# When and How to Canonize: a Generalization Perspective

**Yonatan Sverdlov***
Technion – Israel Institute of Technology
yonatans@campus.technion.ac.il

**Benjamin Friedman**
Technion – Israel Institute of Technology
benjamin.fri@campus.technion.ac.il

**Snir Hordan**
Technion – Israel Institute of Technology
snirhordan@campus.technion.ac.il

**Nadav Dym**
Technion – Israel Institute of Technology
nadavdym@technion.ac.il

## Abstract

While invariant architectures are standard for processing symmetric data, there is growing interest in achieving invariance by applying group averaging or canonization to non-invariant backbones. However, the theoretical generalization properties of these alternative strategies remain poorly understood. We introduce a theoretical framework to analyze the generalization error of these methods by bounding their covering numbers. We establish a rigorous generalization hierarchy: the error bounds of canonized models are at best equal to the error bounds of structurally invariant and group-averaged models, and at worst equal to the bounds of non-invariant baselines. Furthermore, we show that there exist "optimal" canonizations which attain the optimal error bounds, and "poor" canonizations which attain the non-invariant error bounds, and that this depends on the regularity of the canonization. Finally, applying this framework to permutation groups in point cloud processing, we rigorously prove that the covering number of lexicographical sorting grows exponentially with the number of points in the point cloud, whereas Hilbert curve canonization guarantees polynomial growth. This provides the first formal theoretical justification for the empirical success of Hilbert curve serialization in state-of-the-art point cloud architectures. We conclude with experiments that support our theoretical claims. Our code is publicly available on GitHub.

## 1 Introduction

The integration of geometric priors into neural networks has become a cornerstone of modern machine learning, particularly for applications involving graphs, 3D point clouds, and molecular structures. In these domains, the underlying data often exhibits inherent symmetries, meaning the target function to be learned is invariant to a specific group action. Exploiting these symmetries is known to improve sample complexity and generalization [8, 5].

While specialized architectures, which are invariant by design, are a common method for processing symmetric data, there is growing interest in achieving invariance by applying generalized group averaging methods to non-invariant backbones.

This family of methods includes full group averaging, which is theoretically sound but intractable for all but very small groups, group augmentation, which encourages invariance by randomly sampling group transformations during training, frame averaging [23], which enables invariant averaging

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

over subsets of the group, and canonization, which maps every group orbit to a single, consistent "canonical" element.

Canonization is the most efficient of all generalized group averaging methods, as it only involves processing a single orbit representative. At the same time, it enjoys the same universal approximation guarantees as full group averaging [14]. These observations have motivated researchers to define canonizations in several diverse domains. These include sign canonizations for spectral embeddings of graphs [19, 12], canonizations of graphs [17] and point clouds [14, 2, 10, 34] including the celebrated Point Transformer V3 [30], and canonization of the SMILES representation of small molecules [29].

In this paper, we provide a theoretical study of canonization from the perspective of generalization. Our focus is both on comparing canonization with other options for enforcing invariance and on comparing the quality of different canonizations. Our main contributions are threefold:

1. **Canonization vs. Group Averaging:** We prove that the generalization bounds of canonized models are at best equal to the bounds of group-averaged models, and at worst equal to the bounds of non-invariant models (Section 2).
2. **Canonizations and continuity:** We show that continuous, isometric canonizations attain the *optimal* generalization error achieved by group averaging, while discontinuous canonizations can lead to *poor* canonizations whose error bounds are identical to non-invariant models (Section 3).
3. **Hilbert Canonization:** We prove that canonizing point clouds with respect to permutation via lexicographical sorting is inferior to the Hilbert canonizations used in Point Transformer V3 [30], thus establishing the first theoretical justification for the empirical success of this method (Section 4).

### 1.1 Related Work

**Canonization vs. Group Averaging** It was observed that randomized SMILES can often outperform canonized SMILES [1, 3] (see also [13]). Our first contribution gives a theoretical justification for this empirical observation.

**Discontinuity** It was shown in [7, 2] that in many invariant learning scenarios, continuous canonizations are mathematically impossible. Attempts to mitigate this issue were made in [27, 16]. Our second contribution complements these works by showing how discontinuity of canonization hurts the generalization of the canonized model.

**Invariance and Generalization** Many papers have considered generalization analysis of invariant models [28, 22, 20, 9], but they do not consider canonizations. To our knowledge, the only paper considering the generalization of canonizations is [26]. They prove that in some settings the *approximation error* of canonized models is lower than group averaged models, and as a result the *expected error* of canonized models can be lower when the model is presented with enough samples. In contrast, our first conclusion is that the *generalization error bound* of group average models is always lower than the generalization error bound of canonized models.

### 1.2 Notation

**Symmetry space** Throughout the paper, we will consider a metric space $(K, \rho)$, where $K$ is endowed with the action of a group $\mathbb{G}$. We will assume that the metric is $\mathbb{G}$ invariant, i.e., $\rho(g \cdot x, g \cdot y) = \rho(x, y)$ for all $x, y \in K$ and $g \in \mathbb{G}$. We denote the orbit of $x \in K$ by $[x] = \{g \cdot x| \quad g \in \mathbb{G}\}$, and denote the quotient space to be the space of orbits $K/\mathbb{G} := \{[x]|x \in K\}$. On this space, we define a metric $\rho_{\mathbb{G}}([x], [y]) := \min_{g \in \mathbb{G}} \rho(g \cdot x, y)$, where we explicitly assume that the minimum in the definition of $\rho_{\mathbb{G}}$ exists, which is the case in most examples of practical interest. This assumption, together with the $\mathbb{G}$ invariance of $\rho$, guarantees that $\rho_{\mathbb{G}}$ is a metric. We call a triplet $(K, \rho, \mathbb{G})$ satisfying the conditions of this paragraph a *symmetry space*.

**Canonization** Let $(K, \rho, \mathbb{G})$ be a symmetry space. We say that $c : K \to K$ is a canonization if for every $x \in K$, (i) $c(x) \in [x]$, and (ii) all $y \in [x]$ satisfy $c(y) = c(x)$. We will sometimes use

canon. as an abbreviation for canonization. We also add the assumption that the canonization is Borel measurable.

**Covering number** Let $(K, \rho)$ be a metric space, and let $A \subseteq K$. We say that $C \subseteq K$ is an $\epsilon$-cover of $A$ if, for every $x \in A$, there exists $y \in C$ such that $\rho(x, y) \leq \epsilon$. The minimum cardinality of such an $\epsilon$-cover is called the $\epsilon$-covering number of $A$ (with centers in the ambient space $K$), and is denoted by $\mathcal{N}(A, \rho, \epsilon)$.

**Generalization bounds via covering numbers** Our discussion of the generalization properties of canonizations and other invariant models is based on a popular framework from [32] for bounding generalization via covering numbers. We begin with a short review of this framework.

Consider the problem of learning an unknown function $f : X \to Y$ from a finite set of $n$ samples $S = \{(x_i, f(x_i)) \mid i = 1, \ldots, n\}$ of the function drawn from the distribution $\mu$. Assume that we have some algorithm (e.g., gradient descent) which, given $S$, returns a hypothesis $h_S : X \to Y$. Let $\ell : Y \times Y \to \mathbb{R}$ be a function which will be used as a loss function. The *expected loss* $\mathcal{L}_{\text{exp}}$ and the *empirical loss* $\mathcal{L}_{\text{emp}}$ are defined via

$$\mathcal{L}_{\text{exp}}(h_S) \triangleq \mathbb{E}_{x\sim\mu}\ell(h_S(x), f(x)), \quad \mathcal{L}_{\text{emp}}(h_S) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(h_S(x_i), f(x_i))$$

The *generalization* error measures how much the empirical error of $h_S$ on the training data can deviate from its true loss (the expected loss) on the data distribution. It can be bounded using the Lipschitz constants of the functions $h_S, f, \ell$ and the covering number of the domain $X$, as follows:

**Theorem 1.1.** *[Proof in Appendix A.1, based on [32] ] Let $(K, \rho)$ and $(Y, \rho_Y)$ be metric spaces, and assume that $X \subseteq K$ is totally bounded. Let $\mu$ be a Borel probability distribution on $X$. Let $f : X \to Y$ be a $c_f$-Lipschitz function, let $h_S : X \to Y$ be a $c_{h_S}$-Lipschitz function, and let $\ell : Y \times Y \to [0, M]$ be a $c_\ell$-Lipschitz function. Then for any $0 < \epsilon$ and $0 < \delta < 1$ and natural $n$, for a training sample set $S$ generated by $n$ IID draws from $\mu$, with probability of at least $1 - \delta$,*

$$|\mathcal{L}_{\text{exp}}(h_S) - \mathcal{L}_{\text{emp}}(h_S)| \leq 2c_\ell(c_{h_S} + c_f)\epsilon + M\sqrt{\frac{2\mathcal{N}(X, \rho, \epsilon)\ln 2 + 2\ln(1/\delta)}{n}}$$

## 2 Generalization, Canonization and Group Averaging

In this section, we consider a task where the function $f$ to be learned is invariant to a group action. This function is approximated by a 'learned', possibly non-invariant, Lipschitz function $\tilde{h}$, which we will call the *backbone model*. We then consider three strategies of replacing $\tilde{h}$ with an alternative model $h$:

1. **Symmetry Agnostic models** Models that do not incorporate symmetry $h = \tilde{h}$.
2. **Canonized models** Models that incorporate symmetries by canonization, $h = \tilde{h} \circ c$.
3. **Averaged and invariant models** Models that incorporate symmetries by group averaging $h(x) = \int_{\mathbb{G}} \tilde{h}(g \cdot x)dg$, or models where $\tilde{h}$ is already invariant by construction, and so $h = \tilde{h}$. In either case, we have that $h$ and $\tilde{h}$ have the same Lipschitz constant (see Proposition A.2 in the appendix).

In this section we establish a hierarchy for the generalization bounds obtained for these three methods. The first step is the following proposition which bounds the generalization error using Theorem 1.1:

**Proposition 2.1.** *[Proof in Appendix A.1] Let $(K, \rho, \mathbb{G})$ be a symmetry space. Let $(Y, \rho_Y)$ be a metric space, and assume $K$ is totally bounded. Let $\mu$ be a Borel probability distribution on $K$. Let $f : K \to Y$ be a $c_f$ Lipschitz function **which is** $\mathbb{G}$ **invariant**, and let $\ell : Y \times Y \to [0, M]$ be a $c_\ell$ Lipschitz function , then for any $0 < \epsilon$ and $0 < \delta < 1$ and natural $n$, for a training sample set $S$ generated by $n$ IID draws from $\mu$, with probability of at least $1 - \delta$, we have*

$$|\mathcal{L}_{\text{exp}}(h_S) - \mathcal{L}_{\text{emp}}(h_S)| \leq 2c_\ell(c_{h_S} + c_f)\epsilon + M\sqrt{\frac{2\mathcal{N}(h_S)\ln 2 + 2\ln(1/\delta)}{n}}$$

*where*

$$\mathcal{N}(h_S) = \begin{cases} \mathcal{N}(K, \rho, \epsilon) & \text{if } h_S \text{ is } c_{h_S} \text{ Lipschitz} \\ \mathcal{N}(c(K), \rho, \epsilon) & \text{if } h_S = \tilde{h}_S \circ c, \tilde{h}_S \text{ is } c_{h_S} \text{ Lipschitz, and } c \text{ is a canon.} \\ \mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon) & \text{if } h_S \text{ is } c_{h_S} \text{ Lipschitz and } \mathbb{G} \text{ invariant} \end{cases} \tag{1}$$

*Proof idea.* The claim for a Lipschitz function $h_S$ without any invariant structure follows immediately from Theorem 1.1. For a canonized function $h_S = \tilde{h}_S \circ c$, the claim follows from applying Theorem 1.1 to the domain $X = c(K)$ and the function $\tilde{h}_S$. The claim for $h_S$ which is both Lipschitz and invariant follows from the fact that due to invariance $h_S, f$ can be identified with functions $\hat{h}_S, \hat{f} : K/\mathbb{G} \to Y$ satisfying $\hat{h}_S([x]) = h_S(x), \hat{f}([x]) = f(x), \forall x \in K$, and the Lipschitz constant remains unchanged after the identification. □

For each of the three model types considered in Proposition 2.1, a different value of $\mathcal{N}(h_S)$ is provided. The following simple lemma establishes the relationship between these three values:

**Proposition 2.2.** *For any symmetry space $(K, \rho, \mathbb{G})$, and any $\epsilon > 0$, the following inequality holds:*

$$\mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon) \leq \mathcal{N}(c(K), \rho, \epsilon) \leq \mathcal{N}(K, \rho, \epsilon) \tag{2}$$

*Proof.* The inequality $\mathcal{N}(c(K), \rho, \epsilon) \leq \mathcal{N}(K, \rho, \epsilon)$ is immediate since $c(K) \subseteq K$. The inequality $\mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon) \leq \mathcal{N}(c(K), \rho, \epsilon)$ follows since the quotient map $q(x) = [x]$ is 1-Lipschitz and $q(c(K)) = K/\mathbb{G}$. Hence, the image under $q$ of any $\epsilon$-cover of $c(K)$ is an $\epsilon$-cover of $K/\mathbb{G}$.

□

Proposition 2.1 and Proposition 2.2 lead to the following conclusion:

**Conclusion 1**

The generalization bounds of canonized models are at best equal to the bounds of invariant and group-averaged models, and at worst equal to the bounds of non-invariant models.

Some reservations are in order: this conclusion concerns the upper bounds in Proposition 2.1, but these may not be tight. Moreover, the Lipschitz constants obtained by the different methods will not necessarily be identical. Nonetheless, these results lead us to conjecture that in terms of generalization, and overall performance, invariant and averaged models should outperform canonized models, which in turn will outperform non-invariant methods. In Section 5 we give empirical support for this claim, by providing several experiments in which group averaging outperforms canonized models. At the same time, we note that group averaging requires more computational resources and is infeasible for large groups. Invariant models also can lead to more complex and time consuming models. Thus, canonization still can have important advantages in practice. Moreover, the inequality in (2) may not be very large, in which case performance differences may be minor. The remainder of the paper focuses on this gap for different canonizations.

## 3 Canonizations: From Poor to Optimal

The inequality in Equation 2 gives us a theoretical way to measure the quality of a canonization. Informally, 'good canonizations' are those for which the upper bound is 'almost achieved'. Our goal in this section is to understand to what extent 'good canonizations' are possible.

We first show that the ratio between the three different bounds in 2 is at most the cardinality of the group. Thus, for small groups, the difference between canonization and invariant models is not expected to be very substantial. This is an immediate consequence from the following proposition:

**Proposition 3.1.** *[See proof in A.2] Let $(K, \rho, \mathbb{G})$ be a symmetry space, and assume $\mathbb{G}$ is finite. Then*

$$\mathcal{N}(K, \rho, \epsilon) \leq |\mathbb{G}| \cdot \mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon).$$

In the next step, we will study in what cases the upper and lower bounds in (2) are achieved. Formally:

**Definition 3.2.** Let $(K, \rho, \mathbb{G})$ be a symmetry space and let $\epsilon$ be a positive number. We say that $c$ is $(K, \epsilon)$ *optimal* if it attains the lower bound in (2), namely $\mathcal{N}(c(K), \rho, \epsilon) = \mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon)$. We say that $c$ is $(K, \epsilon)$ *poor* if it attains the upper bound in (2), namely $\mathcal{N}(c(K), \rho, \epsilon) = \mathcal{N}(K, \rho, \epsilon)$.

We will next show conditions for optimal and poor canonizations, and show how these conditions relate to the continuity of the canonization.

**Optimal canonizations:** First, we prove that canonizations which are isometries are optimal. We say that a canonization $c : K \to K$ is an *isometry* if $\rho(c(x), c(y)) = \rho_{\mathbb{G}}([x], [y])$ for all $x, y \in K$.

**Proposition 3.3.** *[proof in A.2] Any isometric canonization $c : K \to K$ is $(K, \epsilon)$ optimal for all $\epsilon > 0$.*

We next use this claim to give several examples of optimal canonizations:

**Example 3.4.** The following canonizations are isometries (see proof in A.2)

1. The canon. $c(x) = |x|$ with respect to the action of $\{-1, 1\}$ on $\mathbb{R}$ and the standard metric.
2. The canon. $\mathrm{sort} : \mathbb{R}^n \to \mathbb{R}^n$ with respect to the action of permutations and a metric $\rho$ induced by a $p$ norm.
3. The centralization canon. defined for point clouds in $\mathbb{R}^{d\times n}$ with respect to the action of translation by vectors in $\mathbb{R}^d$, with the metric $\rho$ induced by the Frobenius norm.

In Proposition A.3 in the appendix, we prove that a canonization is isometric if and only if it is 1-Lipschitz in the standard sense. As a result, in the setting where there can be no continuous canonizations, such as those discussed in [7, 2], there can be no isometric canonizations.

**Poor canonizations:** We next show that discontinuities in canonizations can lead to poor canonizations. This occurs under two conditions: firstly, the canonization is 'strongly discontinuous' in the sense that it has $|\mathbb{G}|$ partial limits. Secondly, the domain of the group action is chosen adversarially to exploit the discontinuity of the canonization.

**Proposition 3.5.** *[proof in A.2] Let $(K, \rho, \mathbb{G})$ be a symmetry space, where $\mathbb{G}$ is a finite group. Let $c$ be a canonization that has $|\mathbb{G}|$ partial limits at a point $x \in K$. Then, for any sufficiently small $\epsilon > 0$, there exists a $\mathbb{G}$-invariant set $A \subseteq K$ such that*

$$\mathcal{N}(c(A), \rho, \epsilon) = \mathcal{N}(A, \rho, \epsilon) = |\mathbb{G}| \cdot \mathcal{N}(A/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon).$$

Based on this proposition, we give two examples of poor canonizations.

**Example 3.6.** Consider the action of $\mathbb{G} = \{-1, 1\}$ on $\mathbb{R}$. Consider the canonizations

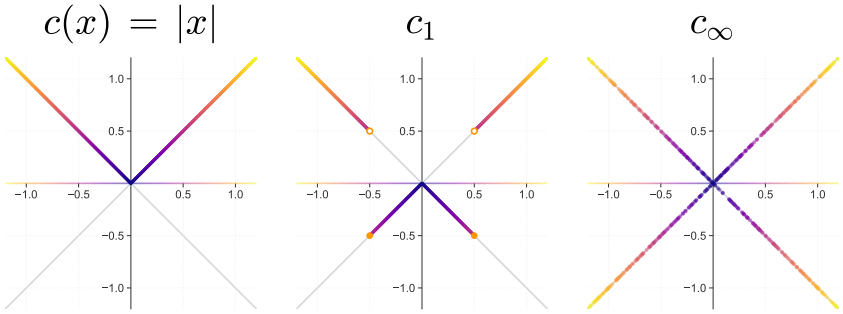
$$c_1(t) = \begin{cases} |t| & \text{if } |t| > 1/2 \\ -|t| & \text{if } |t| \le 1/2 \end{cases}, \quad c_\infty(t) = \begin{cases} |t| & \text{if } t \in \mathbb{Q} \\ -|t| & \text{if } t \notin \mathbb{Q} \end{cases}$$

visualized in Figure 1. Both of these canonizations are discontinuous and have a point with $|\mathbb{G}| = 2$ partial limits. Therefore, by Proposition 3.5 for every small enough $\epsilon > 0$ we can choose an adversarial domain for which the canonization is poor. For $c_1$ we can choose for every $\epsilon \in (0, \frac{1}{2})$ the set $K = B_\epsilon(-\frac{1}{2}) \cup B_\epsilon(\frac{1}{2})$ and then $\mathcal{N}(K, \rho, \epsilon) = \mathcal{N}(c_1(K), \rho, \epsilon) = 2$ and so the canonization is poor. In the canonization $c_\infty$ the discontinuities are much more extreme. In this case, we can choose a non-adversarial domain $K = [-1, 1]$, and the canonization will be poor for any $\epsilon > 0$ because the covering number of a set and its closure are the same, and in our case, the closure of $c_\infty(K)$ will be all of $K$.

We next present the second example.

**Example 3.7.** Consider the action of the permutation group $\mathbb{G} = S_n$ on the space of matrices $\mathbb{R}^{d\times n}$ by permuting the columns of the matrices. We call such matrices 'point clouds'. We assume that $n, d \ge 2$ (in the case $d = 1$, sorting is an isometry by Example 3.4). In this case, it is known that no continuous canonization exists [7]. A natural (discontinuous) canonization in this scenario is lexicographical sorting $c_{\text{lex}}$, which permutes the columns of the matrices so that the first row of the matrix is sorted from small to large, and in the event of ties in the first row, sorts according to the second row, and continues in this fashion until a permutation is uniquely defined. This canonization

Figure 1: *Visualization of three canonizations for the action of $\mathbb{G} = \{-1, 1\}$ on $\mathbb{R}$: the natural canonization $c(x) = |x|$, which is shown to be optimal (Example 3.4), and two discontinuous canonizations, which on an appropriate domain are poor (Example 3.6).*




has $n! = |\mathbb{G}|$ partial limits at any matrix $M \in \mathbb{R}^{d\times n}$ satisfying that the matrix does not contain two identical columns, but $M_{1,1} = M_{1,2} = \ldots = M_{1,n}$. Thus, according to Proposition 3.5, for small enough $\epsilon > 0$ there exists an adversarial domain for which the canonization is poor.

**Conclusion 2**

We show that isometric canonizations are optimal, while discontinuous canonizations can yield poor canonizations. These results give a rigorous justification for the importance of continuous canonizations.

## 4 Case study: Permutation Canonizations

In Example 3.7, we showed that lexicographical sorting is a poor canonization on an adversarial domain. In this section, we analyze the covering numbers of lexicographical sorting when considering a fixed, 'natural' domain of point clouds $K = [0, 1]^{d\times n}$. Throughout this section, we will consider the action of the group $\mathbb{G} = S_n$, and the distance induced by the element-wise infinity norm $\rho_\infty(X, Y) = \|X - Y\|_\infty$. We will show that (for fixed $\epsilon, d$) the covering number of Lexsort grows exponentially in $n$, the number of points in the point cloud, while the covering number of the quotient space only grows polynomially in $n$. Moreover, we will consider an alternative canonization based on Hilbert curves, and show that its covering numbers also grow only polynomially in $n$. The significance of this result is that it gives a theoretical justification for the choice of Point Transformer V3 [30] to use Hilbert canonizations for point cloud serialization.

We begin by proving the exponential growth of Lexsort canonizations:

**Lemma 4.1.** *[Proof in A.3] Let $k, d, n \geq 2$ be natural numbers, and set $\epsilon = 1/(2k)$, $K = [0, 1]^{d\times n}$ and $\mathbb{G} = S_n$. Then the covering number of the lexicographical sorting canonization satisfies*

$$\left(\frac{1}{2\epsilon}\right)^{(d-1)\cdot n+1} \leq \mathcal{N}(c_{\text{lex}}(K), \rho_\infty, \epsilon). \tag{3}$$

*Proof idea.* The Lexsort canonization is ineffective on the set of its 'severe discontinuities'- point clouds whose first row is constant. These point clouds form a $(d-1)\cdot n+1$ dimensional hypercube. The covering number of $c_{\text{lex}}(K)$ is lower bounded by the covering number of this hypercube, which is the left-hand side of (3). □

Next, we prove that for fixed $\epsilon, d$, the quotient space's covering number grows polynomially in $n$:

**Lemma 4.2.** *[Proof in A.3] Let $n, d$ be natural numbers and $\epsilon \in (0, 1)$. Set $K = [0, 1]^{d\times n}$ and $\mathbb{G} = S_n$. Then the covering number of the quotient space satisfies*

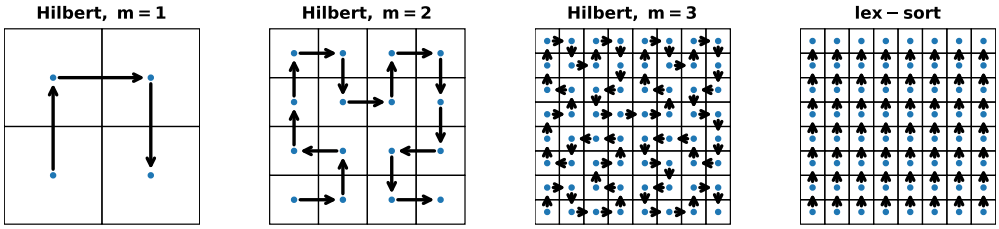
$$\mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon) \leq \binom{n + k^d - 1}{n}, \text{ where } k = \left\lceil \frac{1}{2\epsilon} \right\rceil. \tag{4}$$

*Proof idea.* Set $k = \lceil 1/(2\epsilon) \rceil$. Partition $[0,1]^d$ into $k^d$ hypercubes of side length $1/k$ and radius $1/(2k) \leq \epsilon$. Using their centers for each of the $n$ points yields an $\epsilon$-cover of $K = [0,1]^{d \times n}$. Let $C$ denote the collection of all point clouds of $n$ points whose points are centers of these hypercubes. The image of $C$ under the quotient map is an $\epsilon$ cover $q(C)$ of $K/\mathbb{G}$. Note that points in $C$ which are permutations of each other are mapped by $q$ to the same orbit. Therefore, the cardinality of the cover $q(C)$ is the number of equivalence classes with respect to the relation on $C$ defined by the group: $c_1 \sim c_2 \iff \exists \tau \in S_n : c_1 = \tau \cdot c_2$. As explained in more detail in the proof, the number of equivalence classes is the combinatorial expression in (4). □

### 4.1 Hilbert Curve Canonization

Finally, we will provide an intuitive explanation of the Hilbert canonization and formulate the claim that its covering number exhibits only polynomial growth with respect to the number of points in the point cloud. For a full definition of the canonization and proof, see Appendix A.4.

Figure 2: *The Hilbert curves $H_m$ for $d = 2$ and $m = 1, 2, 3$ induce an ordering on a grid in $[0,1]^d$ with $2^{md}$ points. On the right, we see the discontinuous Lexsort ordering on the grid.*




The Hilbert curve is a space-filling curve, mapping the unit interval $[0,1]$ continuously onto a hypercube $[0,1]^d$. It is defined as a limit of curves $H_m : [0,1] \to [0,1]^d$ which are not onto but whose image becomes more and more 'dense' in $[0,1]^d$ as $m$ grows, as shown in Figure 2 for the case where $d = 2$. The curves $H_m$ can be used to define an order on $[0,1]^d$ (at least on points in the image of the curve) and thus induce a canonization. Intuitively, the advantage of this order over the Lexsort ordering is that adjacently ordered points are mapped to nearby points in the ambient space, due to the continuity of the Hilbert curve (see Figure 2).

However, while previous work [30, 6] cited this continuity as justification for Hilbert ordering, a formal justification of why this continuity is important for learning has not yet been provided. We address this caveat by showing the polynomial growth of the covering number of Hilbert canonizations:

**Theorem 4.3.** *[Proof in A.4] Let $d, n, m \geq 2$ be natural numbers, let $\epsilon \in (0,1)$, and assume that $\epsilon > 2^{-m-1}$. Set $K = [0,1]^{d \times n}$ and $\mathbb{G} = S_n$. Then the covering number of the Hilbert canonization $c_m$ satisfies*

$$\mathcal{N}(c_m(K), \rho_\infty, \epsilon) \leq \binom{n + \lceil 1/(2\delta) \rceil - 1}{n}, \text{ where } \delta = \left( \frac{\epsilon - 2^{-m-1}}{4} \right)^d$$

*Proof idea.* When restricting to $n$-tuples in $K$ which are all in the image of $H_m$, the Hilbert canonization can be written as $H_m \circ \mathrm{sort} \circ H_m^{-1}$. Since sort (Example 3.4) is an isometry, the image of $\mathrm{sort} \circ H_m^{-1}$ can be covered efficiently. This cover can then be pushed forward to cover the image of $H_m \circ \mathrm{sort} \circ H_m^{-1}$, using the fact that $H_m$ is Hölder continuous. □

Theorem 4.3 shows that the covering number of the Hilbert canonization grows at worst polynomially in $n$, for fixed $\epsilon > 0$. This is in contrast with the exponential growth of the sorting canonization, proving a substantial gap between them.

**Conclusion 3**

Our analysis proves that the covering number of Hilbert canonizations is far superior to Lexsort canonization, giving a theoretical justification for its superior performance in practice.

## 5 Experiments

We conduct several experiments to verify our theoretical insights and their manifestation in practical learning scenarios. We focus on validating the hierarchy between group averaging, canonization, and non-invariant models (Conclusion 1), and on the hierarchy between Hilbert and Lexsort canonizations (Conclusion 3).

**Coverage on ModelNet** In our first experiment, we considered the ModelNet [31] point cloud classification task. Our first goal was to examine whether the covering-number hierarchy predicted by our theory is reflected in the geometry of real ModelNet point clouds, when comparing the quotient metric, Hilbert and Lexsort canonizations, and the original Euclidean metric.

As we do not have access to the 'true point cloud' domain $K$, but only to the train and test sets, we do not compute covering numbers, but a related notion we call the coverage. For given sets $T$ (representing the test set) and $R$ (the train set), we define the coverage $\epsilon(R, T)$ to be the smallest $\epsilon$ such that $R$ is an $\epsilon$ cover of $T$. We compute the coverage by choosing, for each test sample $t \in T$, the minimum distance from $t$ to all training samples (for computational efficiency, we consider only samples with the same label as $t$). We call this number $q_t$. The coverage is then given by $\epsilon(R, T) = \max_{t \in T} q_t$. We call this measure the coverage or *max coverage*. We also consider a more robust measure, the *mean coverage* $\epsilon_{\text{mean}}(R, T) = \text{mean}\{q_t | t \in \text{Test}\}$.

Table 1: *Estimates of covering number on ModelNet via mean and max coverage, when using Euclidean distances, Lexsort or Hilbert canonizations, and the group distance.*

| | Mean coverage ↓ | | Max coverage ↓ | |
|---|---|---|---|---|
| **Method** | **ModelNet10** | **ModelNet40** | **ModelNet10** | **ModelNet40** |
| Euclidean distance | 0.4545 | 0.3823 | 0.629 | 0.7304 |
| Lexsort distance | 0.3770 | 0.2989 | 0.5407 | 0.5550 |
| Hilbert distance | 0.2013 | 0.1951 | 0.5283 | 0.5490 |
| Group distance | **0.0917** | **0.0958** | **0.2742** | **0.4880** |

We compute the max and mean coverage on ModelNet10 and ModelNet40, using four different distances: (i) the mean pointwise Euclidean distance, the Euclidean distance after applying (ii) Lexsort or (iii) Hilbert canonizations to the train and test sets, and (iv) the quotient metric $\rho_{\mathbb{G}}$ (in this case, the Wasserstein distance). The results, shown in Table 1, show that the theoretical covering number hierarchy we derived is also preserved when considering coverage of real data: the coverage of $\rho_{\mathbb{G}}$ is the lowest, followed by Hilbert, Lexsort, and finally standard Euclidean distance.

**Classification on ModelNet** Next, we examine whether the hierarchy in covering numbers between the different methods translates into a corresponding hierarchy in accuracy and generalization error. We use a simple, purely non-invariant backbone, a Global MLP, and compare its performance (i) without canonization, (ii) with Lexsort canonization, and (iii) with Hilbert canonization at resolution 12. We additionally evaluate DeepSets as an invariant Lipschitz model (exact group averaging is computationally infeasible here due to the prohibitively large permutation group). The results, shown in Table 2, reveal the expected hierarchy: Hilbert outperforms Lexsort, which in turn outperforms the plain MLP, while the invariant DeepSets model achieves the strongest overall performance. The generalization error column (defined as Train Acc. − Test Acc.) indicates that these performance differences are largely due to improved generalization.

To further verify the generality of our findings, we also trained the three Global MLP alternatives, together with DeepSets, on smaller samples from the ModelNet40 dataset. We find that the DeepSets $>$ Hilbert $>$ Lexsort $>$ no canon. hierarchy is preserved across all training-set sizes, as shown in Table 9 in the appendix.

Table 2: *On the ModelNet classification task, we compare a vanilla MLP without canonization, the same architecture with Lexsort and Hilbert canonization, and DeepSets as an invariant Lipschitz model. As predicted by our theory, improved invariance leads to better generalization and overall accuracy.*

| | ModelNet10 | | ModelNet40 | |
|---|---|---|---|---|
| Model | Test Acc ↑ | Gen. error ↓ | Test Acc ↑ | Gen. error ↓ |
| MLP (no canon.) | 0.573 | 0.427 | 0.441 | 0.559 |
| MLP (Lexsort) | 0.762 | 0.215 | 0.692 | 0.256 |
| MLP (Hilbert) | 0.782 | 0.116 | 0.742 | 0.082 |
| DeepSets | **0.813** | **0.077** | **0.755** | **0.039** |

We note that this experiment is only illustrative; the results attained by simple vanilla MLPs, even with Hilbert canonization, are far from state-of-the-art. In the context of our theory, this is consistent with the fact that successful models for this task are typically permutation invariant, and hence their generalization is governed by the covering number of the quotient space, which is always no larger than the covering number induced by canonization (Conclusion 1).

We next conduct further experiments to corroborate Conclusion 1, which predicts that, in terms of generalization, invariant models or group-averaged predictors should enjoy an advantage over canonized models, which in turn should improve over models that ignore the underlying invariance.

**Group averaging on PCA frames** Our first experiment is still with the ModelNet40 classification task, but while the previous experiments investigate canonization w.r.t. permutations, and implicitly rely on the fact that ModelNet models are aligned w.r.t rotations, in this experiment we use the DeepSets [33] model to deal with permutation invariance, and will focus on rotation invariance: we will assume the rotational degree of freedom of point clouds in ModelNet is initially unknown, and is fixed using the PCA axes as suggested in [23]. Assuming the point cloud singular values are pairwise distinct, this procedure determines the point cloud orientation uniquely up to a $\{-1, 1\}^3$ symmetry.

We consider four ways to deal with this symmetry: (i) Pure PCA, where the symmetry is ignored (ii) Skewness canonization, where the global sign of the $x, y$ and $z$ coordinates is chosen so that their third moment is positive (iii) group averaging and (iv) group augmentation where in each epoch a random group element is applied to the input. This can be seen as an approximation of group averaging. The results of this experiment, detailed in Table 3 demonstrate a clear performance hierarchy. Group averaging achieves the highest overall accuracy, while group augmentation and skewness canonization yield comparable final results, with all three methods significantly outperforming the pure PCA baseline. We note that, as highlighted in Figure 3 in the appendix, the group augmentation approach suffers from a noticeably slower convergence rate compared to the other three methods.

**Group averaging for image classification** Next, we check our approach on image classification. We take the rotated MNIST [15] dataset, and consider the action of the $p4$ group, which is the four-element group generated by $90$ degree rotations. We consider 4 approaches: simple CNN, the canonization from [14] with random or learned weights (named CN(p4 frozen)-CNN and CN(p4)-CNN, respectively), and a group average over the $4$ elements of the group. As shown in Table 4, the canonizations outperform the simple CNN, but are outperformed by group averaging, as predicted in Conclusion 1. We also see in the table that the accuracy is correlated with the mean coverage of the different methods.

Table 3: *Test accuracy on ModelNet40 with a DeepSets model with orientation determined by PCA, and sign ambiguity handled via averaging, sign augmentation, canonization, or Pure PCA (ignoring the ambiguity).*

| **Method** | **Accuracy ↑** |
|---|---|
| Sign Averaging | **0.760** $\pm$ 0.003 |
| Sign Augmentation | 0.742 $\pm$ 0.002 |
| Skewness Canonization | 0.742 $\pm$ 0.005 |
| Pure PCA | 0.729 $\pm$ 0.006 |

Table 4: *Accuracy and mean coverage on Rotated-MNIST, which is invariant with respect to the group $p4$ of $90$ rotations. We consider a non-invariant CNN, two canonizations from [14], and group averaging.*

| **Model** | **Test Acc. ↑** | **Coverage ↓** |
|---|---|---|
| CNN | $94.7 \pm 0.12$ | 0.185 |
| CN(p4)-CNN | $96.5 \pm 0.13$ | 0.175 |
| CN(p4 frozen)-CNN | $95.5 \pm 0.15$ | 0.176 |
| AvgCNN | **97.3** $\pm 0.12$ | **0.167** |

**Canonization in deep weight spaces.** We evaluate canonization on INR classification, where each image is represented by the weights of an implicit neural representation (INR), and the task is to predict the image class directly from these weights [21]. We compare a simple MLP, a canonized MLP, and DWSNet, a specialized permutation-invariant weight-space model [21]. We consider the MNIST-INR and Fashion-MNIST-INR datasets [21]; details of our canonization procedure are given in Appendix C.7.

For completeness, we reran the original code for the MLP and DWSNet baselines and report the results obtained from our runs. The results are shown in Table 5. We observe that DWSNet outperforms the canonized MLP, which in turn outperforms the plain MLP, as predicted by Conclusion 1.

Table 5: *INR classification accuracy on MNIST and Fashion-MNIST. We compare a non-invariant MLP, a canonized MLP, and DWSNet, a specialized permutation-invariant weight-space architecture.*

| **Method** | **MNIST-INR** $\uparrow$ | **Fashion-MNIST-INR** $\uparrow$ |
|---|---|---|
| MLP | $16.34 \pm 0.61$ | $14.41 \pm 0.39$ |
| MLP + Canonization | $31.06 \pm 0.46$ | $40.02 \pm 0.23$ |
| DWSNet | $\mathbf{86.66} \pm 0.73$ | $\mathbf{65.53} \pm 0.29$ |

**Group Averaging for Spectral Embeddings** In the Appendix Section B we present an additional experiment showing the advantage of averaging over previously proposed canonizations, in the context of sign ambiguity present in spectral embeddings for graph learning.

## 6 Conclusion

In this paper, we establish a methodology for comparing different canonizations with group averaging and other invariant models by studying their covering number. We show the covering number of canonizations is at best the covering number of the quotient space (for isometric canonizations), and at worst the covering number of the original domain (for certain discontinuous canonizations). These results suggest that group-averaged models are preferable when computationally feasible. We also show that the covering number of Hilbert canonizations for point clouds is substantially better than that of Lexsort canonizations.

**Limitations and Future Work** A limitation of our analysis is that it relies on upper bounds on the generalization error (Proposition 2.1), which may not be tight. Nonetheless, our experiments indicate that lower covering numbers do correlate with better generalization and learning. Looking towards future developments, a central contribution of this work is the demonstration that a rigorous theoretical comparison between competing canonization methods can be established via their covering numbers. This framework complements the traditional focus on a model's ability to distinguish between complex symmetric examples, instead highlighting properties that arguably exert a more significant influence on the ultimate success of the learning process. In future work, we intend to apply our framework to other canonizations and extend it to enable a rigorous treatment of the generalization of frame averaging.

## Acknowledgments and Disclosure of Funding

**Funding.** This work was supported by the Israel Science Foundation (grant no. 272/23) and the DFG–ISF (grant no. 3597/26). N.D. and S.H. thank Jonathan W. Siegel and Hannah Lawrence for helpful discussions.

**Competing interests.** The authors declare no competing financial interests related to this work.

## References


[1] Josep Arús-Pous, Simon V. Johansson, Oleksii Prykhodko, Esben J. Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Randomized SMILES strings improve the quality of molecular generative models. *Journal of Cheminformatics*, 11(1):71, 2019.

[2] Justin Baker, Shih-Hsin Wang, Tommaso De Fernex, and Bao Wang. An explicit frame construction for normalizing 3D point clouds. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 2456–2473. PMLR, 21–27 Jul 2024.

[3] Esben Jannik Bjerrum. Smiles enumeration as data augmentation for neural network modeling of molecules. *ArXiv*, abs/1703.07076, 2017.

[4] Sergey Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society, 2019.

[5] Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale?, 2025.

[6] Wanli Chen, Xinge Zhu, Guojin Chen, and Bei Yu. Efficient point cloud analysis using hilbert curve. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 730–747, Cham, 2022. Springer Nature Switzerland.

[7] Nadav Dym, Hannah Lawrence, and Jonathan W Siegel. Equivariant frames and the impossibility of continuous canonicalization. In *International Conference on Machine Learning*, pages 12228–12267. PMLR, 2024.

[8] Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[9] Billy J Franks, Christopher Morris, Ameya Velingker, and Floris Geerts. Weisfeiler-leman at the margin: when more expressivity matters. In *Proceedings of the 41st International Conference on Machine Learning*, pages 13885–13926, 2024.

[10] Benjy Friedmann and Michael Werman. Canonnet: Canonical ordering and curvature learning for point cloud analysis, 2025.

[11] Zhijian He and Art B Owen. Extensible grids: uniform sampling on a space filling curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):917–931, 2016.

[12] Snir Hordan, Maya Bechler-Speicher, Gur Lifshitz, and Nadav Dym. Spectral graph neural networks are incomplete on graphs with a simple spectrum. *arXiv preprint arXiv:2506.05530*, 2025.

[13] Michael Ito, Danai Koutra, and Jenna Wiens. Random search neural networks for efficient and expressive graph learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[14] Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, pages 15546–15566. PMLR, 2023.

[15] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480, 2007.

[16] Ya-Wei Eileen Lin and Ron Levie. Adaptive canonicalization with application to invariant anisotropic geometric networks. In *The Fourteenth International Conference on Learning Representations*, 2026.

[17] Yuchao Lin, Jacob Helwig, Shurui Gui, and Shuiwang Ji. Equivariance via minimal frame averaging for more symmetries and efficiency. In *Forty-first International Conference on Machine Learning*, 2024.

[18] George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonicalization perspective on invariant and equivariant learning. *Advances in Neural Information Processing Systems*, 37:60936–60979, 2024.

[19] George Ma, Yifei Wang, and Yisen Wang. Laplacian canonization: A minimalist approach to sign and basis invariant spectral embedding. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[20] Sohir Maskey, Gitta Kutyniok, and Ron Levie. Generalization bounds for message passing networks on mixture of graphons. *SIAM Journal on Mathematics of Data Science*, 7(2):802–825, 2025.

[21] Aviv Navon, Aviv Shamsian, Idan Achituve, Ethan Fetaya, Gal Chechik, and Haggai Maron. Equivariant architectures for learning in deep weight spaces. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25790–25816. PMLR, 23–29 Jul 2023.

[22] Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[23] Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *arXiv preprint arXiv:2110.03336*, 2021.

[24] Jonathan W. Siegel, Snir Hordan, Hannah Lawrence, Ali Syed, and Nadav Dym. Quantitative approximation rates for group equivariant learning, 2026.

[25] Richard P Stanley. Enumerative combinatorics volume 1 second edition. *Cambridge studies in advanced mathematics*, 2011.

[26] Behrooz Tahmasebi and Stefanie Jegelka. Generalization bounds for canonicalization: A comparative study with group averaging. In *The Thirteenth International Conference on Learning Representations*, 2025.

[27] Behrooz Tahmasebi and Stefanie Jegelka. Regularity in canonicalized models: A theoretical perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 4789–4797. PMLR, 2025.

[28] Antonis Vasileiou, Stefanie Jegelka, Ron Levie, and Christopher Morris. Survey on generalization theory for graph neural networks. *arXiv preprint arXiv:2503.15650*, 2025.

[29] David Weininger, Arthur Weininger, and Joseph L. Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.*, 29:97–101, 1989.

[30] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024.

[31] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[32] Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

[33] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

[34] Cai Zhou, Zijie Chen, Zian Li, Jike Wang, Kaiyi Jiang, Pan Li, Rose Yu, Muhan Zhang, Stephen Bates, and Tommi Jaakkola. Rethinking diffusion models with symmetries through canonicalization with applications to molecular graph generation, 2026.

# A Proofs

In this section, we provide proofs and claims that are missing in the main text.

## A.1 Generalization and Covering Numbers

Our generalization results are based on the results of [32]: a combination of Theorem 1 and Example 4 from [32] yields the following theorem:

**Theorem A.1** (Based on [32])**.** *Let $(K, \rho)$ be a metric space, and let $X \subseteq K$ be totally bounded. Suppose that $0 \leq l(h_S, x) \leq M$ for all $x \in X$, and that $l(h_S, \cdot)$ is Lipschitz continuous on $X$ with Lipschitz constant $c(S)$, i.e.,*

$$|l(h_S, x_1) - l(h_S, x_2)| \leq c(S)\rho(x_1, x_2), \quad \forall x_1, x_2 \in X.$$

*Assume that the training sample set $S$ is generated by $n$ IID draws from a Borel probability distribution $\mu$ on $X$. Then, for any $\delta, \epsilon > 0$, with probability at least $1 - \delta$,*

$$|\mathcal{L}_{\exp}(h_S) - \mathcal{L}_{\text{emp}}(h_S)| \leq 2c(S)\epsilon + M\sqrt{\frac{2\mathcal{N}(X, \rho, \epsilon)\ln 2 + 2\ln(1/\delta)}{n}}. \tag{5}$$

We now explain how Theorem 1.1 from the main text follows from Theorem A.1. We recall Theorem 1.1:

**Theorem 1.1.** *[Proof in Appendix A.1, based on [32] ] Let $(K, \rho)$ and $(Y, \rho_Y)$ be metric spaces, and assume that $X \subseteq K$ is totally bounded. Let $\mu$ be a Borel probability distribution on $X$. Let $f : X \to Y$ be a $c_f$-Lipschitz function, let $h_S : X \to Y$ be a $c_{h_S}$-Lipschitz function, and let $\ell : Y \times Y \to [0, M]$ be a $c_\ell$-Lipschitz function. Then for any $0 < \epsilon$ and $0 < \delta < 1$ and natural $n$, for a training sample set $S$ generated by $n$ IID draws from $\mu$, with probability of at least $1 - \delta$,*

$$|\mathcal{L}_{\exp}(h_S) - \mathcal{L}_{\text{emp}}(h_S)| \leq 2c_\ell(c_{h_S} + c_f)\epsilon + M\sqrt{\frac{2\mathcal{N}(X, \rho, \epsilon)\ln 2 + 2\ln(1/\delta)}{n}}$$

*Proof.* We will apply Theorem A.1, where to apply it to the setting discussed in Theorem 1.1, we set

$$l(h_S, x) = \ell(h_S(x), f(x))$$

We equip $Y \times Y$ with the product metric

$$\rho_{Y\times Y}\big((y_1, y_2), (y_1', y_2')\big) = \rho_Y(y_1, y_1') + \rho_Y(y_2, y_2').$$

We note that $l$ is Lipschitz with a Lipschitz constant of $c_\ell(c_{h_S} + c_f)$ as the following shows:

$$\begin{aligned}
|l(h_S, x) - l(h_S, x')| &= |\ell(h_S(x), f(x)) - \ell(h_S(x'), f(x'))| \\
&\leq c_\ell\left(\rho_Y(h_S(x), h_S(x')) + \rho_Y(f(x), f(x'))\right) \\
&\leq c_\ell\left(c_{h_S}\rho(x, x') + c_f \cdot \rho(x, x')\right) \\
&= c_\ell(c_{h_S} + c_f)(\rho(x, x'))
\end{aligned}$$

Now, to conclude the proof, choose any $\epsilon, \delta > 0$. Since $l$ is $c_\ell(c_{h_S} + c_f)$ Lipschitz, we have

$$|\mathcal{L}_{\exp}(h_S) - \mathcal{L}_{\text{emp}}(h_S)| \overset{(5)}{\leq} 2c_\ell(c_{h_S} + c_f)\epsilon + M\sqrt{\frac{2\mathcal{N}(X, \rho, \epsilon)\ln 2 + 2\ln(1/\delta)}{n}}$$

which proves Theorem 1.1. □

Next, we prove Proposition 2.1.

**Proposition 2.1.** *[Proof in Appendix A.1] Let $(K, \rho, \mathbb{G})$ be a symmetry space. Let $(Y, \rho_Y)$ be a metric space, and assume $K$ is totally bounded. Let $\mu$ be a Borel probability distribution on $K$. Let $f : K \to Y$ be a $c_f$ Lipschitz function* ***which is*** *$\mathbb{G}$* ***invariant****, and let $\ell : Y \times Y \to [0, M]$ be a $c_\ell$ Lipschitz function , then for any $0 < \epsilon$ and $0 < \delta < 1$ and natural $n$, for a training sample set $S$ generated by $n$ IID draws from $\mu$, with probability of at least $1 - \delta$, we have*

$$|\mathcal{L}_{\exp}(h_S) - \mathcal{L}_{\text{emp}}(h_S)| \leq 2c_\ell(c_{h_S} + c_f)\epsilon + M\sqrt{\frac{2\mathcal{N}(h_S)\ln 2 + 2\ln(1/\delta)}{n}}$$

*where*

$$\mathcal{N}(h_S) = \begin{cases} \mathcal{N}(K, \rho, \epsilon) & \text{if } h_S \text{ is } c_{h_S} \text{ Lipschitz} \\ \mathcal{N}(c(K), \rho, \epsilon) & \text{if } h_S = \tilde{h}_S \circ c, \tilde{h}_S \text{ is } c_{h_S} \text{ Lipschitz, and } c \text{ is a canon.} \\ \mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon) & \text{if } h_S \text{ is } c_{h_S} \text{ Lipschitz and } \mathbb{G} \text{ invariant} \end{cases} \tag{1}$$

*Proof.* The first case where $h_S$ is Lipschitz (but not invariant) follows directly from Theorem 1.1. Next, we consider the second case $h_S = \tilde{h}_S \circ c$, where $\tilde{h}_S$ is $c_{h_S}$ Lipschitz. We note that since $\tilde{h}_S$ operates on $c(S)$, we can think of it as defined by the set $\tilde{S} = c(S)$, and so we can write $\tilde{h}_S := \tilde{h}_{\tilde{S}}$. This notation is more convenient for the proof.

We want to consider the domain to be $c(K)$, so consider the following distribution measure on $c(K)$:

$$\tilde{\mu}(A) = \mu(c^{-1}(A))$$

Note that

$$\forall x \in c(K), h_S(x) = \tilde{h}_{\tilde{S}}(x)$$

Next, for any fixed $n$ samples from $K$, which we denote by $S = \{x_1, x_2, \ldots, x_n\}$, denote its image by $\tilde{S} = \{c(x_1), \ldots, c(x_n)\}$. Then

$$\mathcal{L}_{\text{emp}}(h_S) = \frac{1}{n}\sum_{i=1}^{n} \ell(\tilde{h}_{\tilde{S}}(c(x_i)), f(x_i)) = \frac{1}{n}\sum_{i=1}^{n} \ell(\tilde{h}_{\tilde{S}}(c(x_i)), f(c(x_i))) = \mathcal{L}_{\text{emp}}(\tilde{h}_{\tilde{S}})$$

So, the empirical losses are the same, and next we show the expected losses are the same:

$$\begin{aligned} \mathcal{L}_{\exp}(\tilde{h}_{\tilde{S}}) &= \mathbb{E}_{x\sim\tilde{\mu}}\ell(\tilde{h}_{\tilde{S}}(x), f(x)) = \int_{c(K)} \ell(\tilde{h}_{\tilde{S}}(x), f(x))d\tilde{\mu} = \int_K \ell(\tilde{h}_{\tilde{S}}(c(x)), f(c(x)))d\mu \\ &= \int_K \ell(h_S(x), f(x))d\mu = \mathcal{L}_{\exp}(h_S) \end{aligned}$$

Since $c(K) \subseteq K$ and $K$ is totally bounded, $c(K)$ is totally bounded. We may therefore apply Theorem A.1 with $X = c(K)$, ambient space $K$, and distribution $\tilde{\mu}$. Since $\tilde{h}_{\tilde{S}}$ is $c_{h_S}$-Lipschitz, with probability at least $1 - \delta$ under $\tilde{\mu}$, we obtain:

$$|\mathcal{L}_{\exp}(\tilde{h}_{\tilde{S}}) - \mathcal{L}_{\text{emp}}(\tilde{h}_{\tilde{S}})| \leq 2c_\ell(c_{h_S} + c_f)\epsilon + M\sqrt{\frac{2\mathcal{N}(c(K), \rho, \epsilon)\ln 2 + 2\ln(1/\delta)}{n}}$$

Let us abbreviate the right hand side of the equation above by $\eta = 2c_\ell(c_{h_S} + c_f)\epsilon + M\sqrt{\frac{2\mathcal{N}(c(K),\rho,\epsilon)\ln 2 + 2\ln(1/\delta)}{n}}$. We then have

$$\begin{aligned} 1 - \delta &\leq \mathbb{P}_{\tilde{\mu}^n}\left(\left|\mathcal{L}_{\exp}(\tilde{h}_{\tilde{S}}) - \mathcal{L}_{\text{emp}}(\tilde{h}_{\tilde{S}})\right| \leq \eta\right) \\ &= \mathbb{P}_{\mu^n}\left(\left|\mathcal{L}_{\exp}(\tilde{h}_{\tilde{S}} \circ c) - \mathcal{L}_{\text{emp}}(\tilde{h}_{\tilde{S}} \circ c)\right| \leq \eta\right) \\ &= \mathbb{P}_{\mu^n}\left(|\mathcal{L}_{\exp}(h_S) - \mathcal{L}_{\text{emp}}(h_S)| \leq \eta\right). \end{aligned}$$

which concludes the proof of the second case.

In the third case, where $h_S$ is $c_{h_S}$-Lipschitz and invariant, we consider the domain $K/\mathbb{G}$ and define $\pi : K \to K/\mathbb{G}$ by $\pi(x) = [x]$. Define the following distribution on $K/\mathbb{G}$:

$$\tilde{\mu}(A) = \mu(\pi^{-1}(A)).$$

Define $\tilde{S} = \pi(S)$ and

$$\tilde{h}_{\tilde{S}}, \tilde{f} : K/\mathbb{G} \to Y$$

by

$$\tilde{h}_{\tilde{S}}([x]) = h_S(x), \qquad \tilde{f}([x]) = f(x).$$

Note that by invariance, these functions are well defined.

Next, for any sample set $S = \{x_1, \ldots, x_n\}$ from $K$, let

$$\tilde{S} = \{\pi(x_1), \ldots, \pi(x_n)\}.$$

Then

$$\begin{aligned}
\mathcal{L}_{\text{emp}}(h_S) &= \frac{1}{n} \sum_{i=1}^{n} \ell(h_S(x_i), f(x_i)) \\
&= \frac{1}{n} \sum_{i=1}^{n} \ell\big(\tilde{h}_{\tilde{S}}(\pi(x_i)), \tilde{f}(\pi(x_i))\big) \\
&= \mathcal{L}_{\text{emp}}(\tilde{h}_{\tilde{S}}).
\end{aligned}$$

Thus, the empirical losses are the same. We next show that the expected losses are also equal:

$$\begin{aligned}
\mathcal{L}_{\text{exp}}(\tilde{h}_{\tilde{S}}) &= \mathbb{E}_{x \sim \tilde{\mu}} \ell(\tilde{h}_{\tilde{S}}(x), \tilde{f}(x)) \\
&= \int_{K/\mathbb{G}} \ell(\tilde{h}_{\tilde{S}}(x), \tilde{f}(x))\, d\tilde{\mu}(x) \\
&= \int_{K} \ell(\tilde{h}_{\tilde{S}}(\pi(x)), \tilde{f}(\pi(x)))\, d\mu(x) \\
&= \int_{K} \ell(h_S(x), f(x))\, d\mu(x) \\
&= \mathcal{L}_{\text{exp}}(h_S).
\end{aligned}$$

Since $K$ is totally bounded and the quotient map $\pi : K \to K/\mathbb{G}$ is 1-Lipschitz and surjective, the quotient space $K/\mathbb{G}$ is also totally bounded. Moreover, $\tilde{f}$ is $c_f$-Lipschitz and $\tilde{h}_{\tilde{S}}$ is $c_{h_S}$-Lipschitz (see Lemma 20 in [24]).

Therefore, all the assumptions of Theorem 1.1 are satisfied, and applying it to $\tilde{h}_{\tilde{S}}$ and $\tilde{f}$ on $K/\mathbb{G}$, we obtain that with probability at least $1 - \delta$,

$$\left|\mathcal{L}_{\text{exp}}(\tilde{h}_{\tilde{S}}) - \mathcal{L}_{\text{emp}}(\tilde{h}_{\tilde{S}})\right| \leq 2c_\ell(c_{h_S} + c_f)\epsilon + M\sqrt{\frac{2\mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon)\ln 2 + 2\ln(1/\delta)}{n}}.$$

Since

$$\left|\mathcal{L}_{\text{exp}}(\tilde{h}_{\tilde{S}}) - \mathcal{L}_{\text{emp}}(\tilde{h}_{\tilde{S}})\right| = \left|\mathcal{L}_{\text{exp}}(h_S) - \mathcal{L}_{\text{emp}}(h_S)\right|,$$

we conclude that with probability at least $1 - \delta$,

$$|\mathcal{L}_{\text{exp}}(h_S) - \mathcal{L}_{\text{emp}}(h_S)| \leq 2c_\ell(c_{h_S} + c_f)\epsilon + M\sqrt{\frac{2\mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon)\ln 2 + 2\ln(1/\delta)}{n}}.$$

□

We next state and prove a result mentioned in the main text: that the Lipschitz constant of a backbone model is preserved under averaging

**Proposition A.2.** *Let $(X, \rho, \mathbb{G})$ be a module, and assume that $\mathbb{G}$ is a compact group acting continuously on $X$, with Haar probability measure $\mu$. Let $Y$ be a Banach space and assume that $h : X \to Y$ is $c_h$ Lipschitz. Then the averaged function*

$$q(x) = \int_{\mathbb{G}} h(gx)\, d\mu(g)$$

*is $c_h$ Lipschitz as well.*

*Proof.* For every $x_1, x_2 \in X$,

$$\begin{aligned}\|q(x_1) - q(x_2)\| &= \left\| \int_{\mathbb{G}} h(gx_1) d\mu(g) - \int_{\mathbb{G}} h(gx_2) d\mu(g) \right\| \\ &\overset{\text{Jensen}}{\leq} \int_{\mathbb{G}} \|h(gx_1) - h(gx_2)\| \, d\mu(g) \\ &\leq \int_{\mathbb{G}} c_h \rho(gx_1, gx_2) d\mu(g) \\ &= \int_{\mathbb{G}} c_h \rho(x_1, x_2) d\mu(g) \\ &= c_h \rho(x_1, x_2)\end{aligned}$$

□

### A.2 Proofs for Section 3

We restate and prove Proposition 3.1.

**Proposition 3.1.** *[See proof in A.2] Let $(K, \rho, \mathbb{G})$ be a symmetry space, and assume $\mathbb{G}$ is finite. Then*

$$\mathcal{N}(K, \rho, \epsilon) \leq |\mathbb{G}| \cdot \mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon).$$

*Proof.* Let $C$ be a minimum-cardinality $\epsilon$-cover of $K/\mathbb{G}$ under $\rho_{\mathbb{G}}$, and define

$$\widehat{C} := \cup_{c \in C} c$$

i.e, $\widehat{C}$ is the union of all orbits in $C$. As each orbit has at most $|\mathbb{G}|$ elements it holds that $|\widehat{C}| \leq |\mathbb{G}| \cdot |C|$.

We claim that $\widehat{C}$ is an $\epsilon$-cover of $K$ under $\rho$. Let $x \in K$. Since $C$ covers $K/\mathbb{G}$, there exists an orbit $O \in C$ such that

$$\rho_{\mathbb{G}}([x], O) \leq \epsilon.$$

Choose $y \in O$. Since $\mathbb{G}$ is finite, the minimum in the definition of $\rho_{\mathbb{G}}$ is attained, so there exists $g \in \mathbb{G}$ such that

$$\rho(x, g \cdot y) \leq \epsilon.$$

By construction, $g \cdot y \in O \subseteq \widehat{C}$, hence $\widehat{C}$ covers $K$. Therefore,

$$\mathcal{N}(K, \rho, \epsilon) \leq |\widehat{C}| \leq |\mathbb{G}| \cdot |C| = |\mathbb{G}| \cdot \mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon).$$

□

We next prove Proposition 3.3.

**Proposition 3.3.** *[proof in A.2] Any isometric canonization $c : K \to K$ is $(K, \epsilon)$ optimal for all $\epsilon > 0$.*

*Proof.* As it always holds that $\mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon) \leq \mathcal{N}(c(K), \rho, \epsilon)$, we just need to prove the opposite direction.

Let $C$ be a cover of $K/\mathbb{G}$ with $\mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon)$ elements. For each orbit $O \in C$, choose a representative $x_O \in O$, and define

$$\hat{C} := \{c(x_O) \mid O \in C\}.$$

We will prove that $\hat{C}$ is an $\epsilon$ cover of $c(K)$.

Let $y \in c(K)$. Then by definition there exists $x \in K$ such that $y = c(x)$. By the definition of cover, there exists $O \in C$ such that

$$\rho_{\mathbb{G}}([x], O) \leq \epsilon.$$

Since $x_O \in O$, we have $[x_O] = O$, and since $c$ is an isometry,

$$\rho(y, c(x_O)) = \rho(c(x), c(x_O)) = \rho_{\mathbb{G}}([x], [x_O]) = \rho_{\mathbb{G}}([x], O) \leq \epsilon.$$

Hence $\hat{C}$ is an $\epsilon$ cover of $c(K)$, as desired.

□

**Example 3.4.** The following canonizations are isometries (see proof in A.2)

1. The canon. $c(x) = |x|$ with respect to the action of $\{-1, 1\}$ on $\mathbb{R}$ and the standard metric.
2. The canon. $\text{sort} : \mathbb{R}^n \to \mathbb{R}^n$ with respect to the action of permutations and a metric $\rho$ induced by a $p$ norm.
3. The centralization canon. defined for point clouds in $\mathbb{R}^{d\times n}$ with respect to the action of translation by vectors in $\mathbb{R}^d$, with the metric $\rho$ induced by the Frobenius norm.

*Proof.* The canonization $c(x) = |x|$ is an isometry as

$$\forall x, y \in \mathbb{R}, \rho_{\mathbb{G}}([x], [y]) = \min\{|x - y|, |x + y|\} = ||x| - |y|| = |c(x) - c(y)|$$

The fact that sorting is an isometry with respect to the metric $\rho_{\mathbb{G}}$ induced from a $p$ norm (which is the $p$-Wasserstein metric) is shown in Lemma 4.2 in [4].

Consider the action of $\mathbb{R}^d$ on $\mathbb{R}^{d\times n}$ by translating all columns of a matrix in $\mathbb{R}^{d\times n}$ by a vector in $\mathbb{R}^d$. Let $\rho$ be the metric induced from the Frobenius norm $\rho(X, Y) = \|X - Y\|_F$, and let $c$ be the centralization mapping $c(X) = X - \frac{1}{n}X \cdot 1_{n\times n}$. This mapping is a canonization and an isometry. To see it is an isometry, note that we have for any canonization $\rho_{\mathbb{G}}([X], [Y]) \leq \rho(c(X), c(Y))$, and we get the converse inequality by noting that $c$ is the orthogonal projection from the inner product space $\mathbb{R}^{d\times n}$ onto the subspace $\{X \in \mathbb{R}^{d\times n} | \quad X1_{n\times n} = 0\}$. Thus, let $g \in \mathbb{G}$ be a translation vector such that $\rho_{\mathbb{G}}([X], [Y]) = \rho(X, gY)$. Then

$$\rho_{\mathbb{G}}([X], [Y]) = \rho(X, gY) = \|X - g\cdot Y\|_F \geq \|c(X) - c(gY)\|_F = \|c(X) - c(Y)\|_F = \rho(c(X), c(Y))$$

□

We next formulate and prove a result mentioned in the main text

**Proposition A.3.** *Let $(K, \rho, \mathbb{G})$ be a module, and $c : K \to K$ a canonization. Then $c$ is an isometric canonization if and only if $c$ is 1-Lipschitz.*

*Proof.* Assume $c$ is an isometric canonization. Recall that this means that $\rho(c(x), c(y)) = \rho_{\mathbb{G}}([x], [y])$. This implies that $c$ is Lipschitz because

$$\rho(c(x), c(y)) = \rho_{\mathbb{G}}([x], [y]) \leq \rho(x, y), \quad \forall x, y \in K$$

In the other direction, assume that $c$ is 1-Lipschitz. This implies that for all $g \in \mathbb{G}$

$$\rho(c(x), c(y)) = \rho(c(gx), c(y)) \leq \rho(gx, y).$$

Since this is true for all $g \in \mathbb{G}$, we can minimize over $g$ and obtain the inequality $\rho(c(x), c(y)) \leq \rho_{\mathbb{G}}([x], [y])$. The converse of this inequality is always true for any canonization, namely

$$\rho(c(x), c(y)) \geq \rho_{\mathbb{G}}([x], [y])$$

and so we obtain that $c$ is an isometric canonization. □

We restate and prove 3.5.

**Proposition 3.5.** *[proof in A.2] Let $(K, \rho, \mathbb{G})$ be a symmetry space, where $\mathbb{G}$ is a finite group. Let $c$ be a canonization that has $|\mathbb{G}|$ partial limits at a point $x \in K$. Then, for any sufficiently small $\epsilon > 0$, there exists a $\mathbb{G}$-invariant set $A \subseteq K$ such that*

$$\mathcal{N}(c(A), \rho, \epsilon) = \mathcal{N}(A, \rho, \epsilon) = |\mathbb{G}| \cdot \mathcal{N}(A/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon).$$

*Proof.* $L$ is a partial limit of $c$ if there is a sequence $x_n \to x$ such that $c(x_n) \to L$. We claim that $L$ must be in the orbit of $x$. Indeed, $c(x_n) = g_n x_n$ for appropriate $g_n \in \mathbb{G}$. Since $\mathbb{G}$ is finite, by passing to a subsequence, we can assume that $g_n = g \in \mathbb{G}$ does not depend on $n$. Since $x \mapsto gx$ is an isometry, we know that $gx_n \to gx = L$. Since by assumption $c$ has $|\mathbb{G}|$ distinct partial limits at $x$, these are exactly the points in the orbit of $x$. In particular, $gx \neq x$ for every $g \neq e$.

Choose some

$$\epsilon < \frac{\min_{g \neq e} \rho(x, g \cdot x)}{2},$$

and define the $\mathbb{G}$-invariant set $A = A(\epsilon)$ by

$$A := \bigcup_{g \in \mathbb{G}} B_\epsilon(g \cdot x)$$

Note that all partial limits of $c$ at $x$ belong to $\overline{c(A)}$, and therefore $gx \in \overline{c(A)}$ for all $g \in \mathbb{G}$. Any finite $\epsilon$-cover of $c(A)$ also covers $\overline{c(A)}$. Since the distance between any two distinct elements in the orbit of $x$ is larger than $2\epsilon$, an $\epsilon$-cover of $c(A)$ must contain at least $|\mathbb{G}|$ elements. On the other hand, $A$ itself is a union of $|\mathbb{G}|$ balls of radius $\epsilon$, and therefore

$$\mathcal{N}(c(A), \rho, \epsilon) \leq \mathcal{N}(A, \rho, \epsilon) \leq |\mathbb{G}| \leq \mathcal{N}(c(A), \rho, \epsilon).$$

Therefore, all these inequalities are equalities and

$$\mathcal{N}(c(A), \rho, \epsilon) = \mathcal{N}(A, \rho, \epsilon) = |\mathbb{G}|.$$

Finally, note that

$$\mathcal{N}(A/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon) = 1,$$

as one ball of radius $\epsilon$ is enough to cover the quotient space, since all $|\mathbb{G}|$ balls are the same up to group action.

Hence,

$$\mathcal{N}(c(A), \rho, \epsilon) = \mathcal{N}(A, \rho, \epsilon) = |\mathbb{G}| \cdot \mathcal{N}(A/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon).$$

□

### A.3 Proofs for Section 4

For proving Lemma 4.2, we need the following lemma:

**Lemma A.4.** *Let $S = [m]^n$ be the set of all vectors of length $n$ with entries in $[m]$. Define an equivalence relation on $S$ by*

$$c_1 \sim c_2 \iff \exists \pi \in S_n : \ c_2 = \pi \cdot c_1.$$

*Then the number of equivalence classes is exactly*

$$\binom{n+m-1}{n}.$$

*Proof.* For $u \in S$ and $i \in [m]$, let $c_i^u$ denote the number of times $i$ appears in $u$.

Then

$$u \sim v \iff c_i^u = c_i^v \quad \text{for all } i \in [m].$$

Thus, the equivalence classes are in one-to-one correspondence with the nonnegative integer solutions of

$$\sum_{i=1}^{m} c_i = n, \qquad c_i \geq 0 \text{ for all } i \in [m].$$

By the standard stars-and-bars formula [25], the number of such solutions is

$$\binom{n+m-1}{n}.$$

□

We now restate and prove 4.2.

**Lemma 4.2.** *[Proof in A.3] Let $n, d$ be natural numbers and $\epsilon \in (0,1)$. Set $K = [0,1]^{d\times n}$ and $\mathbb{G} = S_n$. Then the covering number of the quotient space satisfies*

$$\mathcal{N}(K/\mathbb{G}, \rho_{\mathbb{G}}, \epsilon) \leq \binom{n + k^d - 1}{n}, \text{ where } k = \left\lceil \frac{1}{2\epsilon} \right\rceil. \tag{4}$$

*Proof.* Set $k = \lceil 1/(2\epsilon) \rceil$. The set $K$ can be covered by $k^{dn}$ hypercubes of side length $1/k$ and radius $1/(2k) \leq \epsilon$. Let $C$ denote the collection of centers of these hypercubes, namely

$$C = \left\{ \left[ \frac{2\mathbf{i}_1 + \mathbf{1}_d}{2k}, \frac{2\mathbf{i}_2 + \mathbf{1}_d}{2k}, \ldots, \frac{2\mathbf{i}_n + \mathbf{1}_d}{2k} \right] \;\middle|\; \mathbf{i}_j \in \{0, 1, \ldots, k-1\}^d,\ 1 \leq j \leq n \right\},$$

Applying the quotient map $\pi(c) = [c]$ we obtain points $\pi(C)$ in the quotient space which are an $\epsilon$ cover of $K/\mathbb{G}$. Note that points in $C$ which are permutations of each other are mapped by $\pi$ to the same orbit. Therefore, the cardinality of the cover $\pi(C)$ is the number of equivalence classes with respect to the relation on $C$ defined by the group: $c_1 \sim c_2 \iff \exists \tau \in S_n : c_1 = \tau(c_2)$. Using Lemma A.4, we know that the number of equivalence classes is

$$|\pi(C)| = \binom{n + k^d - 1}{n}$$

□

We now restate and prove 4.1.

**Lemma 4.1.** *[Proof in A.3] Let $k, d, n \geq 2$ be natural numbers, and set $\epsilon = 1/(2k)$, $K = [0,1]^{d\times n}$ and $\mathbb{G} = S_n$. Then the covering number of the lexicographical sorting canonization satisfies*

$$\left(\frac{1}{2\epsilon}\right)^{(d-1)\cdot n+1} \leq \mathcal{N}(c_{\text{lex}}(K), \rho_\infty, \epsilon). \tag{3}$$

*Proof.* We consider the set $U \subset [0,1]^{d\times n}$ defined by

$$U = \{X \in [0,1]^{d\times n} | \quad X_{11} = X_{12} = \ldots = X_{1n}\}.$$

The set $U$ is not a subset of $c_{\text{lex}}(K)$, but it is a subset of its closure $\overline{c_{\text{lex}}(K)}$. This is because for every $X \in U$ we can apply an arbitrarily small perturbation of the first row of $X$ so that it is ordered from small to large, and $X$ will be unaffected by the canonization. We then obtain

$$\mathcal{N}(c_{\text{lex}}(K), \rho_\infty, \epsilon) = \mathcal{N}(\overline{c_{\text{lex}}(K)}, \rho_\infty, \epsilon) \geq \mathcal{N}(U, \rho_\infty, \epsilon) \overset{(*)}{=} \left(\frac{1}{2\epsilon}\right)^{(d-1)\cdot n+1},$$

where $(*)$ follows as follows. The map

$$f : [0,1]^{(d-1)n+1} \to U$$

obtained by repeating the first coordinate $n$ times in the first row and using the remaining coordinates for the other entries is an isometry with respect to the infinity norm. Moreover, the map $r : K \to U$ that replaces all entries in the first row of $X \in K$ by $X_{11}$ and leaves all other entries unchanged is a 1-Lipschitz retraction onto $U$. Hence the isometry $f$ preserves the covering number, and by Lemma A.6,

$$\mathcal{N}(U, \rho_\infty, \epsilon) = \mathcal{N}\left([0,1]^{(d-1)n+1}, \rho_\infty, \epsilon\right) = \left(\frac{1}{2\epsilon}\right)^{(d-1)n+1}.$$

□

### A.4 Hilbert Canonization

In this appendix, we give a formal definition of Hilbert curves and a full proof of Theorem 4.3.

The Hilbert curve $H(x)$ is obtained as a limit of a sequence of increasingly fine Hilbert constructions, and their finite-resolution versions $H_m$ are our main focus. We will now define these constructions, loosely following the terminology from [11]. Firstly, for a given integer $L$, we consider a partition of the unit interval into $2^L$ disjoint intervals: the semi-closed intervals $I_L(\ell) = [\ell 2^{-L}, (\ell+1)2^{-L})$ for $\ell = 0, \ldots, 2^L - 2$, and the closed interval $I_L(2^L - 1) = [(2^L - 1)2^{-L}, 1]$. We denote the collection of all these intervals by $\mathcal{I}_L$. For a natural $d > 1$, we also consider the partition of the unit hypercube $[0,1]^d$ into $2^{L \cdot d}$ disjoint hypercubes of the form $E_L^d(\vec{\ell}) = I_L(\ell_1) \times \ldots \times I_L(\ell_d)$. We denote the collection of these hypercubes by $\mathcal{E}_L^d$.

Next, for any given $m$, we define $H_m$ to be a bijection from $\mathcal{I}_{d \cdot m}$ to $\mathcal{E}_m^d$, such that (i) adjacent intervals are mapped to adjacent hypercubes and (ii) if $I_{d\cdot(m+1)}(k') \subseteq I_{d\cdot m}(k)$ then $H_{m+1}\left(I_{d\cdot(m+1)}(k')\right) \subseteq H_m\left(I_{d\cdot m}(k)\right)$. An illustration of $H_m, m = 1, 2, 3$ for $d = 2$ is given in Figure 2. Each $H_m$ maps a finite collection of intervals bijectively to a finite collection of hypercubes. This naturally induces a bijection from the grid defined by the interval centroids, to the grid defined by the hypercube centroids. We denote these grids by $\mathbb{A}[d \cdot m] \subseteq [0,1]$ and $\mathbb{B}[d, m] \subseteq \mathbb{R}^d$, respectively.

The Hilbert curve $H$ is obtained from $H_m$ by a limiting procedure and shown to be Hölder continuous [11]. Following similar ideas, we show that each Hilbert function $H_m$ is Hölder continuous with a constant that does not depend on $m$. This property will be crucial for the analysis of the covering number of the Hilbert canonization.

**Lemma A.5.** *For every natural $m, d$, we have that*

$$\forall x, y \in \mathbb{A}[d \cdot m], \quad \|H_m(x) - H_m(y)\|_\infty \leq 4|x - y|^{1/d}$$

*Proof.* Let $x \neq y$ be points in the one dimensional grid $\mathbb{A}[d \cdot m]$, and assume without loss of generality that $x < y$. Note that $y - x \geq \frac{1}{2^{d \cdot m}}$, since this is the minimal distance between points in the grid. On the other hand, $y - x \leq 1$. Therefore, there exists some $k$ with $0 \leq k \leq m - 1$, such that

$$2^{-d\cdot(k+1)} \leq |x - y| \leq 2^{-d\cdot k}. \tag{6}$$

It follows that $x$ and $y$ are in two adjacent intervals in the $k$-th level (or in the same interval), and by the nesting property of Hilbert's curve, we know that $H_m(x)$ and $H_m(y)$ are in two adjacent hypercubes (or in the same hypercube). Therefore, their distance is at most the sum of the diameters of those two cubes, so

$$\|H_m(x) - H_m(y)\|_\infty \leq 2 \cdot 2^{-k} \overset{(6)}{\leq} 4|x - y|^{1/d}$$

□

**Hilbert Canonization** As a first step, we define a canonization not on all of $[0,1]^{d \times n}$, but rather on the finite set of $n$ tuples of grid elements $\mathbb{B}^n[d, m]$. This is done by mapping each tuple-element to the unit interval using $H_m^{-1}$, then sorting the resulting $n$-dimensional array, and mapping back to the $d$-dimensional grid using $H_m$. We denote this canonization by $c_m$. We can write $c_m = H_m \circ \mathrm{sort} \circ H_m^{-1}$, with the understanding that $H_m$ and $H_m^{-1}$ are applied elementwise to $n$-tuples of grid elements.

To extend $c_m$ to all of $K = [0,1]^{d \times n}$, we employ the following procedure: For a given $X = (x_1, \ldots, x_n) \in K$, (i) **Rounding:** Each $x_j$ resides in a unique hypercube in $\mathcal{E}_m^d$. Replace $x_j$ with the centroid of this cube $y_j \in \mathbb{B}[d, m]$. (ii) **Sorting:** find the permutation $\tau \in S_n$ which sorts $\left(H_m^{-1}(y_1), \ldots, H_m^{-1}(y_n)\right)$ from small to large. (iii) **Canonizing** Apply the permutation $\tau$ to $X$. We note that if $X$ contains several points in the same hypercube, the permutation $\tau$ is not uniquely defined. In this case, the order of points within the hypercube will be determined by lexicographical sorting.

We now restate and prove our theorem regarding the covering number of the Hilbert canonization:

**Theorem 4.3.** *[Proof in A.4] Let $d, n, m \geq 2$ be natural numbers, let $\epsilon \in (0,1)$, and assume that $\epsilon > 2^{-m-1}$. Set $K = [0,1]^{d \times n}$ and $\mathbb{G} = S_n$. Then the covering number of the Hilbert canonization $c_m$ satisfies*

$$\mathcal{N}(c_m(K), \rho_\infty, \epsilon) \leq \binom{n + \lceil 1/(2\delta) \rceil - 1}{n}, \text{ where } \delta = \left(\frac{\epsilon - 2^{-m-1}}{4}\right)^d$$

*Proof.* We first reduce the problem from covering $K$ to covering the finite grid $\mathbb{B}^n[d,m]$. For convenience, write

$$\mathbb{B}^n := \mathbb{B}^n[d,m], \qquad r := \epsilon - 2^{-m-1} > 0.$$

For any

$$X = (x_1, \dots, x_n) \in K,$$

let

$$Y = (y_1, \dots, y_n) \in \mathbb{B}^n$$

be obtained by the rounding procedure, i.e., each $x_j$ is replaced by the centroid $y_j$ of the unique level-$m$ hypercube containing it. Since every such hypercube has side length $2^{-m}$,

$$\|X - Y\|_\infty \le 2^{-m-1}.$$

Moreover,

$$\|c_m(X) - c_m(Y)\|_\infty \le 2^{-m-1}.$$

Indeed, the Hilbert indices of the rounded points determine the same ordering of the level-$m$ hypercubes for $X$ and $Y$. Within a hypercube, $c_m(X)$ uses lexicographical sorting, whereas all corresponding rounded points in $Y$ are identical. Hence the permutation used to canonize $X$ may also be applied to $Y$ without changing $c_m(Y)$, and the pointwise rounding error remains at most $2^{-m-1}$.

Consequently, any $r$-cover of $c_m(\mathbb{B}^n)$ is an $\epsilon$-cover of $c_m(K)$, and therefore

$$N\big(c_m(K), \rho_\infty, \epsilon\big) \le N\big(c_m(\mathbb{B}^n), \rho_\infty, r\big).$$

It remains to bound the right-hand side. On $\mathbb{B}^n$, the Hilbert canonization can be written as

$$c_m = H_m \circ \mathrm{sort} \circ H_m^{-1},$$

where $H_m$ and $H_m^{-1}$ are applied elementwise. Define

$$S := \mathrm{sort} \circ H_m^{-1}(\mathbb{B}^n) \subseteq [0,1]^n.$$

Thus,

$$c_m(\mathbb{B}^n) = H_m(S).$$

We claim that

$$N\big(H_m(S), \rho_\infty, r\big) \le N\big(S, \rho_\infty, \delta\big), \tag{*}$$

where

$$\delta = \left(\frac{r}{4}\right)^d = \left(\frac{\epsilon - 2^{-m-1}}{4}\right)^d.$$

To prove $(*)$, let $C$ be any external $\delta$-cover of $S$. For each $z \in C$, define

$$S_z := S \cap B_{\rho_\infty}(z, \delta).$$

If $u, v \in S_z$, then

$$\|u - v\|_\infty \le 2\delta.$$

Applying Lemma A.5 elementwise gives

$$\|H_m(u) - H_m(v)\|_\infty \le 4\|u - v\|_\infty^{1/d} \le 4(2\delta)^{1/d}.$$

Therefore,

$$\mathrm{diam}_{\rho_\infty}\big(H_m(S_z)\big) \le 4(2\delta)^{1/d}.$$

Since our covering numbers allow centers in the ambient space, every subset of an $\ell_\infty$ space with diameter $D$ is contained in an $\ell_\infty$ ball of radius $D/2$. Hence $H_m(S_z)$ is contained in a ball of radius

$$2(2\delta)^{1/d} = 2^{1+1/d}\delta^{1/d} \le 4\delta^{1/d} = r.$$

Thus, every ball in the $\delta$-cover $C$ gives rise to one $r$-ball covering the corresponding part of $H_m(S)$. It follows that

$$N\big(H_m(S), \rho_\infty, r\big) \le |C|.$$

Taking $C$ to be a minimum-cardinality $\delta$-cover proves $(*)$.

We now obtain

$$\begin{aligned} N\big(c_m(K), \rho_\infty, \epsilon\big) &\le N\big(c_m(\mathbb{B}^n), \rho_\infty, r\big) \\ &= N\big(H_m(S), \rho_\infty, r\big) \\ &\le N\big(S, \rho_\infty, \delta\big) \\ &\le N\big(\mathrm{sort}([0,1]^n), \rho_\infty, \delta\big). \end{aligned}$$

Finally, one-dimensional sorting is an isometry between the quotient $[0,1]^n/S_n$ and $\mathrm{sort}([0,1]^n)$. Hence, by the quotient-space covering bound in Lemma 4.2, applied with dimension 1,

$$N\big(\mathrm{sort}([0,1]^n), \rho_\infty, \delta\big) \le \binom{n + \lceil \frac{1}{2\delta} \rceil - 1}{n}.$$

Combining the above inequalities yields

$$N\big(c_m(K), \rho_\infty, \epsilon\big) \le \binom{n + \lceil \frac{1}{2\delta} \rceil - 1}{n},$$

where

$$\delta = \left( \frac{\epsilon - 2^{-m-1}}{4} \right)^d.$$

□

**Lemma A.6.** *Let $K = [0,1]^{d \times n}$ and let $\epsilon = \frac{1}{2k}$ for some $k \in \mathbb{N}$. Then*

$$\mathcal{N}(K, \rho_\infty, \epsilon) = \frac{1}{(2\epsilon)^{nd}}$$

*Proof.* For the upper bound, partition each copy of $[0,1]$ into $k$ intervals of length $1/k = 2\epsilon$. Their Cartesian products partition $K$ into $k^{nd}$ hypercubes of side length $2\epsilon$, each contained in an $\epsilon$-ball with respect to $\rho_\infty$. Hence

$$\mathcal{N}(K, \rho_\infty, \epsilon) \le k^{nd} = \frac{1}{(2\epsilon)^{nd}}.$$

For the lower bound, let $F$ be any $\epsilon$-cover of $K$, and let $\mu$ denote Lebesgue measure on $K$. Since

$$K \subseteq \bigcup_{c \in F} B_{\rho_\infty}(c, \epsilon),$$

and every such ball intersects $K$ in a set of measure at most $(2\epsilon)^{nd}$, we obtain

$$1 = \mu(K) \le \sum_{c \in F} \mu(B_{\rho_\infty}(c, \epsilon) \cap K) \le |F| (2\epsilon)^{nd}.$$

Thus $|F| \ge (2\epsilon)^{-nd}$, proving the claim. □

# B Additional Experiments

## B.1 Laplacian Eigenvector Sign Ambiguity: Augmentation vs. Canonization on `ogbg-molpcba`

This appendix presents an additional empirical comparison between deterministic canonization and training-time random group augmentation, on a setting distinct from the main-text experiments and complementary to the PCA experiment of Section 4. The goal is to ask whether the empirical behavior of these two strategies on a standard graph property prediction benchmark is qualitatively consistent

with our Proposition 2.1 and the covering-number inequality (2); we do not claim a quantitative test of the inequality, since the test AP on a finite-sample-trained GIN mixes optimization, capacity, and generalization effects that the inequality does not isolate. The experiment is also consistent with the empirical observation cited in the Related Work that randomized SMILES can outperform canonized SMILES [1, 3].

**Module.** We study graph property prediction on `ogbg-molpcba` using a 5-layer Graph Isomorphism Network (GIN) with Laplacian positional encoding (LapPE). For each graph $G$ with normalized Laplacian $L_G$, we compute the $k$ smallest non-trivial eigenpairs $\{(\lambda_i, v_i)\}_{i=1}^{k}$ and concatenate the eigenvectors as a node-level positional encoding. The resulting LapPE is invariant to graph isomorphism only up to the eigenvector *ambiguity group*

$$\mathbb{G}_{\mathrm{LapPE}} \;=\; \prod_{i\,:\,m_i=1} \{-1,+1\} \;\times\; \prod_{i\,:\,m_i>1} O(m_i),$$

where $m_i$ denotes the multiplicity of eigenvalue $\lambda_i$: a sign flip is admissible for each simple eigenvalue and an arbitrary orthogonal change of basis is admissible within each multiplicity-$m_i$ block. On molecular graphs, the spectrum is almost everywhere simple, so in practice $\mathbb{G}_{\mathrm{LapPE}}$ reduces to $\{-1,+1\}^k$ for almost all graphs. The natural questions are then (i) whether to canonize this ambiguity, and (ii) whether augmentation-based approximations of group averaging behave consistently with our framework.

**Three methods compared.** We compare three strategies for resolving $\mathbb{G}_{\mathrm{LapPE}}$, all implemented as faithful numpy ports of the reference torch code from the Laplacian canonization release of [19]. The first, `map` [19] (Maximal Axis Projection), resolves a simple-eigenvalue sign by forming the rank-one projector $P_i = u_i u_i^\top$, grouping the coordinate indices $j$ by the rounded column norms $\|P_i[:,j]\|$ (14-decimal rounding), and walking these groups from smallest to largest norm to find the first group whose indicator vector $X$ (the sum of standard basis vectors over the group, plus a uniform shift) gives $\|P_i X\| \neq 0$; the canonical sign is then $\mathrm{sign}(u_i^\top X)$. For a multiplicity-$m_i$ block, `map` forms $P = U_i U_i^\top$, selects the $m_i$ *largest*-norm coordinate groups, and builds the canonical orthonormal basis incrementally by projecting each group-indicator vector onto the residual orthogonal complement (with plain QR inside the complementary-space step). The second, `oap` [18] (Orthogonalized Axis Projection), differs from `map` in two places: the sign step rounds the column norms to 6 decimals (coarser tie-grouping under near-degenerate norms), and the basis step replaces the largest-norm groupings of `map` by a hash-based grouping of columns and uses signed-QR in the complementary-space iteration. The third, `random_augmented`, flips the sign of each eigenvector with probability $1/2$ at every training-time forward pass, so the network sees a uniformly random element of $\{-1,+1\}^k$ at each step. This is a finite-group augmentation that approximates group averaging over $\mathbb{G}_{\mathrm{LapPE}}$ without summing over its $2^k$ elements explicitly. At evaluation time the `random_augmented` model is fed the deterministic eigsh-sign convention from the cache (no fresh random draws), yielding a single deterministic prediction per graph; this is the number reported in the `random_augmented` column of Table 6. Note that this is an *adversarial* test-time deployment of an averaging-trained model (it scores a single point of the orbit rather than the orbit mean); the AUG-$K$ column reports the alternative Reynolds-style sample of $K$ orbit elements, and we discuss the gap between the two below.

**Test-time averaging AUG-$K$.** For `random_augmented` models we additionally evaluate a test-time *Reynolds-style* averaged metric. At evaluation, for each test graph we draw $K$ independent group elements $g_1, \ldots, g_K \in \mathbb{G}_{\mathrm{LapPE}}$ (each $g_j$ a sign-flip pattern on the simple-eigenvalue subspace and a Haar-random orthogonal matrix on each multiplicity-$m$ block); we forward-pass the model on each sign-flipped LapPE; we average the predicted per-task probabilities; and we score the averaged predictions with the official OGB AP evaluator. We use $K=8$ at $k=3$ (which approximately covers $\{-1,+1\}^3$ under uniform i.i.d. sampling: 8 draws cover $\sim 65\%$ of the 8-element group on average) and $K=16$ at $k \in \{8, 16\}$ (sparse Monte-Carlo over $\{-1,+1\}^8$ and $\{-1,+1\}^{16}$ respectively). On `ogbg-molpcba` the spectrum is almost everywhere simple, so the multiplicity-block factor $\prod_i O(m_i)$ is degenerate in practice and the $g_j$ samples reduce to sign-flip patterns. This is the empirical group-averaging operator of Section 2, restricted to a finite sample.

**Hyperparameters and protocol.** GIN, 5 layers, hidden dim $h \in \{16, 128, 512\}$, dropout 0.5, batch size 32, optimizer Adam with cosine learning-rate schedule and initial rate $10^{-3}$, 3 seeds per cell

($\{0, 1, 2\}$), eigval-scaling off, eigenvector cache size 15. Checkpoint selection: the model state at the epoch with the highest validation AP across all training epochs is saved (ties broken by the earliest epoch); the final test AP reported in Tables 6 and 7 is this checkpoint's score on the OGB test split. Software stack: PyTorch 2.10.0, CUDA 12.8, `torch_geometric` 2.7.0, `numpy` 2.2.6, `scipy` 1.15.3, `ogb` 1.3.6 (OGB evaluator: `Evaluator(name='ogbg-molpcba')`). Per-arm seeding via `torch.manual_seed`, `numpy.random.seed`, and `random.seed`; `cudnn.deterministic` is not enforced, so two runs at the same seed on different hardware are not guaranteed to be bit-identical (we observe the resulting variability absorbed into the seed std at the magnitudes shown in Table 7). Standard deviations reported below are sample standard deviations ($\mathrm{ddof} = 1$). Code, raw `results.json` files, the `orbit_stability.json` files behind Table 8, and reproduction commands are released at the project repository.

**Compute budget.** At $k \in \{3, 8\}$ all three arms are trained for 200 epochs with early-stop patience 15. At $k = 16$, where validation AP for `random_augmented` continues to improve well past 200 epochs, all three arms are trained for 500 epochs with early stopping disabled (patience 999). Tables 6 and 7 therefore report a matched-compute comparison at every $(k, h)$ cell. Both schedules are released in the project repository.

**Results.** Table 6 reports mean test AP at all nine $(k, h)$ cells for the three methods, alongside the AUG-$K$ test-time-averaged metric for `random_augmented`. Table 7 zooms in on the $k = 16$ slice with sample standard deviations.

Table 6: Mean test AP on `ogbg-molpcba` at matched compute, 3 seeds per cell ($k \in \{3, 8\}$ at 200 epochs / patience 15, $k = 16$ at 500 epochs / patience 999 for all three arms). Per-cell row-best margins are mostly within seed std and none survives a Bonferroni correction across the nine cells at $\alpha = 0.05$ (see Table 7 for the $k = 16$ Welch $p$-values); the bolding, therefore, marks the nominal row-best mean, not a significance claim. The AUG-$K$ column reports test-time averaging over $K$ ambiguity-group draws applied to the `random_augmented`-trained model (see text).

| $k$ | $h$ | `map` ↑ | `oap` ↑ | `random_augmented` ↑ | AUG-$K$ |
|---|---|---|---|---|---|
| 3 | 16 | 0.0626 | 0.0641 | **0.0662** | 0.0663 ($K = 8$) |
| 3 | 128 | 0.1704 | 0.1682 | **0.1762** | 0.1765 ($K = 8$) |
| 3 | 512 | 0.2397 | **0.2431** | 0.2395 | 0.2398 ($K = 8$) |
| 8 | 16 | 0.0622 | 0.0605 | **0.0636** | 0.0637 ($K = 16$) |
| 8 | 128 | 0.1705 | 0.1754 | **0.1766** | 0.1766 ($K = 16$) |
| 8 | 512 | 0.2344 | 0.2371 | **0.2395** | 0.2400 ($K = 16$) |
| 16 | 16 | **0.0716** | 0.0703 | 0.0691 | 0.0693 ($K = 16$) |
| 16 | 128 | **0.1923** | 0.1836 | 0.1916 | 0.1921 ($K = 16$) |
| 16 | 512 | 0.2468 | 0.2419 | **0.2483** | 0.2481 ($K = 16$) |

Table 7: Matched-compute $k = 16$ slice of Table 6: all three arms trained for 500 epochs with patience 999. Reported are mean $\pm$ sample standard deviation over 3 seeds, plus the two-sided Welch $t$-test $p$-value for the row-best mean against the row-second-best mean. None of the three row-best margins reaches $p < 0.05$, and after a Bonferroni correction across the nine cells in Table 6 ($\alpha = 0.05/9 \approx 0.0056$) none of the row-best margins anywhere in the appendix is significant.

| $h$ | `map` | `oap` | `random_augmented` | Welch $p$ (best vs. 2nd) |
|---|---|---|---|---|
| 16 | $\mathbf{0.0716 \pm 0.0067}$ | $0.0703 \pm 0.0070$ | $0.0691 \pm 0.0008$ | $\approx 0.83$ |
| 128 | $\mathbf{0.1923 \pm 0.0005}$ | $0.1836 \pm 0.0025$ | $0.1916 \pm 0.0009$ | $\approx 0.32$ |
| 512 | $0.2468 \pm 0.0021$ | $0.2419 \pm 0.0028$ | $\mathbf{0.2483 \pm 0.0026}$ | $\approx 0.48$ |

**Test-time averaging $\Delta$ is small.** Across all nine $(k, h)$ cells the difference between the single-pass deterministic-eigsh evaluation of the `random_augmented` model and its AUG-$K$ averaged evaluation (Table 6, right-most two columns) lies in $[-0.0002, +0.0005]$ AP, within the seed noise at every cell. The AUG-$K$ $\Delta$ is an aggregate metric, not a direct measurement of model invariance; we test invariance directly in the next paragraph.

**Per-graph orbit stability.** For each test graph we draw $K=64$ i.i.d. elements $g_j$ of the sign-flip subgroup of $\mathbb{G}_{\mathrm{LapPE}}$, forward-pass the trained `random_augmented` model on each $g_j(\mathrm{LapPE})$, and record the per-task predicted probability. The resulting per-graph spread (max $-$ min predicted probability across the $K=64$ draws) is a direct measure of how much the model's output varies across orbit elements. Aggregating across all 43,793 `ogbg-molpcba` test graphs and 128 tasks, at hidden dim $h=128$ and over all 3 training seeds we observe a clean monotone scaling of per-graph spread with $k$ (Table 8): mean per-graph spread is $0.0013\pm0.0005$ at $k=3$, $0.0027\pm0.0012$ at $k=8$, and $0.0048\pm0.0008$ at $k=16$. This is qualitatively consistent with $|\mathbb{G}_{\mathrm{LapPE}}|=2^k$: more orbit elements give more room to vary, and the model is closer to be invariant for smaller $k$. The within-seed standard deviation of the OGB AP across the $K=64$ orbit elements is at most $0.0006$ at $k=16$ and below $0.0004$ at $k\in\{3,8\}$, in every case much smaller than the between-seed standard deviation of the same quantity ($0.0062$ at $k\in\{3,8\}$, $0.0015$ at $k=16$): the model's output varies less when we resample the ambiguity group at fixed seed than when we change the training seed at fixed ambiguity element. The Reynolds-style (mean-probability) AP across the $K=64$ draws is $0.1923\pm0.0017$ at $k=16/h=128$, within seed noise of the matched-compute single-pass number ($0.1916\pm0.0009$ for `random_augmented` and $0.1923\pm0.0005$ for matched-compute `map` at the same cell). We read this as direct evidence that the `random_augmented` model is approximately $\mathbb{G}_{\mathrm{LapPE}}$-invariant on the test distribution at the configurations tested, with the residual non-invariance growing predictably with the size of the ambiguity group. Along the hidden-dim axis at $k=16$ (Table 8, lower block, 3 seeds per cell), the per-graph spread is non-monotone in $h$: $0.0032\pm0.0015$ at $h=16$, $0.0048\pm0.0008$ at $h=128$, $0.0033\pm0.0004$ at $h=512$. We attribute this to a capacity floor and a capacity ceiling: at $h=16$ the network underfits (test AP $\approx0.07$, near a constant-prediction baseline) so per-graph predictions are bunched near a few values and have little room to spread; at $h=512$ the network has enough capacity to learn invariance well; the middle case $h=128$ shows the largest residual non-invariance, although the $h=16$ row's larger seed std ($0.0015$) suggests this conclusion is sensitive to which random initialization one looks at.

Table 8: Per-graph orbit stability of the trained `random_augmented` model under $K=64$ uniform i.i.d. draws from the sign-flip subgroup of $\mathbb{G}_{\mathrm{LapPE}}$, evaluated on the `ogbg-molpcba` test set (43,793 graphs $\times$ 128 tasks). All entries are the mean $\pm$ sample standard deviation over the 3 training seeds. "Per-graph spread" is the mean over graphs and tasks of $(\max-\min)$ predicted probability across the $K=64$ orbit elements. "Within-seed AP std" is the standard deviation of the OGB AP scored on each single orbit draw, averaged across seeds. "Reynolds AP" scores the mean probability over the $K=64$ draws. Per-graph spread scales monotonically with $|\mathbb{G}_{\mathrm{LapPE}}|=2^k$.

| $k$ | $h$ | per-graph spread | within-seed AP std | Reynolds AP |
|---|---|---|---|---|
| *$k$ axis at $h=128$ (3 seeds, $\lvert\mathbb{G}_{\mathrm{LapPE}}\rvert=2^k$):* | | | | |
| 3 | 128 | $0.0013\pm0.0005$ | $\leq 0.0004$ | $0.1764\pm0.0062$ |
| 8 | 128 | $0.0027\pm0.0012$ | $\leq 0.0005$ | $0.1766\pm0.0062$ |
| 16 | 128 | $0.0048\pm0.0008$ | $\leq 0.0006$ | $0.1923\pm0.0017$ |
| *$h$ axis at $k=16$ (3 seeds):* | | | | |
| 16 | 16 | $0.0032\pm0.0015$ | $\leq 0.0003$ | $0.0694\pm0.0010$ |
| 16 | 128 | $0.0048\pm0.0008$ | $\leq 0.0006$ | $0.1923\pm0.0017$ |
| 16 | 512 | $0.0033\pm0.0004$ | $\leq 0.0004$ | $0.2481\pm0.0023$ |

**Reading the tables.** Reading along the hidden-dim axis at fixed $k$ in Table 6, the row-best AP grows monotonically with $h$ at every $k$, and the row-best arm is most often `random_augmented` (six of nine cells), with `map` nominally row-best at $k=16/h\in\{16,128\}$ and `oap` nominally row-best at $k=3/h=512$. The matched-compute $k=16$ slice in Table 7 resolves these margins against seed variance: the smallest sample std on each row is consistently the `random_augmented` std ($0.0008$, $0.0009$, $0.0026$), while `map` carries the largest std at small $h$ ($0.0067$ at $h=16$). Under a two-sided Welch $t$-test for unequal variances, the row-best versus row-second-best margins at $k=16$ have $p$-values of $0.83$, $0.32$, and $0.48$ at $h=16,128,512$ respectively, none significant. At $k\in\{3,8\}$ the row-best margins are similar in magnitude or smaller, with `random_augmented` winning five of six rows there but at margins of the same order as the seed std. We therefore present Table 6 as a nominal-rank summary, not as a benchmark claim.

**Tie to the theory.** On `ogbg-molpcba` at matched compute, with 3 seeds per cell, we did not detect a statistically significant difference between training-time random sign-flip augmentation and projection-based deterministic canonization at any of the nine $(k, h)$ cells; this is qualitatively consistent with (2) of Proposition 2.2, which suggests that augmentation can match canonization on tasks where the covering number of $\mathbb{G}$ is small, but it is not a positive test of the inequality, since AP gaps of order $10^{-3}$ are dominated by optimization and seed noise rather than by the input-space covering complexity, the inequality bounds. The AUG-$K$ column moves AP by at most $+0.0005$ across the nine cells; this is consistent with the cached deterministic eigsh-sign PE already lying close to the orbit-mean the network was trained on, but is not itself a direct test of model invariance.

# C Experimental Setup and Hyperparameter Details

## C.1 Experimental Setup for the covering number experiment

We applied the following preprocessing for *each* sample: we first sample $256$ points, then we shift them to the positive axis, and then divide by the maximum axis along $x, y, z$, obtaining all entries to be positive and between $0$ and $1$. The metric we considered on the point cloud space is

$$\rho(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - Y_i\|_2.$$

The same metric is used for the two canonizations, and its quotient $\rho_{\mathbb{G}}$ is used for the quotient space.

## C.2 Experiment setup: ModelNet40 classification

Table 9: Test accuracy of the Global MLP under varying training set sizes. The performance gap between Hilbert and lexicographical sorting widens as data becomes scarcer, highlighting the regularizing effect of canonization.

| Training Samples | None ↑ | Lexicographical ↑ | Hilbert ↑ | DeepSets ↑ |
|---|---|---|---|---|
| 9840 (full) | 0.433 | 0.700 | 0.748 | **0.759** |
| 4920 | 0.408 | 0.641 | 0.714 | **0.729** |
| 2460 | 0.343 | 0.573 | 0.673 | **0.683** |
| 1230 | 0.307 | 0.504 | 0.601 | **0.644** |

In the ModelNet40 classification tasks, we sampled from each input CAD model point cloud, which consists of $1024$ points, and applied an initial batched affine normalization to $[0, 1]^3$. For the Global MLP models, the input points are processed using one of three strategies: no canonization (baseline), lexicographical sorting (i.e., sorting points sequentially by their $x$, then $y$, then $z$ coordinates), or Hilbert curve sorting (using a space size of $2^{12}$). We additionally evaluate DeepSets as a permutation-invariant model.

For the Global MLP variants, following the ordering step, the coordinates are projected into a higher-dimensional space using Random Fourier Features to capture high-frequency geometric details. The resulting feature vector is then passed through a bottlenecked MLP consisting of residual blocks with dimensions $[256, 128, 64]$, yielding the final classification logits. DeepSets applies the corresponding pointwise MLP independently to each point and aggregates the resulting logits by summation.

To ensure a rigorous and fair comparison, we selected model-specific hyperparameters for the uncanonized baseline (`ply`), lexicographical sorting (`lex`), Hilbert curve sorting (`hilbert`), and DeepSets. The specific hyperparameters used for the final evaluation are detailed in Table 10. The selected hyperparameters for each model on ModelNet40 are also used for the corresponding experiments on ModelNet10.

**Robustness Across Seeds** To account for variance in weight initialization and optimization dynamics, we evaluated the final optimal configurations across 5 independent random seeds. The averaged test accuracy and standard deviations are reported in Table 11. The results demonstrate that the Hilbert curve canonization not only achieves the highest average accuracy but also exhibits the lowest variance across different initializations.

Table 10: Optimal hyperparameters identified via grid search for each ordering strategy.

| Method | Learning Rate | Dropout | Weight Decay | Fourier Scale | Label Smoothing | Bands |
|---|---|---|---|---|---|---|
| No canon. | $1.63 \times 10^{-3}$ | 0.3 | 0.05 | 1.0 | 0.1 | 2 |
| Lexsort | $6.41 \times 10^{-4}$ | 0.1 | 0.05 | 0.1 | 0.2 | 3 |
| Hilbert | $9.61 \times 10^{-4}$ | 0.3 | 0.01 | 0.1 | 0.2 | 3 |

Table 11: Test accuracy across 5 independent random seeds using the optimal hyperparameters. Hilbert canonization yields both higher performance and greater stability.

| Method | Average Accuracy ↑ | Standard Deviation |
|---|---|---|
| No canon. | 0.433 | 0.0249 |
| Lexsort | 0.700 | 0.01358 |
| Hilbert | **0.748** | 0.00944 |

### C.3 Experimental Setup: Modelnet40 Data Scarcity Experiment

To simulate data scarcity, we subsample the ModelNet40 training set using strides of 1, 2, 4, and 8, resulting in training subsets of 9840, 4920, 2460, and 1230 samples, respectively. Results are shown in Table 9.

### C.4 Experimental Setup: PCA group averaging

We use a DeepSets architecture that is fully invariant to point order permutations but strictly non-invariant to the $\{-1, 1\}^3$ sign flips. We intentionally employ a lightweight, low-capacity version of this architecture. A highly over-parameterized model might achieve higher absolute accuracies, potentially masking the generalization gaps we aim to observe; by constraining the network capacity, we ensure that any performance differences are strictly attributable to the chosen alignment strategy. Each input point cloud consists of $1024$ points from the ModelNet40 dataset. We compare four approaches: (1) **Pure PCA**, which leaves the sign ambiguity unresolved (acting as a baseline or potentially *poor* canonization due to eigensolver discontinuities); (2) **Skewness**, a deterministic canonization mapping $c : K \to K$ that resolves the ambiguity by orienting axes based on the third moment; (3) **Frame Averaging**, an explicitly $\mathbb{G}$-invariant model constructed by averaging the predicted class probabilities across all $8$ elements of the group, implemented in log-space for numerical stability; and (4) **Random Frame**, which applies elements of $\mathbb{G}$ as random data augmentation during training. All models are optimized using Adam for 1600 epochs and evaluated across 5 different random seeds.

### C.5 Experimental Setup: Rotated MNIST

In this set of experiments, we compare the performance of several image rotation-invariant models. We considered four models: simple CNN, CN(p4 frozen)-CNN, CN(p4)-CNN, and AvgCNN. The backbone for all four models is the same and additional layers are added for CN(p4)-CNN and AvgCNN. In order to have the same setting, including the data split we rerun all models, including CN(p4)-CNN. We used a learning rate of $10^{-3}$, a batch size of $256$, and an AdamW optimizer. Each experiment is run for $5$ different seeds, and the mean and standard deviation are reported.

### C.6 Experimental Setup: Deep Weight Spaces

We consider the INR classification setting on MNIST-INR and Fashion-MNIST-INR [21]. In these datasets, each image is represented by the parameters of an implicit neural representation (INR), and the objective is to predict the image class directly from the INR weights. For both datasets, we use the original train/validation/test splits containing $55{,}000$, $5{,}000$, and $10{,}000$ INRs, respectively.

We compare three approaches: (1) **Simple MLP**, a non-invariant model operating directly on the flattened INR parameters; (2) **Canonized MLP**, where each INR is first mapped to a deterministic representative of its hidden-neuron permutation orbit and then provided to the MLP; and (3) **DWSNet**,

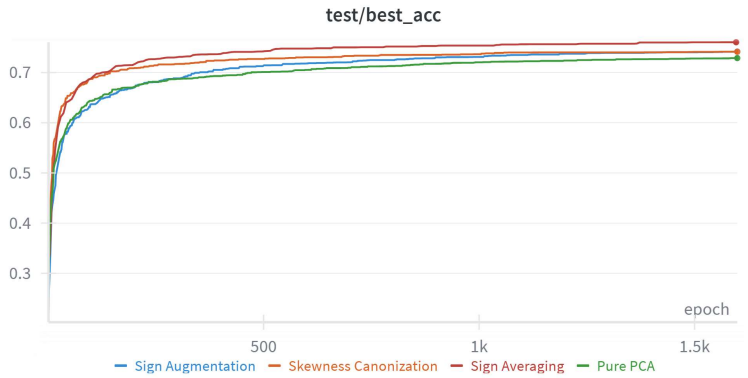



Figure 3: **Rotation Symmetry Accuracy.** The curves illustrate the comparative performance of the four approaches: Random Frame (blue), Frame Averaging (red), Skewness (orange), and the Pure PCA baseline (green). Frame Averaging achieves the strongest overall performance. While Random Frame and Skewness reach comparable final results—both significantly outperforming the Pure PCA baseline—the augmentation-based Random Frame method exhibits a noticeably slower convergence rate than deterministic canonization via Skewness.

a specialized permutation-invariant weight-space model [21]. The Simple MLP and Canonized MLP use the same backbone, so that their comparison isolates the effect of canonization.

For both the Simple MLP and Canonized MLP, generic INR augmentation and hidden-neuron permutation augmentation are disabled. For the Canonized MLP, canonization is performed once as a preprocessing step on all INRs. Since canonization reorders the parameter coordinates, we recompute the normalization statistics on the canonized training set and use these statistics to normalize the canonized training, validation, and test examples. The original dataset statistics are used for the non-canonized models.

Both MLP variants consist of four hidden layers of width $32$ with batch normalization. We train them for $100$ epochs using AdamW with learning rate $5 \times 10^{-3}$, weight decay $5 \times 10^{-4}$, and batch size $512$. Each experiment is repeated for three seeds, $0, 1, 2$. For each seed, we select the checkpoint achieving the highest validation accuracy and report the test accuracy of that checkpoint. Mean accuracy and standard deviation across the three seeds are reported. We now describe the details of the deterministic weight-space canonization procedure.

### C.7 Deep Weight-Space Canonization

We canonize the hidden-neuron permutations sequentially from the input towards the output. After the ordering of layer $\ell - 1$ has been fixed, each neuron in hidden layer $\ell$ is assigned a deterministic lexicographic key constructed from its incoming weights, bias, and an order-invariant summary of its outgoing weights. The hidden neurons are then sorted according to this key, and the induced permutation is propagated to the corresponding dimension of the next weight matrix. Repeating this procedure layer by layer yields a single deterministic representative of the permutation orbit for generic networks. The canonization is performed once as a preprocessing step; the resulting canonized parameters are subsequently provided to exactly the same MLP architecture used for the non-canonized baseline.

### C.8 Dataset provenance and usage terms

The point-cloud classification experiments use ModelNet10 and ModelNet40 [31]; the official ModelNet download page (`https://modelnet.cs.princeton.edu/download.html`) states that the original authors retain copyright to the CAD objects and that the dataset is provided for academic research only. The rotated-MNIST experiments use data based on MNIST [15], whose source images are distributed under the Creative Commons Attribution–ShareAlike 3.0 license according to the Keras dataset documentation (`https://keras.io/api/datasets/mnist/`). The underlying Fashion-MNIST images are made available by Zalando under the MIT license (`https://github.com/zalandoresearch/fashion-mnist`). The MNIST-INR and Fashion-MNIST-INR weight

datasets come from the published deep-weight-space datasets cited in Appendix C.6; their individual redistribution terms should be checked separately from the licenses of the image datasets. The OGB `ogbg-molpcba` experiments use the Open Graph Benchmark data (`https://ogb.stanford.edu/docs/graphprop/`), originally based on the MoleculeNet/PubChem BioAssay collection. The OGB software's license should not be assumed to extend automatically to every underlying dataset. Experiment scripts and setup instructions for the main experiments are publicly available at `https://github.com/yonatansverdlov/Canonization`.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are discussed in the main text and proved in the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discussed the limitation of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All claims contain the assumptions and the consequences, with some of the proofs in the main text and some in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In the appendix, we included all the protocol, and we also added code for reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Experiment code and reproduction commands for the main ModelNet, rotated-MNIST, and INR experiments are available at `https://github.com/yonatansverdlov/Canonization`; the paper describes the experimental protocols in the appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We provide experimental details in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We added error bars for all experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we added the machinery.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conformed with the NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There's no societal impact of the work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper uses ModelNet, rotated MNIST, MNIST-/Fashion-MNIST-INR, and OGB data. Known source and usage terms are documented in the appendix, but not every derived dataset's individual redistribution license could be verified; therefore we do not claim complete license verification.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The authors release experiment code at `https://github.com/yonatansverdlov/Canonization` with installation and reproduction instructions; the repository does not currently include an explicit root-level license file.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.