when ψ is a graph distance (or a stack of graph distances), so their one layer RPE transformer can simulate one step of ψ-WL in an injective manner.

Recall that GD-WL is equivalent to ψ-WL due to the fact that ψ takes value 0 on diagonal entries and ψ takes non-zero values on off-diagonal entries. This is the reason why one can drop the first argument in $(\chi^{(t)}_{\psi_G}(v), \{\!\{(\chi^{(t)}_{\psi_G}(u),\, \psi_G(v,u)) : u \in V\}\!\})$ without losing distinguishing power. Hence, the same argument in the proof of (Zhang et al., 2023, Theorem 4) can be applied to the setting of diagonally-aware RPEs due to the following simple result.

Lemma A.6. *When ψ is a diagonally-aware RPE, then ψ-WL is equivalent to the following variant of ψ-WL:*

$$\widehat{\chi}^{(0)}_{\psi_G}(v) = X_G(v),$$
$$\widehat{\chi}^{(t+1)}_{\psi_G}(v) = \{\!\{(\widehat{\chi}^{(t)}_{\psi_G}(u),\, \psi_G(v,u)) : u \in V\}\!\}.$$

Now, to prove our theorem, we simply let $D := \{\psi_G(u,v) : u, v \in V_G\} \cup \{\psi_H(x,y) : x, y \in V_H\}$ and apply the same argument as in the proof of (Zhang et al., 2023, Theorem 4) to conclude the proof. □

A.7. Mapping APEs to RPEs: Proof of Lemma 3.7 and Theorem 3.8

We first recall a key lemma from Xu et al. (2018). For any set S, we let $\mathrm{Mul}(S)$ denote the collection of all finite multi-subsets of S and recall that $\mathrm{Mul}_2(S)$ denotes the collection of all multi-subsets of size 2 in S.

Lemma A.7 ((Xu et al., 2018, Lemma 5))**.** *If S is countable, then there exists a function $f : S \to \mathbb{R}$ such that the map $h_f : \mathrm{Mul}(S) \to \mathbb{R}$ defined by sending $h_f(X) = \sum_{x \in X} f(x)$ is injective.*

As a direct consequence, we immediately have the following corollary.

Corollary A.8. *If S is countable, then there exists a function $f : S \to \mathbb{R}$ such that the map $h_f : \mathrm{Mul}_2(S) \to \mathbb{R}$ defined by sending $h_f(\{\!\{x, y\}\!\}) = f(x) + f(y)$ is injective.*

Lemma 3.7. *Let ϕ be an APE. Then there is a function f such that any two featured graphs (G, X_G) and (H, X_H) are indistinguishable by ϕ iff (G, X_G) and (H, X_H) cannot be distinguished by the ψ^f-2-WL test.*

Proof. Let $S = \cup_{G \in \mathcal{G}} \operatorname{im} \phi_G$ where $\mathcal{G}$ is the set of all finite graphs. We first note that S is countable as the set of all finite graphs is countable. Let f be any function such that h_f that is injective on S; such a function must exist by Corollary A.8 as S is countable. Assume that G and H are indistinguishable by the ψ^f-2-WL test. This, in particular, implies that the multisets of initial colors are the same:

$$\{\!\{(X_G(u), X_G(v), \psi^f_G(v,w)) : u, v \in V_G\}\!\} = \{\!\{(X_H(x), X_H(y), \psi^f_H(x,y)) : x, y \in V_H\}\!\}.$$

By the definition of ψ^f and the fact that h_f is injective, one has that

$$\{\!\{(X_G(u), X_G(v), \phi_G(u), \phi_G(v)) : u, v \in V_G\}\!\} = \{\!\{(X_H(x), X_H(y), \phi_H(x), \phi_H(y)) : x, y \in V_H\}\!\}.$$

Hence, the following two Cartesian products of multisets are equal:

$$\{\!\{(X_G(u), \phi_G(u)) : u \in V_G\}\!\}^2 = \{\!\{(X_H(x), \phi_H(x)) : x \in V_H\}\!\}^2.$$

By Fact 3, we have that $\{\!\{(X_G(u), \phi_G(u)) : u \in V_G\}\!\} = \{\!\{(X_H(x), \phi_H(x)) : x \in V_H\}\!\}$.

Conversely, assume that (G, X_G) and (H, X_H) are indistinguishable by ϕ. Then $\{\!\{(X_G(u), \phi_G(u) : u \in V_G\}\!\} = \{\!\{(X_H(x), \phi_H(x)) : x \in V_H\}\!\}$. Let $\sigma : V_G \to V_H$ be a bijection such that $\phi_G(v) = \phi_H(\sigma(v))$ and $X_G(v) = X_H(\sigma(v))$ for all $v \in V_G$. We will prove that (G, X_G) and (H, X_H) are indistinguishable by ψ^f-WL for any function f. In fact, we will prove the stronger statement that $\chi^{(t)}_{\psi^f_G}(v) = \chi^{(t)}_{\psi^f_H}(\sigma(v))$ for all $t \geq 0$. This is stronger than proving they are ψ^f-WL indistinguishable as we use the *same* bijection for all iterations t.

We prove this by induction on the iteration t. For $t = 0$, this is true as the initial ψ^f-WL colors $\chi^{(t)}_{\psi^f_H}(v)$ are just the node features $X_G(v)$, which are implied by the definition of indistinguishability by an APE. Now suppose this is true for some iteration t; we will show it is true for $t+1$. Let $v \in V_G$. The inductive hypothesis implies that the first element of the colors are equal $\chi^{(t)}_{\psi^f_G}(v) = \chi^{(t)}_{\psi^f_H}(\sigma(v))$. Moreover, if we consider the second element of the colors, we find that

$$\begin{aligned}
\{\!\{(\chi^{(t)}_{\psi^f_G}(w), \psi^f_G(v,w)) : w \in V_G\}\!\} &= \{\!\{(\chi^{(t)}_{\psi^f_G}(w), f(\phi_G(v)) + f(\phi_G(w))) : w \in V_G\}\!\} && \text{(defintion of } \psi^f_G\text{)} \\
&= \{\!\{(\chi^{(t)}_{\psi^f_G}(w), f(\phi_H(\sigma(v))) + f(\phi_H(\sigma(w)))) : w \in V_G\}\!\} && \text{(defintion of } \sigma\text{)} \\
&= \{\!\{(\chi^{(t)}_{\psi^f_H}(\sigma(w)), f(\phi_H(\sigma(v)) + f(\phi_H(\sigma(w))))) : w \in V_G\}\!\} && \text{(inductive hypothesis)} \\
&= \{\!\{(\chi^{(t)}_{\psi^f_H}(\sigma(w)), \psi^f_H(\sigma(v), \sigma(w))) : w \in V_G\}\!\} && \text{(defintion of } \psi^f_H\text{)} \\
&= \{\!\{(\chi^{(t)}_{\psi^f_H}(w), \psi^f_H(\sigma(v), w)) : w \in V_H\}\!\}. && (\sigma \text{ is a bijection})
\end{aligned}$$

Therefore, $\chi^{(t+1)}_{\psi^f_G}(v) = \chi^{(t+1)}_{\psi^f_H}(\sigma(v))$ as claimed. □

Theorem 3.8. *For any APE ϕ, there exists a function f such that any two graphs (G, X_G) and (H, X_H) are indistinguishable by all ϕ-APE-GTs iff they are indistinguishable by all ψ^f-RPE-GTs.*

Proof. By Lemma 3.1, two graphs (G, X_G) and (H, X_H) are indistinguishable by all APE transformers with APE ϕ iff they are indistinguishable by ϕ. By Lemma 3.7, this is equivalent to being indistinguishable by ψ^f-2-WL. Hence, by Lemma 3.6, this is equivalent to being indistinguishable by any RPE transformer with RPE ψ^f. □

A.8. Mapping RPE to APE: Proof of Theorem 3.10

Lemma 3.9. *Let ψ be a* diagonally-aware *RPE.*

For any 2-EGN g, if (G, X_G) and (H, X_H) are indistinguishable by the ψ-2-WL test then (G, X_G) and (H, X_H) are indistinguishable by ϕ^g.

Moreover, for any finite set of unfeatured graphs $\mathcal{G}$, there is a 2-EGN g such that if $G, H \in \mathcal{G}$ are indistinguishable by ϕ^g, then G and H are indistinguishable by the ψ-2-WL test.

Proof. Let (G, X_G) and (H, X_H) be featured graphs that are indistinguishable by ψ-2-WL. By Lemma A.5, this means that for any 2-EGN g of the appropriate dimension, $g([\psi_G, \mathrm{row}(X_G), \mathrm{col}(X_G)]) = g([\psi_G, \mathrm{row}(X_G), \mathrm{col}(X_G)])$. In particular, for any 2-EGN h of appropriate dimension, there is a 2-EGN g_h such that $g_h([\psi_G, X_G X_G^T]) = [h(\psi_G), X_G^2]$; the 2-EGN g_h runs h on the first columns of $[\psi_G, \mathrm{row}(X_G), \mathrm{col}(X_G)]$ and then copies the mean of the rows of $\mathrm{row}(X_G)$ on the remaining columns. The output of g_h is exactly the APE ψ^h_G concatenated to the node features X_G.

For the backward direction of the case of graphs without node features, this follows directly from Lemma A.4. □

Theorem 3.10. *Let ψ be a* diagonally-aware *RPE.*

For any 2-EGN g, if (G, X_G) and (H, X_H) are indistinguishable by all ψ-RPE-GTs, then (G, X_G) and (H, X_H) are indistinguishable by ϕ^g.

Moreover, for any finite set of unfeatured *graphs $\mathcal{G}$, there is a 2-EGN g such that if $G, H \in \mathcal{G}$ are indistinguishable by all ϕ^g-APE-GTs, then G and H are indistinguishable by all ψ-RPE-GTs.*

Proof. By Lemma 3.6 and Theorem 3.5, G and H are indistinguisable by an RPE transformer with RPE ψ iff G and H are indistinguisable by the ψ-2-WL test. By Lemma 3.9, this is equivalent to G and H being indistinguishable by ϕ^g. By Lemma 3.1, this is equivalent to G and H being indistinguishable by any APE transformer with APE ϕ^g. □

We now show that RPEs are strictly stronger than APEs derived from RPEs for graphs with node features.

Example 1. There exists an RPE ψ and featured graphs (G, X_G) and (H, X_H) that are distinguishable by ψ-2-WL but are indistinguishable by ϕ^g for any 2-EGN g.

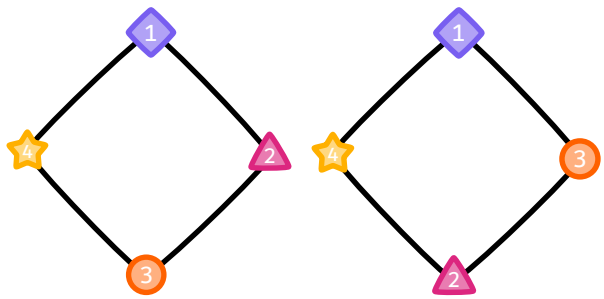


Figure 5. Left: (G, X_G). Right: (H, X_H).

Proof. The graph G and H are both be the cycle graph on 4-vertices C_4. The RPE ψ is the adjacency matrix. The node features are $X_G = [1, 2, 3, 4]$ and $X_H = [1, 3, 2, 4]$. See Figure 5.

We first prove these graphs are distinguishable by ψ-WL. If we consider the first ψ-WL colors $\chi_\psi^{(1)}$, we see that G has a node of color $(1, \{\!\{(2, 1), (3, 0), (4, 1)\}\!\})$; however, H has no node of that color, as the only node with feature 1 has color $(1, \{\!\{(3, 1), (2, 0), (4, 1)\}\!\})$.

We next prove that are indistinguishable by the APE ϕ^g for any 2-EGN g. The first thing to observe is that the APE ϕ^g will assign the same value to each node in G and H; this follows from the definition of APE and the fact that for any two nodes $u, v \in V_G$, there is a graph isomorphism $\sigma : V_G \to V_G$ such that $\sigma(u) = v$; likewise for any two nodes $x, y \in V_H$ or nodes $u \in V_G$ and $x \in V_H$. Therefore, as (G, X_G) and (H, X_H) have the same multiset of node features and ϕ^g has uniform entries, they cannot be distinguished by ϕ^g. □

B. Details from Section 4

B.1. A Lemma for RPE-augWL Composition.

In this section, we will repeatedly use the following useful property of composing the RPE-augWL test. Let ψ be an RPE valued in $\mathbb{R}^d$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be a function. Consider the composition $f \circ \psi$ defined by $(f \circ \psi)_G(u, v) := f(\psi_G(u, v))$ for any graph G. One naturally wonders about the distinguishing power of ψ-WL versus $f \circ \psi$-WL. In fact, ψ-WL will always be at least as strong as $f \circ \psi$-WL, and under mild conditions on the input, there will exist a function f such that ψ-WL and $f \circ \psi$-WL are equally strong.

Proposition B.1. *Let ψ be an RPE. Let G and H be graphs. Let $u \in V_G$ and $v \in V_H$.*

1. *If $\chi_{\psi_G}^{(t)}(u) = \chi_{\psi_H}^{(t)}(v)$, then for any function $f : \mathbb{R}^d \to \mathbb{R}$, $\chi_{f\circ\psi_G}^{(t)}(u) = \chi_{f\circ\psi_H}^{(t)}(v)$. In particular, ψ-WL is as strong as $f \circ \psi$-WL.*

2. *For any set $\Omega \subset \mathbb{R}^d$ and any function $f : \mathbb{R}^d \to \mathbb{R}$ injective on Ω, if $\operatorname{im} \psi_G \cup \operatorname{im} \psi_H \subset \Omega$ and $\chi_{\psi_G}^{(t)}(u) \neq \chi_{\psi_H}^{(t)}(v)$, then $\chi_{f\circ\psi_G}^{(t)}(u) \neq \chi_{f\circ\psi_H}^{(t)}(v)$. In particular, $f \circ \psi$-WL is as strong as ψ-WL.*

Proof. For item 1, it follows from (Zhu et al., 2023, Theorem 3).

For item 2, we choose the function f to be any function that is injective on Ω; such a function exists as Ω are finite. The lemma follows by induction of the iteration t. □

Remark B.2. It is quite common that RPEs will be drawn from a finite set $\Omega \subset \mathbb{R}^d$. In particular, if we consider any RPE defined on unweighted graphs with n vertices that is purely a function of its topology, e.g., the shortest-path distance or Laplacian eigenspace projection, then because there are only finitely many unweighted graphs on n vertices, then the image of the PE in $\mathbb{R}^d$ will be finite.

B.2. Proof of Theorem 4.2 and Corollary 4.3

Euclidean Distance and Gram Matrices. Let $P = \{p_1, \dots, p_n\} \subset \mathbb{R}^d$ be a point cloud. The ***Gram matrix*** of P is denoted $G_P \in \mathbb{R}^{n\times n}$ where $(G_P)_{ij} = p_i^T p_j$. The ***distance matrix*** of P is denoted $D_P \in \mathbb{R}^{n\times n}$ where $(D_P)_{ij} = \|p_i - p_j\|$. A point cloud P is ***centered*** if $\sum_{i=1}^n p_i = 0$.

Theorem B.3. *Assume that for any graph G, there is an embedding of the nodes $P_G : V \to \mathbb{R}^d$ such that $\operatorname{im} P_G$ is centered. Let ψ_d and ψ_g be RPEs that are the distance and Gram matrix of the point set $\operatorname{im} P_G$, respectively. Then, ψ_d-WL is at least as strong as are as ψ_g-WL. Conversely, ψ_g-WL with diagonal augmentation is at least as strong as ψ_d.*

The proof is based on the following key lemma. We define the ***squared distance matrix*** of P to be the matrix $D_P^2 \in \mathbb{R}^{n\times n}$ where $(D_P^2)_{ij} = \|p_i - p_j\|^2$.

Lemma B.4. *Let ψ_d and ψ_g be RPEs such that, for any graph G, $\psi_{d,G}$ and $\psi_{g,G}$ are the squared distance and Gram matrix of a centered point set $P_G \subset \mathbb{R}^d$. Then ψ_d-WL is at least as strong as are as ψ_g-WL. Conversely, ψ_g-WL with diagonal augmentation is at least as strong as ψ_d.*

Proof of Theorem B.3 assuming Lemma B.4. As the square root function is injective, then by Proposition B.1, G and H are indistinguishable by ψ_d-WL if and only if D_P^2 and D_Q^2 are indistinguishable by ψ_g-WL. □

Now we proceed to prove Lemma B.4.

Lemma B.5. *Let $P = \{p_1, \dots, p_n\} \subset \mathbb{R}^d$ be a centered point cloud. For any $1 \le k \le n$, $(G_P)_{ii} = \frac{1}{n}\sum_{i=1}^n (D_P^2)_{ik} - \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n (D_P^2)_{ij}$.*

Proof. Observe that $(D_p^2)_{ij} = p_i^T p_i + p_j^T p_j - 2p_i^T p_j$. Therefore, the first term simplifies to

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (D_P^2)_{ik} =& p_k^T p_k + \frac{1}{n}\sum_{i=1}^n p_i^T p_i - \frac{2}{n}\sum_{i=1}^n p_i^T p_k \\ =& p_k^T p_k + \frac{1}{n}\sum_{i=1}^n p_i^T p_i. \qquad \text{(as } \textstyle\sum_{i=1}^n p_i = 0\text{)}\end{aligned}$$

The second term simplifies to:

$$\begin{aligned}\frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n (D_P^2)_{ij} =& \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n \left(p_i^T p_i + p_j^T p_j - 2p_i^T p_j\right) \\ =& \frac{1}{2n^2}\sum_{i=1}^n \left(np_i^T p_i + \sum_{j=1}^n p_j^T p_j - \sum_{j=1}^n 2p_i^T p_j\right) \\ =& \frac{1}{2n^2}\sum_{i=1}^n \left(np_i^T p_i + \sum_{j=1}^n p_j^T p_j\right) \qquad \text{(as } \textstyle\sum_{i=1}^n p_i = 0\text{)} \\ =& \frac{1}{2n^2}\sum_{i=1}^n 2np_i^T p_i \\ =& \frac{1}{n}\sum_{i=1}^n p_i^T p_i.\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (D_P^2)_{ik} - \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n (D_P^2)_{ij} =& p_k^T p_k + \frac{1}{n}\sum_{i=1}^n p_i^T p_i - \frac{1}{n}\sum_{i=1}^n p_i^T p_i \\ =& p_k^T p_k \\ =& (G_P)_{kk}.\end{aligned}$$

□

Lemma B.6. *Let ψ_d and ψ_g be RPEs such that, for any graph G, $\psi_{d,G}$ and $\psi_{g,G}$ are the squared distance and Gram matrix of a centered point set $P_G \subset \mathbb{R}^d$. Let (G, X_G) and (H, X_H) be featured graphs. If $\chi^{(2)}_{\psi_{d,G}}(u) = \chi^{(2)}_{\psi_{d,H}}(x)$, then $(G_{P_G})_{uu} = (G_{P_H})_{xx}$.*

Proof. We will prove this using Lemma B.5. First, the fact that $\chi^{(1)}_{\psi_{d,G}}(u) = \chi^{(1)}_{\psi_{d,H}}(x)$ implies $\{\!\{(D^2_{P_G})_{uv} : v \in V_G\}\!\} = \{\!\{(D^2_{P_H})_{xy} : y \in V_H\}\!\}$, which in particular implies that $\frac{1}{n}\sum_{v\in V_G}(D^2_{P_G})_{uv} = \frac{1}{n}\sum_{y\in V_H}(D^2_{P_H})_{xy}$. Second, as $\chi^{(2)}_{\psi_{d,G}}(u) = \chi^{(2)}_{\psi_{d,H}}(x)$, then as the second coordinate of these colors are equal, there is a bijection $\sigma : V_G \to V_H$ such that $\chi^{(1)}_{\psi_{d,G}}(v) = \chi^{(1)}_{\psi_{d,H}}(\sigma(v))$ for all $v \in V_G$. In particular, this implies that $\{\!\{(D^2_{P_G})_{vw} : w \in V_G\}\!\} = \{\!\{(D^2_{P_H})_{\sigma(v)z} : x \in V_H\}\!\}$ for all $v \in V_G$. Therefore, P_G and P_H have the same multiset of distances, and $\frac{1}{2n^2}\sum_{v\in V_G}\sum_{w\in V_G}(D^2_{P_G})_{vw} = \frac{1}{2n^2}\sum_{y\in V_H}\sum_{z\in V_H}(D^2_{P_H})_{yz}$. These two observations and Lemma B.5 imply the lemma. □

Proposition B.7 (Forward Direction of Lemma B.4)**.** *Let ψ_d and ψ_g be RPEs such that, for any graph G, $\psi_{d,G}$ and $\psi_{g,G}$ are the squared distance and Gram matrix of a centered point set $P_G \subset \mathbb{R}^d$. Let (G, X_G) and (H, X_H) be featured graphs. Let $u \in V_G$ and $x \in V_H$. If $\chi^{(t+2)}_{\psi_{d,G}}(u) = \chi^{(t+2)}_{\psi_{d,H}}(x)$, then $\chi^{(t)}_{\psi_{g,G}}(u) = \chi^{(t)}_{\psi_{g,H}}(x)$. Moreover, if (G, X_G) and (H, X_H) are indistinguishable by ψ_d-WL, then (G, X_G) and (H, X_H) are indistinguishable by ψ_g-WL.*

Proof. We prove this by induction on t. For $t = 0$, this is obvious as $\chi^{(2)}_{\psi_{d,G}}(u) = \chi^{(2)}_{\psi_{d,H}}(x)$ implies that $\chi^{(0)}_{\psi_{g,G}}(u) = X_G(u) = X_H(v) = \chi^{(0)}_{\psi_{g,H}}(x)$

Now suppose the proposition is true for a natural number t, and let $u \in V_G$ and $x \in V_H$ such that $\chi^{(t+3)}_{\psi_{d,G}}(u) = \chi^{(t+3)}_{\psi_{d,H}}(x)$. By induction, we know that $\chi^{(t+2)}_{\psi_{d,G}}(u) = \chi^{(t+2)}_{\psi_{d,H}}(x)$ implies $\chi^{(t)}_{\psi_{g,G}}(u) = \chi^{(t)}_{\psi_{g,H}}(x)$, so the first coordinates of $\chi^{(t+1)}_{\psi_{g,G}}(u)$ and $\chi^{(t+1)}_{\psi_{g,H}}(x)$ are equal.

We now show the second coordinates are equal. The fact that $\chi^{(t+3)}_{\psi_{d,G}}(u) = \chi^{(t+3)}_{\psi_{d,H}}(x)$ implies there is a bijection $\sigma : V_G \to V_H$ such that $(\chi^{(t+2)}_{\psi_{d,G}}(v),\, (D^2_{P_G})_{uv}) = (\chi^{(t+2)}_{\psi_{d,H}}(\sigma(v)),\, (D^2_{P_H})_{x\sigma(v)})$ for all $v \in V_G$. As $\chi^{(t+2)}_{\psi_{d,G}}(v) = \chi^{(t+2)}_{\psi_{d,H}}(\sigma(v))$ implies $(G_{P_G})_{vv} = (G_{P_H})_{\sigma(v)\sigma(v)}$ by Lemma B.6, then we know that

$$(G_{P_G})_{uv} = \frac{1}{2}\left((G_{P_G})_{vv} + (G_{P_G})_{uu} - (D^2_{P_G})_{uv}\right) = \frac{1}{2}\left((G_{P_H})_{\sigma(v)\sigma(v)} + (G_{P_H})_{xx} - (D^2_{P_H})_{x\sigma(v)}\right) = (G_{P_Q})_{x\sigma(v)}.$$

Moreover, $\chi^{(t+2)}_{\psi_{d,G}}(v) = \chi^{(t+2)}_{\psi_{d,H}}(\sigma(v))$ implies $\chi^{(t)}_{\psi_{g,G}}(v) = \chi^{(t)}_{\psi_{g,H}}(\sigma(v))$ by the inductive hypothesis. Therefore, as σ is a bijection, the second coordinates of $\chi^{(t+1)}_{\psi_{g,G}}(u)$ and $\chi^{(t+1)}_{\psi_{g,H}}(x)$ are equal:

$$\{\!\{(\chi^{(t)}_{\psi_{g,G}}(v),\, (G_{P_G})_{uv}) : v \in V_G\}\!\} = \{\!\{(\chi^{(t)}_{\psi_{g,H}}(y),\, (G_{P_H})_{xy}) : y \in V_H\}\!\}.$$ □

Proposition B.8 (Backward Direction of Lemma B.4)**.** *Let ψ_d and ψ_g be RPEs such that, for any graph G, $\psi_{d,G}$ and $\psi_{g,G}$ are the squared distance and Gram matrix of a centered point set $P_G \subset \mathbb{R}^d$. Let $\psi'_g = (\psi_g, I)$ be the diagonal augmentation of ψ_g. Let (G, X_G) and (H, X_H) with featured graphs. Let $u \in V_G$ and $x \in V_H$. If $\chi^{(t+1)}_{\psi'_{g,G}}(u) = \chi^{(t+1)}_{\psi'_{g,H}}(x)$, then $\chi^{(t)}_{\psi_{d,G}}(u) = \chi^{(t)}_{\psi_{d,H}}(x)$. In particular, ψ'_g-WL with diagonal augmentation is at least as strong as ψ_d-WL.*

Proof. We prove this by induction on t. For $t = 0$, this is obvious as $\chi^{(1)}_{\psi'_{g,G}}(u) = \chi^{(1)}_{\psi'_{g,H}}(x)$ implies $\chi^{(0)}_{\psi'_{g,G}}(u) = X_G(u) = X_H(v) = \chi^{(0)}_{\psi'_{g,H}}(x)$ and equivalently $\chi^{(0)}_{\psi_{d,G}}(u) = X_G(u) = X_H(v) = \chi^{(0)}_{\psi_{d,H}}(x)$.

Now suppose the proposition is true for a natural number t, and suppose that $\chi^{(t+2)}_{\psi'_{g,G}}(u) = \chi^{(t+2)}_{\psi'_{g,H}}(x)$. By induction, we know that $\chi^{(t+1)}_{\psi'_{g,G}}(u) = \chi^{(t+1)}_{\psi'_{g,H}}(x)$ implies $\chi^{(t)}_{\psi_{d,G}}(u) = \chi^{(t)}_{\psi_{d,H}}(x)$, so the first coordinates of $\chi^{(t+1)}_{\psi_{d,G}}(u)$ and $\chi^{(t+1)}_{\psi_{d,H}}(x)$ are equal.

We now show the second coordinates are equal. As $\chi^{(t+2)}_{\psi'_{g,G}}(u) = \chi^{(t+2)}_{\psi'_{g,H}}(x)$, then there is a bijection $\sigma : V_G \to V_H$ such that $(\chi^{(t+1)}_{\psi'_{g,G}}(v), (G_{P_G}, I)_{uv}) = (\chi^{(t+1)}_{\psi'_{g,H}}(\sigma(v)), (G_{P_G}, I)_{x\sigma(v)})$ for all $v \in V_G$. The diagonal augmentation implies that

$(G_{P_G}, I)_{uu} = (G_{P_H}, I)_{xx}$ as a diagonal element of (G_{P_G}, I) can only equal a diagonal element of (G_{P_H}, I). By a similar argument, the fact that $\chi^{(t+1)}_{\psi'_{g,G}}(v)) = \chi^{(t+1)}_{\psi'_{g,H}}(\sigma(v))$ implies that $(G_{P_G}, I)_{vv} = (G_{P_H}, I)_{\sigma(v)\sigma(v)}$. Therefore,

$$\begin{aligned}(D^2_{P_G})_{uv} =& (G_{P_G})_{uu} + (G_{P_G})_{vv} - 2(G_{P_G})_{uv} \\ =& (G_{P_H})_{xx} + (G_{P_H})_{\sigma(v)\sigma(v)} - 2(G_{P_H})_{x\sigma(v)} = (D^2_{P_H})_{x\sigma(v)}.\end{aligned}$$

Therefore, the second coordinates of $\chi^{(t+1)}_{\psi_{d,G}}(u)$ and $\chi^{(t+1)}_{\psi_{d,H}}(x)$ are equal, i.e.,

$$\{\!\{(\chi^{(t)}_{\psi_{d,G}}(v),\, (D^2_{P_G})_{uv}) : v \in V_G\}\!\} = \{\!\{(\chi^{(t)}_{\psi_{d,H}}(v),\, (D^2_{P_H})_{uv}) : v \in V_H\}\!\}. \qquad \square$$

Now we can prove Theorem 4.2 below:

Theorem 4.2. *Let $f : \mathbb{R}^+ \to \mathbb{R}^+$. D^f-WL is at least as strong as K^f-WL. K^f-WL with diagonal augmentation is at least as strong as D^f-WL.*

Proof of Theorem 4.2. Note that for given graph G with n points, any spectral decomposition of the Laplacian $L = \sum_{i=2}^n \lambda_i z_i z_i^T$, and any function $f : \mathbb{R}^+ \to \mathbb{R}^+$, we can define the point cloud $P_G = \{\sqrt{f(L)}1_v : v \in V\}$, where $\sqrt{f(L)} = \sum_{i=2}^n \sqrt{f(\lambda_i)} z_i z_i^T$ The spectral kernel K^f_G is the Gram matrix of this point cloud and the spectral distance D^f_G is the distance matrix of this point cloud. (While different choice of bases for the eigenspace result in different point clouds P_G, these point clouds will always have the same Gram and distance matrices regardless of choice of basis.) This point cloud is centered as $\sum_{v\in V} \sqrt{f(L)}1_v = \sum_{v\in V}\sum_{i=2}^n \sqrt{f(L)}\lambda_i z_i z_i^T 1_v = 0$ by the fact that each eigenvector z_i is orthogonal to the all-1s vector, which is the eigenvector of L corresponding to $\lambda_1 = 0$. The result for RPE-augWL with RPEs K^f_G and D^f_G holds by directly applying Theorem B.3. $\square$

RD-WL and $L^\dagger$-WL In this section, we prove Corollary 4.3 by showing that it is unnecessary to perform diagonal augmentation on $L^\dagger$ in order for $L^\dagger$-WL to be as strong as RD-WL. In particular, we show that vertices u and v with the same $L^\dagger$-WL color after one iteration have the same diagonal entry $L^\dagger(u,u) = L^\dagger(v,v)$. This is the point in the proof of Proposition B.8 where diagonal augmentation is used, so because this holds for $L^\dagger$-WL, we can drop the diagonal augmentation.

For background, the ***pseudoinverse*** of the Laplacian is $L^\dagger = \sum_{i=2}^n \frac{1}{\lambda_i} z_i z_i^T$, and the ***resistance distance*** between nodes u and v in a graph is $RD(u,v) = (1_u - 1_v)^T L^\dagger (1_u - 1_v)$, where 1_u is the vector whose u-th entry is 1 and other entries are 0. The resistance distance is a **squared** spectral distance, but remarkably, it is also a metric (see for example (Spielman, 2019, Section 12.8)), which is not true in general for the square of distances.

In particular, we prove the following lemma:

Lemma B.9. *Let G and H be graphs. Let $u \in V_G$ and $v \in V_H$ such that $\chi^{(1)}_{L^\dagger_G}(u) = \chi^{(1)}_{L^\dagger_H}(v)$. Then $L^\dagger_G(u,u) = L^\dagger_H(v,v)$.*

The proof is based on several auxiliary results.

Lemma B.10. *Let G be a connected graph with more than 1 vertex and let L denote its graph Laplacian. Then, for any $u \neq v \in V$, we have that $L^\dagger(v,v) > L^\dagger(u,v)$.*

Proof. By the fact that $RD(u,v) = L^\dagger(u,u) + L^\dagger(v,v) - 2L^\dagger(u,v)$, we have that

$$L^\dagger(v,v) - L^\dagger(u,v) = \frac{1}{2}(RD(u,v) + L^\dagger(v,v) - L^\dagger(u,u))$$

Therefore, to prove that $L^\dagger(v,v) > L^+_{uv}$, it suffices to prove that

$$RD(u,v) > L^\dagger(u,u) - L^\dagger(v,v).$$

By Lemma B.5, we have that

$$\begin{aligned}
L^\dagger(u,u) - L^\dagger(v,v) &= \frac{1}{n}\left(\sum_{w\in V} RD(u,w) - \sum_{w\in V} RD(v,w)\right) \\
&= \frac{1}{n}\sum_{w\in V} RD(u,w) - RD(v,w) \\
&= \frac{1}{n}\sum_{w\in V\setminus\{u,v\}} RD(u,w) - RD(v,w) && \text{(as } RD(v,u)=0\text{)} \\
&\leq \frac{1}{n}\sum_{w\in V\setminus\{u,v\}} RD(u,v) && \text{(Triangle inequality.)} \\
&= \frac{n-2}{n} RD(u,v) < RD(u,v). && \square
\end{aligned}$$

Lemma B.11. *Let G be a graph and let L denote its graph Laplacian. Let $u \in V$. Then either*

1. $L^\dagger(u,u) > L^\dagger(u,v)$ *for all* $v \in V$*, or*
2. $L^\dagger(u,v) = 0$ *for all* $v \in V$*.*

Proof. Let $G_1, \cdots, G_k$ denote the connected components of G. Then, we have that $L = L_1 \oplus \cdots \oplus L_k$ and thus $L^\dagger = L_1^\dagger \oplus \cdots \oplus L_k^\dagger$. For each connected component G_l, if it contains more than 1 vertex, then by Lemma B.10, $L_l^\dagger$ satisfies the condition that the diagonal entries are strictly larger than all other entries in the corresponding rows. If G_l contains only 1 vertex, then $L_l^\dagger = 0$ and hence the corresponding row and column in $L^\dagger$ is always 0. This concludes the proof. □

Now, we finish the proof of Lemma B.9.

Proof of Lemma B.9. By the definition of the $L^\dagger$-WL test, $\chi^{(1)}_{L_G^\dagger}(u) = \chi^{(1)}_{L_H^\dagger}(v)$ implies that $\{\!\{L_G^\dagger(u,w) : w \in V_G\}\!\} = \{\!\{L_H^\dagger(v,x) : x \in V_H\}\!\}$. By Lemma B.11, we know that either the multisets $\{\!\{L_G^\dagger(u,w) : w \in V_G\}\!\} = \{\!\{L_H^\dagger(v,x) : x \in V_H\}\!\}$ are all 0s, in which case $L_G^\dagger(u,u) = 0 = L_H^\dagger(v,v)$, or they have the same maximal element, which is $L_G^\dagger(u,u) = L_H^\dagger(v,v)$. □

B.3. SPE APE-GT is Stronger than RD-RPE-GT

Background. First, we provide the definition of SPE (Huang et al., 2024). Denote the eigenvalues and eigenvectors of the Laplacian as $\lambda_1 \leq \ldots \leq \lambda_{|V|}$ and $z_1, \ldots, z_{|V|}$. Let $\Lambda := [\lambda_1, \ldots, \lambda_{|V|}] \in \mathbb{R}^{|V|}$ and $V := [z_1 | \cdots | z_{|V|}] \in \mathbb{R}^{|V|\times|V|}$, so $L = V\,\mathrm{diag}(\Lambda)V^T$.

Stable and expressive positional encoding (SPE) is the APE

$$\mathrm{SPE}(V,\Lambda) = g(V\,\mathrm{diag}(f(\Lambda))V^T),$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a permutation-equivariant function like an elementwise MLP or DeepSets (Zaheer et al., 2017) and $g : \mathbb{R}^{n\times n} \to \mathbb{R}^n$ is any graph neural network. In our context, we assume that f is a DeepSets network and g is a 2-EGN.

By removing the final layer of 2-EGN from SPE, one obtains a natural RPE that we call the ***relative-SPE (RSPE)***: $\mathrm{RSPE}(V,\Lambda) = V\,\mathrm{diag}(f(\Lambda))V^T$.

Lemma B.12. *If a chosen RSPE is diagonally-aware (or diagonally augmented), then there exists a 2-EGN g such that RSPE-RPE-GT is equivalent to SPE-APE-GT in terms of distinguishing power.*

Proof. This follows from Theorem 3.10. □

Notice that when f is an elementwise function, RSPE reduces to a spectral kernel and we hence have the following result; however, RSPE is more general than spectral kernels as $f(\Lambda)_i$ may be a function of all the eigenvalues, not just the ith eigenvalue. Accordingly, we have the following lemma.

Lemma B.13. *RSPE can compute any spectral kernel.*

Based on the lemma, we are able to prove our main result in this section.

Theorem 4.4. *For unfeatured graphs, diagonally augmented SPE-APE-GT is at least as strong as RD-RPE-GT.*

Proof. The precise statement of the result is that there exists a choice of RSPE with the diagonal augmentation such that its corresponding SPE-APE-GT is at least as strong as RD-RPE-GT.

By Lemma B.12, we know that there exists a 2-EGN g such that RSPE-RPE-GT is equivalent to SPE-APE-GT in terms of distinguishing power. Therefore, it suffices to show that RSPE-RPE-GT is at least as strong as RD-RPE-GT. Since the pseudoinverse of the Laplacian $L^\dagger$ is a spectral kernel, by Lemma B.13 RSPE can be chosen so that its corresponding RSPE-WL is at least as strong as $L^\dagger$-WL. Since the extra diagonal augmentation of the two RPEs will not change the relative distinguishing power of their RPE-augWL test, by Lemma 3.6 and the fact that RD is diagonally-aware, we conclude the proof. □

B.4. Powers of Matrices and Spectral Kernels: Proofs from Section 4.3

We want to show that the multidimensional RPE $(I, L, \ldots, L^k)$ is equivalent to spectral kernels in distinguishing power. To prove this, we will use a basic result from the interpolation literature showing that any function on k points can be fit exactly by a $(k-1)$-degree polynomial p.

Lemma B.14. *Let $f : \mathbb{R} \to \mathbb{R}$ be any function and let $\{x_1, \ldots, x_n\} \subset \mathbb{R}$. Then there exists an $(n-1)$-degree polynomial $p : \mathbb{R} \to \mathbb{R}$ such that $p(x_i) = f(x_i)$ for all $1 \leq i \leq n$.*

Theorem 4.6. *$(I, L, \ldots, L^{2n-1})$-WL is at least as strong as K^f-WL on graphs with at most n nodes.*

Proof. Let G and H be graphs such that K^f-WL distinguishes G and H. Let $\{\lambda_{G,1}, \ldots, \lambda_{G,n}\}$ and $\{\lambda_{H,1}, \ldots, \lambda_{H,n}\}$ be the spectra of the Laplacians L_G and L_H respectively, and denote their union $\{\lambda_1^G, \ldots, \lambda_n^G\} \cup \{\lambda_1^G, \ldots, \lambda_n^G\} = \{\lambda_0, \ldots, \lambda_{2n-1}\}$ (without loss of generality, we assume these elements are distinct). Let $p(x) = \sum_{i=0}^{2n-1} c_i x^i$ be the polynomial guaranteed by Lemma B.14 for the point cloud $\{\lambda_0, \ldots, \lambda_{2n-1}\}$ and for the function f in defining K^f. Let $p_l : \mathbb{R}^{2n} \to \mathbb{R}$ be the linear function with the same coefficients as p, i.e. $p_l(x) = \sum_{i=0}^{2n-1} c_i x_i$. By linearity, if we apply p_l elementwise on the tensor $(I, L_G, \ldots, L_G^{2n-1})$, we find that it equals K^f for G:

$$p_l(I, L_G, \ldots, L_G^{2n-1}) = \sum_{i=0}^{2n-1} c_i L_G^i = \sum_{i=0}^{2n-1} \sum_{j=2}^{n} c_i \lambda_{G,j}^i z_j z_j^T = \sum_{j=2}^{n} p(\lambda_{G,j}) z_j z_j^T = \sum_{j=2}^{n} f(\lambda_{G,j}) z_j z_j^T = K_G^f$$

The same calculation holds for H. Therefore, as K^f-WL distinguishes G and H, then $p_l \circ (I, L, \ldots, L^{2n-1})$-WL distinguishes G and H.

Moreover, Proposition B.1 implies $(I, L, \ldots, L^{2n-1})$-WL distinguishes G and H. As the above argument can be repeated (using different polynomials p) to show that $(I, L, \ldots, L^{2n-1})$ can distinguish any pair of graphs that K^f can, then $(I, L, \ldots, L^{2n-1})$-WL is as strong as K^f-WL on graphs with n nodes. □

The following theorem uses normalized versions of the adjacency matrix and Laplacian. Recall that D denotes the degree matrix, A denotes the adjacency matrix, and L denotes the graph Laplacian. The symmetrically-normalized adjacency matrix is $\widehat{A} = D^{-1/2} A D^{-1/2}$, and its symmetrically-normalized Laplacian is $\widehat{L} = D^{-1/2} A D^{-1/2} = I - \widehat{A}$. We let $\mu_1 \leq \cdots \leq \mu_n$ denote eigenvalues of $\widehat{A}$ and let $\nu_1 \leq \cdots \leq \nu_n$ denote eigenvalues of $\widehat{L}$

Note that $\nu_i = 1 - \mu_i$. Moreover, $\widehat{A}$ and $\widehat{L}$ they have the same set of eigenvectors $\{x_1, \ldots, x_n\}$ (up to choice of orthonormal basis). For a function $f : \mathbb{R}^+ \to \mathbb{R}^+$, we define the ***normalized spectral kernel*** as $\widehat{K}^f = \sum_{i=2}^{n} f(\nu_i) x_i x_i^T$.

Theorem 4.8. *$(I, \widehat{A}, \ldots, \widehat{A}^{2n-1})$-WL is at least as strong as $\widehat{K}^f$-WL on graphs with at most n nodes.*

Proof. This theorem follows as the eigenvalues μ_i of $\widehat{A}$ and ν_i of $\widehat{L}$ can be matched as $\mu_i = 1 - \nu_i$. The rest of the proof is the same as Theorem 4.6 by using polynomials to fit the function $\widehat{f}(x) = f(1-x)$. □

Theorem 4.9. *$(I, H^{(1)}, \ldots, H^{(2n-1)})$-WL is at least as strong as K^f-WL on graphs with at most n nodes.*

Proof. This theorem follows as the eigenvalues of $H^{(t)}$ are $e^{-t\lambda_i} = (e^{-\lambda_i})^t$, so the proof is the same as the proof of Theorem 4.6 by using polynomials to fit the function $\widehat{f}(x) = f(-\log(x))$. □

B.5. Common Matrices: Proof of Proposition 4.10

Proposition 4.10. *$A, \widehat{A}, \widetilde{A}, L, \widehat{L}$ and $\widetilde{L}$ induce equivalent RPE-augWL tests. In particular, all of these RPE-augWL tests are equally strong as the WL test.*

Proof. By Proposition B.16 below, we know that A-WL is equivalent to the WL test. Hence, we only need to show that all other RPE-augWL tests are equivalent to A-WL.

Choose any featured graphs (G, X_G) and (H, X_H). Pick $u \in V_G$ and $v \in V_H$. First of all, we observe the following fact.

Fact 4. $\chi^{(1)}_{A_G}(u) = \chi^{(1)}_{A_H}(v)$ *simplies that* $\deg_G(u) = \deg_H(v)$.

$\widetilde{A}$: We inductively prove that $\chi^{(t)}_{A_G}(u) = \chi^{(t)}_{A_H}(v)$ iff $\chi^{(t)}_{\widetilde{A}_G}(u) = \chi^{(t)}_{\widetilde{A}_H}(v)$. The case when $t = 0$ trivially holds. Assume that the statement holds for some $t \geq 0$. Then, we first assume that $\chi^{(t+1)}_{A_G}(u) = \chi^{(t+1)}_{A_H}(v)$. This implies that $\chi^{(1)}_{A_G}(u) = \chi^{(1)}_{A_H}(v)$. By Fact 4, we have that $\deg_G(u) = \deg_H(v)$. By induction, we have that $\chi^{(t)}_{\widetilde{A}_G}(u) = \chi^{(t)}_{\widetilde{A}_H}(v)$ and

$$\{\!\{(\chi^{(t)}_{\widetilde{A}_G}(x), \widetilde{A}_G(u,x)) : x \in V_G\}\!\} = \{\!\{(\chi^{(t)}_{\widetilde{A}_H}(y), \widetilde{A}_H(v,y)) : y \in V_H\}\!\}.$$

Hence, we have that

$$\{\!\{(\chi^{(t)}_{\widetilde{A}_G}(x), A_G(u,x)/\deg_G(u)) : x \in V_G\}\!\} = \{\!\{(\chi^{(t)}_{\widetilde{A}_H}(y), A_H(v,y)/\deg_H(v)) : y \in V_H\}\!\},$$

and thus $\chi^{(t+1)}_{\widetilde{A}_G}(u) = \chi^{(t+1)}_{\widetilde{A}_H}(v)$.

Conversely, assume that $\chi^{(t+1)}_{\widetilde{A}_G}(u) = \chi^{(t+1)}_{\widetilde{A}_H}(v)$. This implies that $\chi^{(1)}_{\widetilde{A}_G}(u) = \chi^{(1)}_{\widetilde{A}_H}(v)$. Similarly to Fact 4, we have that $\deg_G(u) = \deg_H(v)$. Then, the above argument can be repeated to show that $\chi^{(t+1)}_{A_G}(u) = \chi^{(t+1)}_{A_H}(v)$.

Therefore, $\widetilde{A}$-WL is equivalent to A-WL.

$\widehat{A}$: We prove that $\widehat{A}$-WL is equivalent to $\widetilde{A}$-WL and hence A-WL by showing that for any $u \in V_G$ and $v \in V_H$, one has for all $t \geq 0$ that

$$\chi^{(t)}_{\widehat{A}_G}(u) = \chi^{(t)}_{\widehat{A}_H}(v) \text{ implies } \chi^{(t)}_{\widetilde{A}_G}(u) = \chi^{(t)}_{\widetilde{A}_H}(v) \text{ and } \chi^{(t+1)}_{\widetilde{A}_G}(u) = \chi^{(t+1)}_{\widetilde{A}_H}(v) \text{ implies } \chi^{(t)}_{\widehat{A}_G}(u) = \chi^{(t)}_{\widehat{A}_H}(v)$$

The case when $t = 0$ holds trivially.

Now assume that the statement holds for any $t \geq 1$. $\chi^{(t+1)}_{\widehat{A}_G}(u) = \chi^{(t+1)}_{\widehat{A}_H}(v)$ implies that

$$\begin{aligned}&(\chi^{(t)}_{\widehat{A}_G}(u), \{\!\{(\chi^{(t)}_{\widehat{A}_G}(x), A_G(u,x)/\sqrt{\deg_G(u)\deg_G(x)}) : x \in V_G\}\!\})\\ &= (\chi^{(t)}_{\widehat{A}_H}(v), \{\!\{(\chi^{(t)}_{\widehat{A}_H}(y), A_H(v,y)/\sqrt{\deg_H(v)\deg_H(y)}) : y \in V_H\}\!\}).\end{aligned}$$

By counting the number of non-zero elements on both sides, we have that $\deg_G(u) = \deg_H(v)$. Hence,

$$\begin{aligned}&(\chi^{(t)}_{\widehat{A}_G}(u), \{\!\{(\chi^{(t)}_{\widehat{A}_G}(x), \deg_G(u), A_G(u,x)/\sqrt{\deg_G(x)}) : x \in V_G\}\!\})\\ &= (\chi^{(t)}_{\widehat{A}_H}(v), \{\!\{(\chi^{(t)}_{\widehat{A}_H}(y), \deg_H(v), A_H(v,y)/\sqrt{\deg_H(y)}) : y \in V_H\}\!\})\end{aligned}$$

and thus

$$\begin{aligned}&(\chi^{(t)}_{\widehat{A}_G}(u), \{\!\{(\chi^{(t)}_{\widehat{A}_G}(x), \deg_G(u), A_G(u,x)) : x \in V_G\}\!\})\\ &= (\chi^{(t)}_{\widehat{A}_H}(v), \{\!\{(\chi^{(t)}_{\widehat{A}_H}(y), \deg_H(v), A_H(v,y)) : y \in V_H\}\!\}).\end{aligned}$$

Finally, by the inductive hypothesis we have that

$$\begin{aligned}\chi^{(t+1)}_{\widetilde{A}_G}(u) &= (\chi^{(t)}_{\widetilde{A}_G}(u), \{\!\{(\chi^{(t)}_{\widetilde{A}_G}(x), \widetilde{A}_G(u,x)) : x \in V_G\}\!\}) \\ &= (\chi^{(t)}_{\widetilde{A}_H}(v), \{\!\{(\chi^{(t)}_{\widetilde{A}_H}(y), \widetilde{A}_H(v,y)) : y \in V_H\}\!\}) = \chi^{(t+1)}_{\widetilde{A}_H}(v).\end{aligned}$$

For the other direction, assume that $\chi^{(t+2)}_{\widetilde{A}_G}(u) = \chi^{(t+2)}_{\widetilde{A}_H}(v)$. This implies that

$$\begin{aligned}&(\chi^{(t+1)}_{\widetilde{A}_G}(u), \{\!\{(\chi^{(t+1)}_{\widetilde{A}_G}(x), \widetilde{A}_G(u,x)) : x \in V_G\}\!\}) \\ &= (\chi^{(t+1)}_{\widetilde{A}_H}(v), \{\!\{(\chi^{(t+1)}_{\widetilde{A}_H}(y), \widetilde{A}_H(v,y)) : y \in V_H\}\!\}).\end{aligned}$$

It is obvious that $\deg_G(u) = \deg_H(v)$. Let $\sigma : V_G \to V_H$ denote a bijection such that $(\chi^{(t+1)}_{\widetilde{A}_G}(x), \widetilde{A}_G(u,x)) = (\chi^{(t+1)}_{\widetilde{A}_H}(\sigma(x)), \widetilde{A}_H(v,\sigma(x)))$ for all $x \in V_G$. Then, we have that $\deg_G(x) = \deg_H(\sigma(x))$. Hence, $(\chi^{(t+1)}_{\widehat{A}_G}(x), \widehat{A}_G(u,x)) = (\chi^{(t+1)}_{\widehat{A}_H}(\sigma(x)), \widehat{A}_H(v,\sigma(x)))$. By the inductive hypothesis, we have that $(\chi^{(t)}_{\widehat{A}_G}(x), \widehat{A}_G(u,x)) = (\chi^{(t)}_{\widehat{A}_H}(\sigma(x)), \widehat{A}_H(v,\sigma(x)))$. Therefore,

$$\begin{aligned}\chi^{(t+1)}_{\widehat{A}_G}(u) &= (\chi^{(t)}_{\widehat{A}_G}(u), \{\!\{(\chi^{(t)}_{\widehat{A}_G}(x), \widehat{A}_G(u,x)) : x \in V_G\}\!\}) \\ &= (\chi^{(t)}_{\widehat{A}_H}(v), \{\!\{(\chi^{(t)}_{\widehat{A}_H}(y), \widehat{A}_H(v,y)) : y \in V_H\}\!\}) = \chi^{(t+1)}_{\widehat{A}_H}(v).\end{aligned}$$

$L, \widehat{L}, \widetilde{L}$: Note that the graph Laplacian L is combinatorially-aware. Hence by Theorem 4.16, L-WL is at least as strong as the WL test and hence A-WL test. For the other direction, we prove the stronger result that $\chi^{(t)}_{A_G}(u) = \chi^{(t)}_{A_H}(v)$ implies that $\chi^{(t)}_{L_G}(u) = \chi^{(t)}_{L_H}(v)$ for any $u \in V_G$ and $v \in V_H$.

The statement trivially holds when $t = 0$. Assume that the statement holds for some $t \geq 0$. Then, we assume that $\chi^{(t+1)}_{A_G}(u) = \chi^{(t+1)}_{A_H}(v)$. This implies that $\chi^{(1)}_{A_G}(u) = \chi^{(1)}_{A_H}(v)$. By Fact 4, we have that $\deg_G(u) = \deg_H(v)$. By definition, $\chi^{(t)}_{A_G}(u) = \chi^{(t)}_{A_H}(v)$ and hence by induction hypothesis, $\chi^{(t)}_{L_G}(u) = \chi^{(t)}_{L_H}(v)$. Now, from the fact that $\{\!\{(\chi^{(t)}_{A_G}(x), A_G(u,x)) : x \in V_G\}\!\} = \{\!\{(\chi^{(t)}_{A_H}(y), A_H(v,y)) : y \in V_H\}\!\}$, we obtain a bijection $\sigma : V_G \to V_H$ such that $(\chi^{(t)}_{A_G}(x), A_G(u,x)) = (\chi^{(t)}_{A_H}(\sigma(x)), A_H(v,\sigma(x)))$ for all $x \in V_G$. We assume that $\sigma(u) = v$ (otherwise we can modify σ so that the assumption holds). Then, by inductive hypothesis we have that $\chi^{(t)}_{L_G}(x) = \chi^{(t)}_{L_H}(\sigma(x))$ for all $x \in V_G$. Furthermore, from the fact that $A_G(u,x) = A_H(v,\sigma(x))$, $\sigma(u) = v$ and $\deg_G(u) = \deg_H(v)$, we have that $L_G(u,x) = L_H(v,\sigma(x))$ for all $x \in V_G$. Hence, $\{\!\{(\chi^{(t)}_{L_G}(x), L_G(u,x)) : x \in V_G\}\!\} = \{\!\{(\chi^{(t)}_{L_H}(y), L_H(v,y)) : y \in V_H\}\!\}$. Therefore, $\chi^{(t+1)}_{L_G}(u) = \chi^{(t+1)}_{L_H}(v)$. This concludes the proof.

Therefore, L-WL is equivalent to A-WL. Similarly, $\widehat{L}$-WL is equivalent to $\widehat{A}$-WL and $\widetilde{L}$-WL is equivalent to $\widetilde{A}$-WL. Hence, all of these RPE-augWL tests are equivalent to the WL test. □

B.6. Combinatorial-Awareness: Proof of Theorem 4.16

Theorem 4.16. *Let ψ be a combinatorially-aware RPE. Then, ψ-WL is at least as strong as WL.*

For any two featured graphs (G, X_G) and (H, X_H), we prove the following stronger lemma that immediately implies Theorem 4.16.

Lemma B.15. *Let ψ be a combinatorially-aware RPE. Let $u \in G$ and $v \in H$. If $\chi^{(t)}_{\psi_G}(u) = \chi^{(t)}_{\psi_H}(v)$, then $\chi^{(t)}_G(u) = \chi^{(t)}_H(v)$.*

Proof. We prove this by induction on the iteration. The case $t = 0$ holds trivially by definition of the two WL tests.

Now inductively suppose the theorem is true for iteration t. Let $u \in G$ and $v \in H$ such that $\chi^{(t+1)}_{\psi_G}(u) = \chi^{(t+1)}_{\psi_H}(v)$.

First, by definition, $\chi^{(t)}_{\psi_G}(u) = \chi^{(t)}_{\psi_H}(v)$. This and the inductive hypothesis imply $\chi^{(t)}_G(u) = \chi^{(t)}_H(v)$.

Next, by definition, there is a bijection $\sigma : V_G \to V_H$ such that $(\chi^{(t)}_{\psi_G}(w), \psi_G(u,w)) = (\chi^{(t)}_{\psi_H}(\sigma(w)), \psi_H(v, \sigma(w)))$ for all $w \in V_G$. This implies $\psi_G(u,w) = \psi_H(v, \sigma(w))$ if and only if both $\{u,w\}$ and $\{v, \sigma(w)\}$ are edges or non-edges; therefore, $\{\!\{\chi^{(t)}_{\psi_G}(w) : \{u,w\} \in E_G\}\!\} = \{\!\{\chi^{(t)}_{\psi_H}(x) : \{v,x\} \in E_H\}\!\}$. This and the inductive hypothesis imply $\{\!\{\chi^{(t)}_{G}(w) : \{u,w\} \in E_G\}\!\} = \{\!\{\chi^{(t)}_{H}(x) : \{v,x\} \in E_H\}\!\}$. This and the observation from the previous paragraph imply $\chi^{(t+1)}_G(u) = \chi^{(t+1)}_H(v)$. □

Note that the adjacency matrix A is trivially combinatorially-aware and hence A-WL is at least as strong as the WL. In fact, the other direction also holds and hence we have that the A-WL and WL are equally strong.

Proposition B.16 (RPE-augWL generalizes WL)**.** *Consider the RPE given by the adjacency matrix A. Given two graphs G and H, then $\chi(G) = \chi(H)$ iff $\chi_A(G) = \chi_A(H)$.*

Proof. We only need to prove the direction that the WL is at least as strong as the A-WL. Let (G, X_G) and (H, X_H) be two featured graphs. We prove the following stronger statement:

Lemma B.17. *Assume that $\{\!\{\chi^{(t)}_G(u) : u \in V_G\}\!\} = \{\!\{\chi^{(t)}_H(v) : v \in V_H\}\!\}$. Let $u \in G$ and $v \in H$. If $\chi^{(t)}_G(u) = \chi^{(t)}_H(v)$, then $\chi^{(t)}_{A_G}(u) = \chi^{(t)}_{A_H}(v)$ and hence $\{\!\{\chi^{(t)}_{A_G}(u) : u \in V_G\}\!\} = \{\!\{\chi^{(t)}_{A_H}(v) : v \in V_H\}\!\}$.*

Proof of Lemma B.17. We prove this by induction on the iteration. The case $t = 0$ holds trivially by definition of the two WL tests.

Now inductively suppose the theorem is true for iteration t. Let $u \in G$ and $v \in H$ such that $\chi^{(t+1)}_G(u) = \chi^{(t+1)}_H(v)$. By definition, $\chi^{(t)}_G(u) = \chi^{(t)}_H(v)$ and $\{\!\{\chi^{(t)}_G(w) : \{u,w\} \in E_G\}\!\} = \{\!\{\chi^{(t)}_H(x) : \{v,x\} \in E_H\}\!\}$. By induction assumption, we have that $\chi^{(t)}_G(u) = \chi^{(t)}_H(v)$. Furthermore, there is a bijection $\sigma : N_G(u) \to N_H(v)$ such that $\chi^{(t)}_G(w) = \chi^{(t)}_H(\sigma(w))$ for all $w \in N_G(u)$. Now, since $\{\!\{\chi^{(t)}_G(x) : x \in V_G\}\!\} = \{\!\{\chi^{(t)}_H(y) : y \in V_H\}\!\}$, we have that $\{\!\{\chi^{(t)}_G(x) : x \in V_G \backslash N_G(u)\}\!\} = \{\!\{\chi^{(t)}_H(y) : y \in V_H \backslash N_H(v)\}\!\}$. Hence there exists a bijection $\tau : V_G \backslash N_G(u) \to V_H \backslash N_H(v)$ such that $\chi^{(t)}_G(x) = \chi^{(t)}_H(\tau(x))$ for all $x \in V_G \backslash N_G(u)$. Now, we define a bijection $\Phi : V_G \to V_H$ as the union of σ and τ. In this way, we have that for any $x \in V_G$, $A_G(u,x) = A_H(v, \Phi(x))$ and $\chi^{(t)}_A(x) = \chi^{(t)}_A(\Phi(x))$. Hence, we have that

$$(\chi^{(t)}_{A_G}(u), \{\!\{(\chi^{(t)}_{A_G}(x), A_G(u,x)) : x \in V_G\}\!\}) = (\chi^{(t)}_{A_H}(v), \{\!\{(\chi^{(t)}_{A_H}(y), A_H(v,y)) : y \in V_H\}\!\}).$$

This implies that $\chi^{(t+1)}_{A_G}(u) = \chi^{(t+1)}_{A_H}(v)$ and we hence concludes the proof. □

This concludes the proof. □

B.7. Resistance Distance is not Combinatorially-Aware

In Figure 6, we show a graph G so that $\mathrm{RD}(1,3) = \mathrm{RD}(4,5) = 1$ but $\{4,5\} \in E_G$ and $\{1,3\} \notin E_G$. This proves that the resistance distance is not combinatorially-aware.

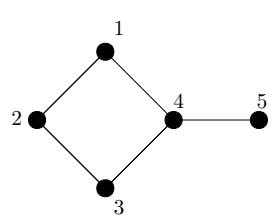


Figure 6. $\mathrm{RD}(1,3) = \mathrm{RD}(4,5) = 1$.

B.8. Resistance Distance and Block Cut-Edge Component Trees

For a connected graph G, a ***cut edge*** is an edge $e \in E_G$ such that $G \setminus \{e\}$ is disconnected. An ***(edge-)biconnected component*** of G is a maximal subset $U \subset V_G$ such that the subgraph of G induced by U is connected and has no cut edges. The ***block cut-edge tree*** of G is the tree whose vertices are the edge-biconnected components $BCC(G)$ of G and its edges are $\{\{C_1, C_2\} : C_1 \neq C_2 \in BCC(G), \exists \{u, v\} \in E_G$ such that $u \in C_1 \,\&\, v \in C_2\}$, i.e. the edges of the block cut-edge tree correspond to cut edges in G that connect two edge-biconnected components.

Similarly, a ***cut vertex*** is a vertex $v \in V_G$ such that $G \setminus \{v\}$ is disconnected.

In this section, we prove the following theorem.

Theorem B.18. *Let G and H be unfeatured graphs.*

1. *Let $\{u, v\} \in E_G$ and $\{x, y\} \in E_H$. If $\{\!\{\chi^{(t)}_{RD_G}(u), \chi^{(t)}_{RD_G}(v)\}\!\} = \{\!\{\chi^{(t)}_{RD_H}(x), \chi^{(t)}_{RD_H}(y)\}\!\}$ for all $t \geq 0$, then $\{u, v\}$ is a cut edge if and only if $\{x, y\}$ is a cut edge.*
2. *If $\chi_{RD}(G) = \chi_{RD}(H)$, then the block cut-edge trees of G and H are isomorphic.*

Background The following are well-known facts about effective resistance that will be useful for proving this theorem.

Lemma B.19. *Let G be a graph with $\{u, v\} \in E_G$. Then $RD_G(u, v) \leq 1$. Moreover, $RD_G(u, v) = 1$ if and only if $\{u, v\}$ is a cut edge.*

Lemma B.20. *Let G be a graph with $u, v, w \in V$. Then $RD_G(u, v) + RD_G(v, w) = RD_G(u, w)$ if and only if v is a cut vertex such that u and w are disconnected in $G \setminus \{v\}$.*

Notation For any pair of vertices u and v of a graph G, their resistance distance is denoted by $RD_G(u, v)$; however, we may drop the subscript G when it is clear from the context.

Recall that the notation $\chi_{RD}(G) = \chi_{RD}(H)$ means that two graphs G and H have the same multisets of the RD-WL colorings of their vertices for all $t \geq 0$, i.e. $\{\!\{\chi^{(t)}_{RD_G}(v) : v \in V_G\}\!\} = \{\!\{\chi^{(t)}_{RD_H}(v) : v \in V_H\}\!\}$, where χ_{RD_G} and χ_{RD_H} are RD-WL colorings of G and H, respectively. Similarly, we define $\chi_{RD_G}(u)$ as the ordered tuple $(\chi^{(0)}_{RD_G}(u), \chi^{(1)}_{RD_G}(u), \ldots)$. In other words, $\chi_{RD_G}(u) = \chi_{RD_H}(v)$ if and only if $\chi^{(t)}_{RD_G}(u) = \chi^{(t)}_{RD_H}(v)$ for all $t \geq 0$. We call these the ***colors*** of the graphs and nodes, respectively. We prove the following lemma about this new notation.

Lemma B.21. *Let $u \in V_G$ and $u' \in V_H$ be such that $\chi_{RD_G}(u) = \chi_{RD_H}(u')$. Then $\{\!\{(\chi_{RD_G}(v), RD_G(u, v)) : v \in V_G\}\!\} = \{\!\{(\chi_{RD_H}(v'), RD_H(u', v')) : v' \in V_H\}\!\}$.*

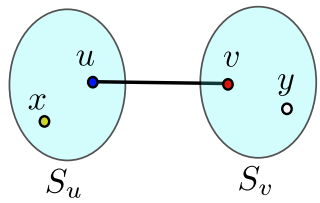
Proof. If not, then there is some iteration t such that $\{\!\{(\chi^{(t)}_{RD_G}(v), RD_G(u, v)) : v \in V_G\}\!\} \neq \{\!\{(\chi^{(t)}_{RD_H}(v'), RD_H(u', v')) : v \in V_H\}\!\}$. In this case, clearly $\chi^{(t+1)}_{RD_G}(u) \neq \chi^{(t+1)}_{RD_H}(u')$. □

Lemma B.22. *Let C be any set of colors. Let $u \in V_G$ and $u' \in V_H$ be such that $\chi_{RD_G}(u) = \chi_{RD_H}(u')$. Let $D(u)$ (respectively, $D(u')$) be the multisets of the resistance distances of all vertices of G (respectively, H) with color C from u (respectively, u'), i.e. $D(u) = \{\!\{RD_G(u, v) : v \in V_G, \chi_{RD_G}(v) \in C\}\!\}$ and $D(u') = \{\!\{RD_H(u', v') : v' \in V_H, \chi_{RD_H}(v) \in C\}\!\}$. Then $D(u) = D(u')$. In particular, the mean, minimum, and maximum of the multisets $D(u)$ and $D(u')$ are equal.*

Proof. If these sets are not the same, then by Lemma B.21, it cannot be the case that $\chi_{RD_G}(u) = \chi_{RD_H}(u')$. □

Lemma B.23. *Let $\{u, v\}$ be a cut edge in G, and let S_u and S_v be the vertex sets of the connected components of $G \setminus \{u, v\}$ such that $u \in S_u$ and $v \in S_v$. Let $C_{uv} \subset V$ be the subset of all vertices that have RD-WL-color $\chi_{RD_G}(u)$ or $\chi_{RD_H}(v)$. Then $S_u \cap C_{uv} = \{u\}$ or $S_v \cap C_{uv} = \{v\}$.*

Proof. Let x be the farthest vertex of C_{uv} (with respect to the resistance distance) from the set $\{u, v\}$ in G. Assume $x \neq u$ and $x \neq v$, otherwise the statement of the lemma trivially holds. Without loss of generality assume $x \in S_u$. We show that $C_{uv} \cap S_v = \{v\}$.



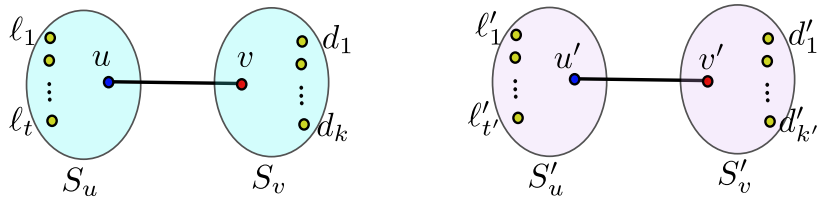
Let $y \in S_v$, and $y \neq v$. Let m_y, m_u an m_v be the maximum (resistance) distance of y, u and v from any vertex in C_{uv}. Note $m_u = RD(u,x)$, and $m_v = RD(v,x) = 1 + RD(u,x)$ because of the choice of x. Further, we have $m_y \geq RD(y,v) + RD(v,x) = RD(y,v) + m_v > m_v$, and that $m_y \geq RD(y,u) + RD(u,x) = RD(y,u) + m_u = RD(y,v) + 1 + m_u > m_u$ as $\{u,v\}$ is a cut edge. Thus, $m_y \neq m_u$ and $m_y \neq m_v$. In particular, the list of distances of both v and u to vertices in $C_{u,v}$ are different from the list of distances of y to these vertices. Hence, by Lemma B.22, the color $\chi_{RD_G}(y)$ is different from both of the colors $\chi_{RD_G}(u)$ and $\chi_{RD_G}(v)$. □

Theorem B.24. *Let G and H be two graphs with the same RD-WL colorings. Let $\{u,v\} \in E_G$ be a cut edge, and let $\{u',v'\} \in E_H$ be such that $\chi_{RD_G}(u) = \chi_{RD_H}(u')$ and $\chi_{RD_G}(v) = \chi_{RD_H}(v')$. Then, $\{u',v'\}$ is a cut edge in H.*

Proof. Let S_u and S_v be the vertex sets of the connected components of $G \backslash \{u,v\}$ such that $u \in S_u$ and $v \in S_v$. Also, let C^G_{uv} (respectively, $C^H_{u'v'}$) be the set of all vertices that are colored $\chi_{RD_G}(u)$ or $\chi_{RD_G}(v)$ in G (respectively, in H). By Lemma B.23, and without loss of generality, we assume $C^G_{uv} \cap S_v = \{v\}$. Therefore, the closest vertex of C^G_{uv} to v is u, which is at distance 1 to v since $\{u,v\}$ is a cut edge; the vertex u is the closest to v as there are no other vertices besides v in $C^G_{uv} \cap S_v$, and any vertex in $C^G_{uv} \cap S_u$ is at distance at least 1 from v via the series formula of the resistance distance.

Since $\chi_{RD_G}(v) = \chi_{RD_H}(v')$, the closest vertex in $C^H_{u'v'}$ to v' has distance 1 from v'—otherwise $\chi_{RD_H}(v')$ would be different from $\chi_{RD_G}(v)$ by Lemma B.22. Moreover, $RD_H(u',v') \leq 1$ as $\{u',v'\} \in E_H$. Thus, $RD_H(u',v') = 1$ and, by Lemma B.19, $\{u',v'\}$ is a cut edge. □

Lemma B.25. *Let G and H be two graphs with the same RD-WL colorings. Let $\{u,v\}$ be a cut edge in G and S_u, S_v be the partitions of $G \backslash \{u,v\}$. Let $\{u',v'\}$ be a cut edge in H such that $\chi_{RD_G}(u) = \chi_{RD_H}(u')$ and $\chi_{RD_G}(v) = \chi_{RD_H}(v')$, and let $S'_{u'}$, $S'_{v'}$ be partitions of $H \backslash \{u',v'\}$. Let C be any subset of colors used in these equivalent colorings. Let $D_v = \{\!\{d_1, \ldots, d_k\}\!\}$ be the list of distances of vertices with color C in S_v from v and let $L_u = \{\!\{\ell_1, \ldots, \ell_t\}\!\}$ be the list of distances of vertices with color C in S_u from u. Similarly, let $D'_{v'} = \{\!\{d'_1, \ldots, d'_{k'}\}\!\}$ be the list of distances of vertices with color C in $S'_{v'}$ from v' and let $L'_{u'} = \{\!\{\ell'_1, \ldots, \ell'_{t'}\}\!\}$ be the list of distances of vertices with color C in $S'_{u'}$ from u'. Then, we have $D_v = D'_{v'}$ and $L_u = L'_{u'}$. In particular, $|D_v| = |D'_{v'}|$ and $|L_u| = |L'_{u'}|$.*



Proof. Without loss of generality, by permuting indices, let $D_I = D_v \cap D'_{v'} = \{\!\{d_1, \ldots, d_p\}\!\} = \{\!\{d'_1, \ldots, d'_p\}\!\}$, and $L_I = L_u \cap L'_{u'} = \{\!\{\ell_1, \ldots, \ell_q\}\!\} = \{\!\{\ell'_1, \ldots, \ell'_q\}\!\}$. Hence,

$$D'_{v'} = \{\!\{d_1, \ldots, d_p, d'_{p+1}, \ldots, d'_{k'}\}\!\},$$

and

$$L'_{u'} = \{\!\{\ell_1, \ldots, \ell_q, \ell'_{q+1}, \ldots, \ell'_{t'}\}\!\}.$$

By the definition of D_I,

$$\{\!\{d'_{p+1}, \ldots, d'_{k'}\}\!\} \cap \{\!\{d_{p+1}, \ldots, d_k\}\!\} = \emptyset, \tag{3}$$

and

$$\{\!\{\ell'_{q+1}, \ldots, \ell'_{t'}\}\!\} \cap \{\!\{\ell_{q+1}, \ldots, \ell_t\}\!\} = \emptyset. \tag{4}$$

Since $\chi_{RD_G}(v) = \chi_{RD_H}(v')$, by Lemma B.22, the list of distances of vertices with color C from v in G is the same as the list of vertices with color C from v' in H. As $\{u, v\}$ and $\{u', v'\}$ are cut edges, then $D_v \cup (L_u + 1) = D'_{v'} \cup (L'_{u'} + 1)$, where $+1$ is applied to all elements of a multiset. Specifically,

$$\begin{aligned}\{\!\{d_1, \dots, d_k, \ell_1 + 1, \dots, \ell_t + 1\}\!\} &= \{\!\{d'_1, \dots, d'_{k'}, \ell'_1 + 1, \dots, \ell'_{t'} + 1\}\!\} \\ &= \{\!\{d_1, \dots, d_p, d'_{p+1}, \dots d'_{k'}, \ell_1 + 1, \dots, \ell_q + 1, \ell'_{q+1} + 1 \dots \ell'_{t'} + 1\}\!\}\end{aligned}$$

Therefore, removing D_I and L_I from these multisets, we conclude

$$\{\!\{d_{p+1}, \dots, d_k, \ell_{q+1} + 1, \dots, \ell_t + 1\}\!\} = \{\!\{d'_{p+1}, \dots d'_{k'}, \ell'_{q+1} + 1 \dots \ell'_{t'} + 1\}\!\}$$

Equations (3) and (4) imply

$$\{\!\{d_{p+1}, \dots, d_k\}\!\} = \{\!\{\ell'_{q+1} + 1 \dots \ell'_{t'} + 1\}\!\},$$

and

$$\{\!\{\ell_{q+1} + 1, \dots, \ell_t + 1\}\!\} = \{\!\{d'_{p+1}, \dots d'_{k'}\}\!\}. \tag{5}$$

On the other hand, since the list of distances of vertices with color C from u in G is the same as the list of vertices with color C from u' in H, i.e. $(D_v + 1) \cup L_u = (D'_{v'} + 1) \cup L'_{u'}$. Specifically,

$$\begin{aligned}\{\!\{d_1 + 1, \dots, d_k + 1, \ell_1, \dots, \ell_t\}\!\} &= \{\!\{d'_1 + 1, \dots, d'_{k'} + 1, \ell'_1, \dots, \ell'_{t'}\}\!\} \\ &= \{\!\{d_1 + 1, \dots, d_p + 1, d'_{p+1} + 1, \dots d'_{k'} + 1, \ell_1, \dots, \ell_q, \ell'_{q+1} \dots \ell'_{t'}\}\!\}.\end{aligned}$$

Therefore,

$$\{\!\{d_{p+1} + 1, \dots, d_k + 1, \ell_{q+1}, \dots, \ell_t\}\!\} = \{\!\{d'_{p+1} + 1, \dots d'_{k'} + 1, \ell'_{q+1} \dots \ell'_{t'}\}\!\}.$$

Again, Equations (3) and (4), imply

$$\{\!\{d_{p+1} + 1, \dots, d_k + 1\}\!\} = \{\!\{\ell'_{q+1} \dots \ell'_{t'}\}\!\},$$

and

$$\{\!\{\ell_{q+1}, \dots, \ell_t\}\!\} = \{\!\{d'_{p+1} + 1, \dots d'_{k'} + 1\}\!\}. \tag{6}$$

Combining Equations (5) and (6), we conclude

$$\{\!\{\ell_{q+1}, \dots, \ell_t\}\!\} = \{\!\{d'_{p+1} + 1, \dots d'_{k'} + 1\}\!\} = \{\!\{\ell_{q+1} + 2, \dots, \ell_t + 2\}\!\}.$$

Thus, $\{\!\{\ell_{q+1}, \dots, \ell_t\}\!\}$ is empty, implying $L_u = L'_{u'}$, and $D_v = D'_{v'}$, as desired. □

Lemma B.26. *Let G and H be two graphs with the same RD-WL colorings. Let (i) $\{u, v\}$ be a cut edge in G, (ii) $u' \in V_H$ such that $\chi_{RD_G}(u) = \chi_{RD_H}(u')$, (iii) $v' \in V_H$ such that $\chi_{RD_H}(v') = \chi_{RD_G}(v)$, and $RD_H(u', v') = 1$. Then, $\{u', v'\}$ must be a cut edge in H.*

Note that the lemma does not trivially follow from Lemma B.19 since $\{u', v'\}$ is not assumed to be an edge a priori.

Proof. Let C_{uv} be the set of all vertices that are colored $\chi_{RD_G}(u)$ or $\chi_{RD_G}(v)$ in G. By Lemma B.23, $S_u \cap C_{uv} = \{u\}$ or $S_v \cap C_{uv} = \{v\}$. Assume that $S_v \cap C_{uv} = \{v\}$; the other case is similar.

Suppose, to derive a contradiction, that $\{u', v'\}$ is not a cut edge. Let w' be a neighbor of v' that resides in the same connected component as u' in $H \backslash \{v'\}$. (In particular, if v' is not a cut vertex w' can be any neighbor of v'). Such a w' exists and is distinct from u', otherwise $\{u', v'\}$ would be a cut edge. We know that $RD_H(w', u') < RD_H(w', v') + RD_H(v', u') = RD_H(w', v') + 1$; the triangle inequality implies the inequality, and it is a strict inequality because, in the case of equality, v' would a cut vertex separating w' and u', contradicting the choice of w'. Let $\ell = RD_H(w', v')$, and note that $\ell < 1$, as otherwise $\{w', v'\}$ is a cut edge between v' and u' implying that $RD_H(u', v') > 1$.

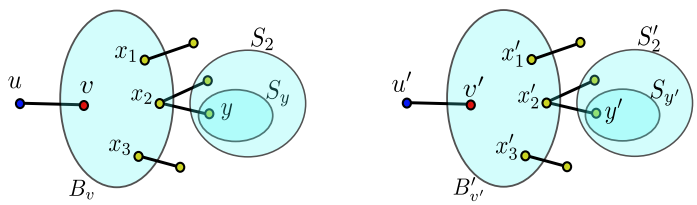
Since $\chi_{RD_G}(v) = \chi_{RD_H}(v')$, there must exist a vertex $w \in V_G$ such that $\chi_{RD_G}(w) = \chi_{RD_H}(w')$ and $RD_G(v, w) = \ell < 1$. Since all vertices in S_u are at a distance at least one from v, we have $w \in S_v$. Let L_w be the list of distances of vertices in C_{uv} from w. Since $S_v \cap C_{uv} = \{v\}$, the smallest distance in L_w is ℓ and the second smallest distance in $\ell + 1$, to v and u, respectively.

Let $C_{u'v'}$ be the set of all vertices in H that are colored $\chi_{RD_H}(u') = \chi_{RD_G}(u)$ or $\chi_{RD_H}(v') = \chi_{RD_G}(v)$. Then, let $L_{w'}$ be the list of all distances of w' to vertices in $C_{u'v'}$. Since $\chi_{RD_H}(w') = \chi_{RD_G}(w)$, then $L_{w'} = L_w$ by Lemma B.22. In particular, the smallest two numbers in $L_{w'}$ are ℓ and $\ell + 1$. We know $RD_H(w', v') = \ell$, and by the triangle inequality $RD_H(w', u') \leq RD_H(w', v') + 1 = \ell + 1$. Since, the second smallest number in the list is $\ell + 1$, and not smaller, the inequality must be tight, i.e. $RD_H(w', u') = RD_H(w', v') + 1 = RD_H(w', v') + RD_H(v', u')$. Therefore, v' is a cut vertex separating u' and w', contradicting the choice of w'. □

Lemma B.27. *Let G and H be two graphs with the same RD-WL colorings. Let $\{u, v\}$ be a cut edge in G and $\{u', v'\}$ a cut edge in H such that $\chi_{RD_G}(u) = \chi_{RD_H}(u')$ and $\chi_{RD_G}(v) = \chi_{RD_H}(v')$. Furthermore, let B_v (resp. $B'_{v'}$) be the biconnected component of v (resp. v') in G (resp. H). Finally, let K (resp. K') be the set of all cut edges except $\{u, v\}$ (resp. $\{u', v'\}$) with one endpoint in B_v (resp. $\{u', v'\}$). Then there exists a bijection between $f : K \to K'$, such that for any $\{x, y\} \in K$ and $\{w, z\} = f(\{x, y\})$, if x in on v's side in $G \backslash \{x, y\}$ and w is on v''s side in $H \backslash \{w, z\}$, then $\chi_{RD_G}(x) = \chi_{RD_H}(w)$ and $\chi_{RD_G}(y) = \chi_{RD_H}(z)$.*

Proof. Let S_v (resp. $S'_{v'}$) be the vertex set of the connected component of $G \backslash \{u, v\}$ (resp. $H \backslash \{u', v'\}$) that contains v (resp. v').

Let $X = \{x_1, \ldots, x_k\}$ be the set of endpoints of K that are in B_v ordered such that $RD(v, x_1) \leq \ldots \leq RD(v, x_k)$. (Note that each x_i may be connected to multiple cut edges of K.)



By Lemma B.25, there exists $X' = \{x'_1, \ldots, x'_k\} \subset S'_{v'}$ such that for every $i \in [k]$

(1) $\chi_{RD_G}(x_i) = \chi_{RD_H}(x'_i)$ and

(2) $RD_G(v, x_i) = RD_H(v', x'_i)$.

(Let C_X be the set of all vertices that are colored by the colors $\{\chi_{RD_G}(x_1), \ldots, \chi_{RD_G}(x_k)\}$ in S_v. Note, C_X is a (possibly proper) superset of $\{x_1, \ldots, x_k\}$. Therefore, by Lemma B.25, there exists a set $C'_X \subseteq S_{v'}$, such that there is a one-to-one correspondence between C_X and C'_X with corresponding pairs have the same color and distance to v and v'. The existence of X' with the aforementioned properties follows as $X \subseteq C_X$.)

For each $i \in [k]$, let Y_i be the set of all vertices that are adjacent to x_i with a cut edge in K. Since $\chi_{RD_G}(x_i) = \chi_{RD_H}(x'_i)$, by Lemma B.26, for each $i \in [k]$, there exists Y'_i such that (1) for all $y' \in Y'_i$, $\{x'_i, y'\}$ is a cut edge, and (2) the multisets $\{\!\{\chi_{RD_G}(y)\}\!\}_{y \in Y_i} = \{\!\{\chi_{RD_H}(y')\}\!\}_{y' \in Y'_i}$.

Next, we show for every $i \in [k]$ and every $y' \in Y'_i$, $RD_H(v', x'_i) < RD_H(v', y'_i)$, or equivalently, v' is in the connected component of x'_i in $H \backslash \{x'_i, y'\}$. We prove this by induction on i.

For each $i \in [k]$, and each $y \in Y_i$, let S_y be the set of all vertices on the y side in $G \backslash \{x_i, y\}$, and let $S_i = \bigcup_{y \in Y_i} S_y$. For all $h < i$, and each $y' \in Y'_h$, let $S'_{y'}$ be the set of all vertices on the y' side in $H \backslash \{x'_h, y'\}$, and let $S'_h = \bigcup_{y' \in Y'_h} S'_{y'}$. Note $S_v \backslash \bigcup_{h<i} S_h$ contains v and, by the induction hypothesis, $S'_{v'} \backslash \bigcup_{h<i} S'_h$ contains v'.

By Lemma B.25 and that $RD_G(v, x_h) = RD_H(v', x'_h)$, S_h and S'_h induce the same set of color/distance pairs from v and v', respectively; i.e. there is a one-to-one map f from S_h to S'_h such that for each pair $(q, f(q))$, (i) q and $f(q)$ have the same color, and (ii) $RD_G(v, q) = RD_H(v', f(q))$. Thus, $B_{v,i} = S_v \backslash \bigcup_{h<i} S_h$ and $B'_{v,i} = S'_{v'} \backslash \bigcup_{h<i} S'_h$ induce the

same set of color/distance pairs from v and v', respectively. Let C be the set of colors of vertices that are incident to cut edges. The ith smallest distance with color from C in $B_{v,i}$ corresponds to x_i. We know x'_i has the same color as x_i, and $RD_G(v, x_i) = RD_H(v', x'_i)$. If there exists $y' \in Y'_i$ such that the $RD_H(v', y') < RD_H(v', x_i)$ then the ith smallest number in $B'_{v,i}$ would be smaller than $RD_G(v, x_i)$ which is a contradiction. Hence, $RD_H(v', x'_i) < RD_H(v', y')$ for all $y' \in Y'_i$, as desired.

Therefore, $B'_{v'} = S'_{v'} \setminus \bigcup_{i\in[k]} S'_i$ is connected, and it has the same set of colors as $B_v = S_v \setminus \bigcup_{i\in[k]} S_i$. So, it must be the biconnected component that contains v'. □

Theorem B.28. *If two graphs G and H have the same RD-WL colorings, then their block cut-edge trees are isomorphic.*

Proof. Let $\{u, v\}$ be a cut edge in G such that S_u is a biconnected component. Let $\{u', v'\}$ be an edge in H with $\chi_{RD_G}(u) = \chi_{RD_H}(u')$ and $\chi_{RD_G}(v) = \chi_{RD_H}(v')$. By Lemma B.27, $S'_{u'}$ is a biconnected component of H. Root the block cut-edge tree of G at S_u and the block cut-edge tree of H at $S'_{u'}$ so all cut edges are directed away from S_u and $S'_{u'}$.

For a directed cut edge $u \to v$, define $T_G[v]$ the be the block cut-edge subtree rooted at the biconnected component of v. We use induction on the size of $T_G[v]$ to show that for two same colored edges with the same direction $u \to v$ and $u' \to v'$, $T_G[v]$ and $T_H[v']$ are isomorphic.

Let B_v and $B'_{v'}$ be the biconnected components of v in G and v' in H respectively. For the base case that $B_v = S_v$, then by Lemma B.27, $B'_{v'} = S'_{v'}$ and the statement holds. Alternatively, also by Lemma B.27, the outgoing directed edges of B_v and $B'_{v'}$ are in one-to-one correspondence with matching head and tail colors. Let these sets of edges be K and K', respectively, and let f be the bijection between them that preserves colors. For any $x \to y \in K$, and its image $f(x) \to f(y) = w \to z \in K'$, by the induction hypothesis, the block cut-edge subtrees $T_G[y]$ and $T_H[z]$ are isomorphic. Therefore, $T_G[v]$ and $T_H[v']$ are isomorphic. □

C. Experiments

In this section, we carry out experiments to validate our two main results Theorem 3.8 and Theorem 3.10. Our code is adapted from the GraphGPS module (Rampášek et al., 2022) and subsequent fork from Müller et al. (2024).

C.1. Graph Isomorphism: CSL

Table 1. Test performance on the CSL dataset of different APEs. Results are shown in the form of mean ± standard deviation. Experiments are averaged over 5 runs.

CSL Dataset: Classification Accuracy (↑)		
	HKdiagSE	RWSE
APE-GT	100 ± 0.00	100 ± 0.00
DeepSet RPE-GT	100 ± 0.00	100 ± 0.00

Table 2. Test performance on the CSL dataset of different RPEs. Results are shown in the form of mean ± standard deviation. Experiments are averaged over 5 runs. The result for GIN was taken from Müller et al. (2024).

CSL Dataset: Classification Accuracy (↑)					
	RSPE	Powers of Adjacency	RD	SPD	Adjacency
RPE-GT	100 ± 0.00	100 ± 0.00	100 ± 0.00	90 ± 0.00	10 ± 0.00
EGN APE-GT	100 ± 0.00	100 ± 0.00	100 ± 0.00	90 ± 0.00	10 ± 0.00
GIN	10 ± 0.00				

We consider the graph classification task on the Circular Skip Links (CSL) dataset (Dwivedi et al., 2023). The goal of the CSL dataset is to classify graphs according to their isomorphism type. This is classified as a "Hard" task by Müller et al. (2024).

To validate Theorem 3.8, we consider two different APEs: Random Walk Structural Encoding (RWSE) and Heat Kernel Diagonal Structural Encoding (HKdiagSE) from the GraphGPS Library (Rampášek et al., 2022). For both APEs, we used

times $\{1, ..., 20\}$. We test these APEs with APE-GTs and their corresponding RPEs by applying a DeepSet and the result GTs are denoted DeepSet RPE-GTs.

To validate Theorem 3.10, we consider four different RPEs: RSPE (Huang et al., 2024), a stack of 20 powers of the adjacency matrix (inspired by RRWP (Ma et al., 2023)), the resistance distance (RD), the shortest-path distance (SPD), and the adjacency matrix. We then consider their corresponding APEs by applying a 2-EGN and the result GTs are denoted EGN APE-GTs. We ran the RPE-GT and APE-GT for 1000 epochs each. Our results are averaged over 5 random seeds. See Table 1 and Table 2 for the results.

As our theoretical results (Theorems 3.8 and 3.10) predict, RPE-GTs and EGN APE-GTs have the same classification accuracy when using the same RPE. Likewise, APE-GT and DeepSets RPE-GT have the same classification accuracy when using the same APE. This validates our main theorems.

Moreover, note that the accuracy of Adjacency-RPE-GT and GIN are equal, which agrees with Corollary 4.17.

C.2. Graph Isomorphism: BREC

Table 3. Performance of different RPEs on the BREC dataset. Basic, Regular, Extension, and CFI are subsets of the BREC dataset. Results are shown in the form of mean $\pm$ standard deviation. Experiments are averaged over 5 runs.

BREC Dataset: Pair-Distinguishing Accuracy (↑)					
	Basic	Regular	Extension	CFI	Total
RPE: Resistance Distance (RD)					
RPE-GT	100.00 ± 0.00	35.71 ± 0.00	100.00 ± 0.00	7.60 ± 1.14	54.40 ± 0.29
EGN APE-GT	100.00 ± 0.00	35.71 ± 0.00	100.00 ± 0.00	7.00 ± 1.23	54.25 ± 0.31
RPE: Shortest-Path Distance (SPD)					
RPE-GT	26.67 ± 0.00	9.29 ± 0.00	40.8 ± 0.45	6.8 ± 0.11	19.15 ± 0.22
EGN APE-GT	26.67 ± 0.00	9.42 ± 0.28	41.00 ± 0.00	11.80 ± 0.45	20.50 ± 0.18
RPE: Stable Positional Encoding (SPE)					
RPE-GT	97.00 ± 2.17	34.71 ± 0.57	96.00 ± 1.87	3.2 ± 4.47	51.5 ± 0.73
EGN APE-GT	0.97 ± 2.17	33.14 ± 0.97	98.80 ± 0.84	8.80 ± 1.30	53.05 ± 0.54
RPE: Stack of Powers of Adjacency Matrix					
RPE-GT	100.00 ± 0.00	35.00 ± 00	94.6 ± 0.00	9.80 ± 0.84	53.55 ± 0.28
EGN APE-GT	83.00 ± 1.39	32.86 ± 0.00	88.20 ± 1.79	3.00 ± 00	46.75 ± 0.47
RPE: Adjacency Matrix					
RPE-GT	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
EGN APE-GT	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 4. Performance of different APEs on the BREC dataset. Basic, Regular, Extension, and CFI are subsets of the BREC dataset. Results are shown in the form of mean $\pm$ standard deviation. Experiments are averaged over 5 runs.

BREC Dataset: Pair-Distinguishing Accuracy (↑)					
	Basic	Regular	Extension	CFI	Total
APE: Heat Kernel Diagonal Structural Encoding (HKdiagSE)					
APE-GT	56.00 ± 4.50	26.00 ± 0.57	38.00 ± 3.16	3.00 ± 00	27.75 ± 1.43
DeepSets RPE-GT	59.00 ± 4.01	18.86 ± 1.84	59.60 ± 6.11	1.2 ± 1.30	30.65 ± 1.57
APE: Random Walk Structural Encoding (RWSE)					
APE-GT	84.33 ± 3.65	31.29 ± 0.83	62.40 ± 4.39	1.00 ± 0.00	39.45 ± 1.10
DeepSets RPE-GT	69.67 ± 5.32	21.29 ± 2.73	63.80 ± 2.39	0.4 ± 0.55	33.95 ± 1.81

We consider the graph isomorphism benchmark dataset BREC (Wang & Zhang, 2024). The BREC dataset tests a graph neural network architecture's ability to distinguish pairs of non-isomorphic graphs. Unlike the CSL dataset which poses the graph isomorphism problem as a graph classification problem, the BREC dataset considers pairs of non-isomorphic graphs and tests whether a graph neural networks can learn different representations for the two graphs. The graph neural network is trained via contrastive learning on sets of graphs pairs with randomly permuted node indices. The graph neural network

is then evaluated using the T^2 test on the learned features for each of the two graphs. The exact training and evaluation procedures can be found in the original BREC dataset paper (Wang & Zhang, 2024).

We compared several RPEs on the BREC dataset. We found that RD, SPE, and a stack of 20 powers of the adjacency matrix all perform similarly well and perform better than SPD in terms of total accuracy. Moreover, the adjacency matrix achieves 0% accuracy, which matches the theoretical results of Proposition 4.10 as all pairs of graphs in the BREC dataset are WL indistinguishable.

We also compared two APEs on the BREC dataset: RWSE and HKdiagSE from the GraphGPS library (Rampášek et al., 2022). For both APEs, we used times $\{1, ..., 20\}$. We found that RWSE slightly outperformed HKdiagSE.

Finally, for the most part, both RPEs and APEs achieved similar performance for their different architectures, i.e. RPE-GT and EGN APE-GT or APE-GTs and DeepSets RPE-GT. This matches our theoretical results (Theorems 3.8 and 3.10).

However, one interesting thing we found is that there are several graph pairs in the CFI subset that SPD-GTs were able to learn to distinguish while RD-GTs did not. This suggests that SPD-WL and RD-WL may be incomparable. Zhang et al. (2023) show there are graphs that RD-WL can distinguish that SPD-WL cannot; this suggests the converse is true too, although it is not a definitive proof of this fact and should not be interpreted as such. A question for future research would be to prove that these pairs of graphs are indistinguishable by RD-WL or to find a RD-GT that could distinguish these pairs of graphs.

Hyperparameter search was performed for each pair of PE and architecture.

C.3. Graph Regression: ZINC

Table 5. Test performance on the Small ZINC dataset. Results are shown in the form of mean $\pm$ standard deviation. Experiments are averaged over 3 runs.

ZINC Dataset: MAE (↓)				
	# Parameters	RD	SPD	SPE
RPE-GT	573922	0.096 ± 0.002	0.130 ± 0.010	0.132 ± 0.007
EGN APE-GT	520514	0.196 ± 0.004	0.217 ± 0.011	0.200 ± 0.007
	# Parameters	RD+EF	SPD+EF	SPE+EF
RPE-GT	696226	0.069 ± 0.004	0.092 ± 0.010	0.092 ± 0.001
EGN APE-GT	664818	0.110 ± 0.003	0.132 ± 0.004	0.126 ± 0.003

Although our theoretical results are for the distinguishing power of graph transformers and apply most directly to graph classification/isomorphism tasks, we would also like to compare RPEs and APEs on real-world tasks beyond those. In this experiment, we compare RPE-GTs and EGN APE-GTS for graph regression on the small ZINC dataset containing 12k graphs (Dwivedi et al., 2023). We consider the RPEs resistance distance (RD), shortest-path distance (SPD), and SPE (Huang et al., 2024). See Table 5 for the results.

We test the models both with and without the use of edge features of graphs in the ZINC dataset. Edge features are distinct from RPEs as they are not dependent on the graph structure and are associated to edges and not a pairs of vertices. For the ZINC dataset, the features are one-hot encodings of the bond type (e.g. single, double, triple). When we include edge features, we concatenate them to the RPE and pass the concatenated tensor to the model in place on the RPE. This use of edge features is not captured by our theoretical results, and it is not clear if this is the best or most natural way of including edge features in a graph transformer. Other papers have proposed different ways of incorporating edge features into a graph transformer (Ma et al., 2023; Jin et al., 2023).

From Table 5, we see that the performance of EGN APE-GT is worse than the corresponding RPE-GT. This could be partially explained by Theorem 3.10, which states that RPE-GTs have stronger distinguishing power than EGN APE-GTs for graphs with node features (as the graphs in ZINC do). Additionally, we suspect that the training of the extra EGN layer introduces additional difficulty in the training process, leading to worse performance of APE-GT compared to RPE-GT. This result aligns with our expectations, suggesting that instead of converting an RPE into an APE, it may be better to use the RPE directly in a RPE-GT.

Hyperparameters & experimental setup. Our results are averaged over 3 random seeds. We ran the RPE-GT and APE-GT for 1000 epochs each.

Following Zhang et al. (2023), the functions f_1 and f_2 in the RPE-GT head are embeddings into Gaussian kernels followed by an MLP.

Hyperparameter search was performed for RPE-GTs and EGN APE-GTs. We chose the final hyperparameters for each architecture to have similar numbers of parameters. Hyperparameters involved in our models are listed in Table 6.

Table 6.

Hyperparameters for ZINC Experiment	
RPE GTs	
# Transformer Layers:	14
# Transformer Heads:	8
# Gaussian Kernels	16
# MLP Layers	2
MLP Hidden Dimension (No Edge Features)	16
MLP Hidden Dimension (Edge Features)	16
EGN APE GTs	
# Transformer Layers:	8
# Transformer Heads:	8
# EGN Layers	6
EGN Hidden Dim (No Edge Features)	48
EGN Hidden Dim (Edge Features)	64
APE Type	Add
SPE	
# DeepSets Layers	3
DeepSets Hidden Dimension	64
# Parameters	17217

C.4. Code

Code for all experiments can be found at `https://github.com/blackmit/comparing_graph_transformers_via_positional_encodings`