Karczewski, Rafał; Souza, Amauri H.; Garg, Vikas

On the Generalization of Equivariant Graph Neural Networks

Published in:
Proceedings of Machine Learning Research

Published: 01/01/2024

Document Version
Publisher's PDF, also known as Version of record

On the Generalization of Equivariant Graph Neural Networks

Rafał Karczewski [1] **Amauri H. Souza** [1 2] **Vikas Garg** [1 3]

Abstract

$E(n)$-Equivariant Graph Neural Networks (EGNNs) are among the most widely used and successful models for representation learning on geometric graphs (e.g., 3D molecules). However, while the expressivity of EGNNs has been explored in terms of geometric variants of the Weisfeiler-Leman isomorphism test, characterizing their generalization capability remains open. In this work, we establish the first generalization bound for EGNNs. Our bound depicts a dependence on the weighted sum of logarithms of the spectral norms of the weight matrices (EGNN parameters). In addition, our main result reveals interesting novel insights: i) the spectral norms of the initial layers may impact generalization more than the final ones; ii) ε-normalization is beneficial to generalization — confirming prior empirical evidence. We leverage these insights to introduce a spectral norm regularizer tailored to EGNNs. Experiments on real-world datasets substantiate our analysis, demonstrating a high correlation between theoretical and empirical generalization gaps and the effectiveness of the proposed regularization scheme.

1. Introduction

Leveraging symmetries of the underlying domains/signals is a key design principle underlying successful neural network architectures for structured data (Bronstein et al., 2021; Kipf & Welling, 2017; Cohen et al., 2018; Gilmer et al., 2017). Typically, this boils down to identifying relevant symmetries captured in the form of groups (e.g., groups of translations) and then building predictive models by composing layers of equivariant (or invariant) transformations to the actions of such groups on the inputs. As classic examples, linear layers in convolutional neural networks (CNNs) (Fukushima, 1980; LeCun et al., 1990) implement shift-equivariant functions; graph neural network (GNN) (Gori et al., 2005; Scarselli et al., 2009; Gilmer et al., 2017) layers are equivariant to node permutations. Remarkably, equivariant/invariant architectures have led to breakthroughs in tasks such as drug discovery (Stokes et al., 2020), weather modeling (Verma et al., 2024), simulation of physical systems (Sanchez-Gonzalez et al., 2020), traffic forecasting (Derrow-Pinion et al., 2021), and recommender systems (Ying et al., 2018).

Recently, $E(n)$-equivariant graph neural networks (EGNNs) (Satorras et al., 2021) have emerged as an efficient and powerful approach for learning representations of geometric graphs — i.e., graphs embedded in Euclidean space. EGNNs employ a GNN-like message-passing scheme (Gilmer et al., 2017), where the embeddings of each node are refined using messages from its neighbors in the graph at each layer. Their design ensures that EGNNs inherent the permutation-equivariance property of regular GNNs while also being equivariant to actions of the Euclidean group $E(n)$, which comprise all translations, rotations, and reflections in $\mathbb{R}^n$. EGNNs have been successfully applied to molecular property prediction (Satorras et al., 2021), drug binding structure prediction (Stärk et al., 2022), generative modeling (Garcia Satorras et al., 2021), structure-based drug design (Fu et al., 2022) and molecular dynamics (Arts et al., 2023).

Uncovering the strengths and limits of machine learning models from a theoretical standpoint is imperative to seeding a path to novel principled approaches. For instance, one of the most important aspects of any learning machine is its ability to generalize beyond seen data. Outlining generalization guarantees for a given model class can profoundly impact its applicability. Another relevant aspect concerns the functions a model class can approximate. In this regard, Joshi et al. (2023) have recently analyzed the expressive power of EGNNs in terms of a geometric version of the Weisfeiler-Leman isomorphism test (or 1-WL test) (Weisfeiler & Lehman, 1968) — which has been extensively used to study the expressivity of regular GNNs (Maron et al., 2019; Morris et al., 2019; Sato et al., 2019; Xu et al., 2019). In contrast, establishing theoretical guarantees for the generalization capability of EGNNs remains an open problem.

In this paper, we study the generalization of the EGNNs

[1]Department of Computer Science, Aalto University, Finland [2]Federal Institute of Ceará (Brazil) [3]YaiYai Ltd. Correspondence to: Rafał Karczewski <rafal.karczewski@aalto.fi>, Vikas Garg <vgarg@csail.mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

from a theoretical perspective and derive a high probability bound for their generalization error using the PAC-learning framework (Valiant, 1984; Mohri et al., 2012). Our analysis reveals three new insights: (i) the generalization gap depends on the weighted sum of logarithms of spectral norm of weight matrices; (ii) bottom layers have higher weight (aligning with common knowledge that bottom layers generalize better); and (iii) ε-normalization is essential to obtain a bound polynomial in depth instead of exponential. We compare this bound with existing results on Multilayer Perceptrons (Bartlett et al., 2017; Neyshabur et al., 2018).

Furthermore, we validate our results empirically on a real-world problem of molecular property prediction (Ramakrishnan et al., 2014), a task particularly well-suited for EGNNs. Specifically, we first establish that our theoretical bound highly correlates with the empirical one across different model hyperparameters. Second, we support our claims that ε-normalization reduces the generalization gap when increasing the depth of the model.

Finally, inspired by our theoretical findings, we propose a new regularization method. We evaluate it experimentally and compare it with a commonly used regularization based on the spectral norm. We find that ours leads to lower test loss and generalization gap across different tasks and choices of model hyperparameters (e.g., number of layers).

In summary, our main contributions are: i) we derive the first generalization bounds for E(n)-Equivariant Graph Neural Networks; ii) we validate our theoretical analysis with experiments on real-world data; iii) we show theoretically and empirically that ε-normalization helps generalization; iv) we propose a new regularization method tailored to EGNNs and assess its performance on twelve regression tasks.

2. Related Work

Expressivity and Generalization of GNNs. Understanding the expressivity and generalization capabilities of Graph Neural Networks (GNNs) is crucial for their application across diverse domains. Xu et al. (2019) demonstrated the potential of GNNs to capture complex graph structures, setting a benchmark for their expressivity. However, challenges in expressivity and generalization are highlighted by Oono & Suzuki (2020) and Loukas (2020), who show that GNNs can lose expressive power or face limitations based on their depth and width. Theoretical advances by Barceló et al. (2020) and Garg et al. (2020) further dissect the logical expressiveness and representational boundaries of GNNs, respectively. Recent studies, such as (Tang & Liu, 2023) and (Yang et al., 2023), offer new insights into GNNs' generalization, suggesting inherent advantages of GNN architectures over MLPs in graph learning tasks.

Expressivity of geometric GNNs. Joshi et al. (2023) develops a geometric Weisfeiler-Leman (GWL) test to evaluate the expressivity of geometric GNNs, revealing how equivariant and invariant layers affect their ability to distinguish complex geometric graphs. Meanwhile, Wang et al. (2024) introduces ViSNet, an efficient model that enhances molecular structure modeling through geometric information, showcasing improved performance on molecular dynamics benchmarks.

Equivariant neural networks. Equivariant Neural Networks have reshaped model design by embedding data symmetries directly into network architectures. Introduced by Cohen & Welling (2016), these networks ensure equivariance to group actions, a concept expanded in Euclidean spaces (Weiler & Cesa, 2019; Satorras et al., 2021) and Lie groups (Finzi et al., 2020). For 3D data, Tensor Field Networks (Thomas et al., 2018) and SE(3)-Transformers (Fuchs et al., 2020) exemplify their utility. The approach is theoretically deepened by Lang & Weiler (2021) and adapted to dynamics with imperfect symmetries (Wang et al., 2022) and particle physics (Bogatskiy et al., 2020).

Generalization of equivariant GNNs. Recent studies have advanced our understanding of the generalization capabilities of equivariant GNNs. Petrache & Trivedi (2023) explore approximation-generalization trade-offs under group equivariance, while Behboodi et al. (2022) establish a PAC-Bayesian generalization bound for these networks. Bulusu et al. (2021) focus on translation-equivariant neural networks, and Kondor & Trivedi (2018) consider equivariance to the action of compact groups. Additionally, Elesedy & Zaidi (2021) demonstrate a strict generalization benefit for equivariant models, and Liu et al. (2023) highlight how physical inductive biases can enhance generalization. Sannai et al. (2021) further provide improved generalization bounds through quotient feature spaces.

3. Background

This section overviews $E(n)$-equivariant graph neural networks and provides definitions and results from learning theory we leverage in our analysis. For readability, we summarize the notation used in this work in Appendix A.

3.1. E(n)-Equivariant Graph Neural Network

We consider *geometric graphs* and denote them by $G = (\mathcal{V}, \mathcal{E}, \mathcal{C}, \mathcal{Z})$, where $\mathcal{V}$ is the set of vertices, $\mathcal{E} \subseteq \mathcal{V}^2$ is the set of edges, $\mathcal{C} = \{c_v\}_{v\in\mathcal{V}} \subset \mathbb{R}^{d_c}$ is an indexed set of vertex attributes (or colors), and $\mathcal{Z} = \{z_v\}_{v\in\mathcal{V}} \subset \mathbb{R}^{d_z}$ is an indexed set of vertex *coordinates*. The neighborhood of a vertex v is given by $\mathcal{J}(v) = \{u : (u, v) \in \mathcal{E}\}$.

Let $\mathfrak{G}$ be a group acting on two sets $\mathcal{X}$ and $\mathcal{X}'$ by the representations Φ and Ψ, respectively. We say a function

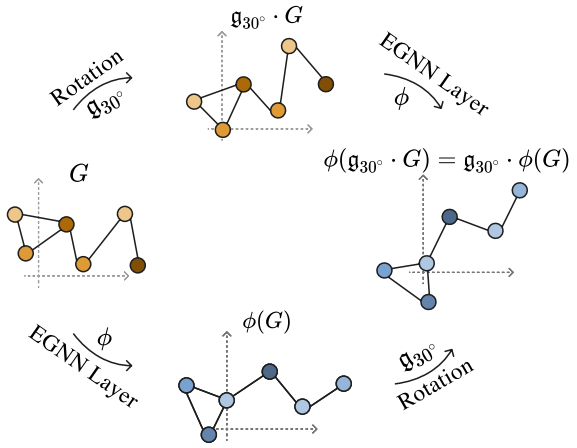

Figure 1. Visualization of rotation equivariance of the EGNN model. Colors depict node features.

$f : \mathcal{X} \to \mathcal{X}'$ is $\mathfrak{G}$-equivariant if it commutes with the group actions, i.e., for all $\mathfrak{g} \in \mathfrak{G}$ and $x \in \mathcal{X}$, we have that $f(\Phi(\mathfrak{g})x) = \Psi(\mathfrak{g})f(x)$. Here, we are interested in graph models that are equivariant to: i) the Euclidean group $E(n)$, which comprises all translations, rotations, and reflections of the n-dimensional Euclidean space; ii) the permutation group Σ_n. To represent actions of these groups on graphs, it is convenient to assume (WLOG) that $\mathcal{V} = \{1, 2, \dots, n\}$, the sets features $\mathcal{C}$, $\mathcal{Z}$ are organized as matrices $C \in \mathbb{R}^{n \times d_c}$, $Z \in \mathbb{R}^{n \times d_z}$, and the graph connectivity is given by the adjacency matrix $A \in \{0, 1\}^{n \times n}$ — in this case, we denote graphs as $G = (A, C, Z)$. Thus, any $\mathfrak{g} \in \Sigma_n$ can be represented by a corresponding permutation matrix $P_\mathfrak{g}$ that acts on G by $(P_\mathfrak{g} A P_\mathfrak{g}^\top, P_\mathfrak{g} C, P_\mathfrak{g} Z)$. On the other hand, each element $\mathfrak{g} \in E(n)$ (or more precisely $E(d_z)$ in our case) is a translation (represented by a vector $t_\mathfrak{g} \in \mathbb{R}^{d_z \times 1}$) followed by a linear transformation via an orthogonal matrix $Q_\mathfrak{g} \in \mathbb{R}^{d_z \times d_z}$ that acts on G by $(A, C, (Z + 1_n t_\mathfrak{g}^\top) Q_\mathfrak{g})$, where 1_n is a n-dimensional column vector of ones.

$E(n)$-Equivariant GNNs (EGNNs, Satorras et al., 2021) are arguably the most popular models for geometric graphs. The basic idea consists of modifying message-passing GNNs by incorporating distances between vertices based on the geometric features into messages and recursively updating the geometric features in an $E(n)$-equivariant fashion. This way, EGNNs maintain the permutation equivariance of GNN layers while also being equivariant to $E(n)$.

Let $h_v^{(0)} = c_v$ and $z_v^{(0)} = z_v$ $\forall v \in \mathcal{V}$. Also, assume that each edge $(u, v) \in \mathcal{E}$ has an associated feature $a_{uv} \in \mathbb{R}^{d_e}$. At each layer $\ell = 0, 1, \dots, L_{\text{egnn}} - 1$, EGNNs compute the incoming **messages** to each vertex v from its neighbors $u \in \mathcal{J}(v)$ as

$$\mu_{u \to v}^{(\ell)} = \phi_\mu^\ell \left(h_u^{(\ell)}, h_v^{(\ell)}, \|z_u^{(\ell)} - z_v^{(\ell)}\|, a_{uv} \right), \tag{1}$$

where ϕ_μ^ℓ is commonly parameterized by multilayer perceptrons (MLPs). Whenever an MLP has multiple inputs, we assume that they are concatenated into a single vector as in the original implementation [1].

Next, the messages to each vertex v are used to recursively **update its coordinates** using an auxiliary MLP ϕ_z^ℓ and a normalization term $\gamma(z_u^{(\ell)}, z_v^{(\ell)}, \varepsilon)$ as

$$z_v^{(\ell+1)} = z_v^{(\ell)} + \frac{1}{|\mathcal{J}(v)|} \sum_{u \in \mathcal{J}(v)} \frac{z_v^{(\ell)} - z_u^{(\ell)}}{\gamma(z_u^{(\ell)}, z_v^{(\ell)}, \varepsilon)} \phi_z^\ell(\mu_{u \to v}^{(\ell)}) \tag{2}$$

and then combined using **sum aggregation**:

$$\mu_v^{(\ell)} = \sum_{u \in \mathcal{J}(v)} \mu_{u \to v}^{(\ell)}. \tag{3}$$

We then recursively **update** the embedding of vertex v using a MLP ϕ_h^ℓ as

$$h_v^{(\ell+1)} = \phi_h^\ell(h_v^{(\ell)}, \mu_v^{(\ell)}). \tag{4}$$

Remark 3.1 (ε-normalization). In their original formulation (Satorras et al., 2021), EGNNs do not include normalization, i.e. $\gamma(z_u^{(\ell)}, z_v^{(\ell)}, \varepsilon) \equiv 1$. In our analysis, we consider *ε-normalization*, i.e. $\gamma(z_u^{(\ell)}, z_v^{(\ell)}, \varepsilon) = \|z_u^{(\ell)} - z_v^{(\ell)}\|_2 + \varepsilon$. This variant is available in the original implementation, but not discussed in the manuscript. We will see later in Section 5 that ε-normalization plays a crucial role regarding the generalization of EGNNs.

In graph-level prediction tasks, we often obtain a representation for the entire graph by applying a *mean readout* function to the output of the last EGNN layer, i.e.,

$$h_G = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} h_v^{(L_{\text{egnn}})}. \tag{5}$$

Finally, we send the graph embedding h_G through a final MLP ϕ_{out} to achieve the graph-level prediction

$$g(G) = \phi_{\text{out}}(h_G). \tag{6}$$

Hereafter, we refer to the full mapping $g(\cdot)$ (i.e., EGNN + final MLP) as the *scoring model*.

3.2. Generalization bounds via Rademacher Complexity

The first important notion is that of generalization error (or gap), defined as the difference between the expected (or true) and empirical risks w.r.t. an arbitrary loss function.

[1] github.com/vgsatorras/egnn/blob/main/models/gcl.py, lines 203, 216

Definition 3.1 (Generalization Error)**.** Let $f : \mathcal{X} \to \mathcal{Y}$ and $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function. Let $S = \{(x_i, y_i)\}_{i=1}^m$ be a finite collection of i.i.d. samples from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The generalization error of f is defined as the difference between the expected loss and the sample loss:

$$\mathcal{R}_{S,\mathcal{L}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f(x), y)] - \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(f(x_i), y_i).$$

Deriving generalization bounds for a function class often involves obtaining some measure of its size (or capacity). In this regard, the *Empirical Rademacher Complexity* (ERC) is one of the most popular tools. In particular, ERC measures how well a function class can fit random noise.

Definition 3.2 (Empirical Rademacher complexity)**.** Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a class of bounded functions and $S = \{x_i\}_{i=1}^m \subseteq \mathcal{X}$ a fixed set of size m. The empirical Rademacher complexity of $\mathcal{F}$ with respect to S is

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(x_i) \right], \tag{7}$$

where $\sigma = [\sigma_1, \ldots, \sigma_m]$ is random vector of i.i.d. random variables such that $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 0.5$.

Notably, a fundamental result in learning theory bounds the generalization error in terms of the ERC (Theorem 3.1).

Theorem 3.1 (Mohri et al. (2012))**.** *Let $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be loss function and $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ a class of functions. Then for any $\delta > 0$, with probability at least $1 - \delta$ over choosing a m-sized sample $S \sim \mathcal{D}^m$ from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the following holds for any $f \in \mathcal{F}$:*

$$\mathcal{R}_{S,\mathcal{L}}(f) \le 2\widehat{\mathfrak{R}}_S(\mathcal{F}_\mathcal{L}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}, \tag{8}$$

where

$$\mathcal{F}_\mathcal{L} = \{(x, y) \mapsto \mathcal{L}(f(x), y) : f \in \mathcal{F}\}.$$

From Theorem 3.1, finding a generalization bound reduces to bounding the ERC. We will use standard tools for bounding ERC, for which we need to introduce a concept of a covering number.

Definition 3.3 (Covering number)**.** Let Θ be a set and $\|\cdot\|$ be a norm. We say that Θ is r-covered by a set Θ', with respect to $\|\cdot\|$, if for all $\theta \in \Theta$ there exists $\theta' \in \Theta'$ with $\|\theta - \theta'\| \le r$. We define the covering number of Θ as the cardinality of the smallest Θ' that r-covers Θ, and denote it by $\mathcal{N}(\Theta, r, \|\cdot\|)$.

Importantly, the result in Lemma 3.1 relates the ERC of a function class with its covering number.

Lemma 3.1 (Bartlett et al. (2017))**.** *Let $\mathcal{F} \subseteq [-\beta, \beta]^{\mathcal{X}}$ be a class of functions taking values in $[-\beta, \beta]$. Also, assume that $f_0 \in \mathcal{F}$, where $f_0(x) = 0\ \forall x \in \mathcal{X}$. Define $\|f\|_\infty = \sup_{x\in\mathcal{X}} \|f(x)\|_2$. Then, for any set $S = \{x_i\}_{i=1}^m \subseteq \mathcal{X}$*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \le \inf_{\alpha>0} \left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_\alpha^{2\beta\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}, r, \|\cdot\|_\infty)}\, dr \right). \tag{9}$$

We will analyze function classes parametrized by weight matrices, and the following result regarding the bound of the covering number of sets of matrices will be useful.

Lemma 3.2 (Chen et al. (2020))**.** *Let $\mathcal{W} = \{W \in \mathbb{R}^{d_1\times d_2} : \|W\|_2 \le \lambda\}$ be a set of matrices with bounded spectral norm, and $r > 0$ a constant. The covering number of $\mathcal{W}$ can be bounded in terms of the Frobenius norm $\|\cdot\|_F$ as*

$$\mathcal{N}(\mathcal{W}, r, \|\cdot\|_F) \le \left(1 + 2\frac{\min\{\sqrt{d_1}, \sqrt{d_2}\}\lambda}{r}\right)^{d_1 d_2}. \tag{10}$$

4. Main results

To derive our generalization bound, we make the following mild assumptions.

Assumption 4.1 (Inputs are bounded)**.** The elements of the input graphs are contained in an Euclidean ball with a radius β. More specifically, there exists a $\beta \ge 1$ such that for all graphs $G = (\mathcal{V}, \mathcal{E}, \mathcal{C}, \mathcal{Z})$, and all $v \in \mathcal{V}$ and $(v, u) \in \mathcal{E}$ with feature vector a_{vu}, we have that

$$\max\{\|c_v\|_2, \|z_v - z_u\|_2, \|a_{uv}\|_2\} \le \beta.$$

Assumption 4.2 (EGNNs are parametrized with MLPs)**.** For all $\ell = 0, \ldots, L_{\text{egnn}}$, the functions ϕ_z^ℓ, ϕ_h^ℓ, and ϕ_μ^ℓ are MLPs with identical number of layers, denoted by L_ϕ. The scoring model ϕ_{out} is also an MLP with L_{out} layers. In addition, all activation functions $\psi(\cdot)$ are K_ψ-Lipschitz, for some $K_\psi \ge 1$, and $\psi(0) = 0$.

Assumption 4.3 (Weights have bounded spectral norm)**.** Let W_i^ϕ denote the weight matrix of the i-th (linear) layer of the MLP ϕ. For all layers i of the MLPs $\phi \in \{\phi_z^\ell\}_{\ell=0}^{L_{\text{egnn}}} \cup \{\phi_h^\ell\}_{\ell=0}^{L_{\text{egnn}}} \cup \{\phi_\mu^\ell\}_{\ell=0}^{L_{\text{egnn}}} \cup \{\phi_{\text{out}}\}$, there exists a $\beta_{i,\phi} \ge 1$ such that $\|W_i^\phi\|_2 \le \beta_{i,\phi}$.

These assumptions are standard in the generalization literature (Chen et al., 2020; Bartlett et al., 2017). In particular, commonly used activation functions are 1-Lipschitz and vanish at 0 (e.g. ReLU, tanh, LeakyReLU, SiLU, and ELU).

Our derivation of generalization bounds proceeds through the following steps: (i) Show that the scoring function is

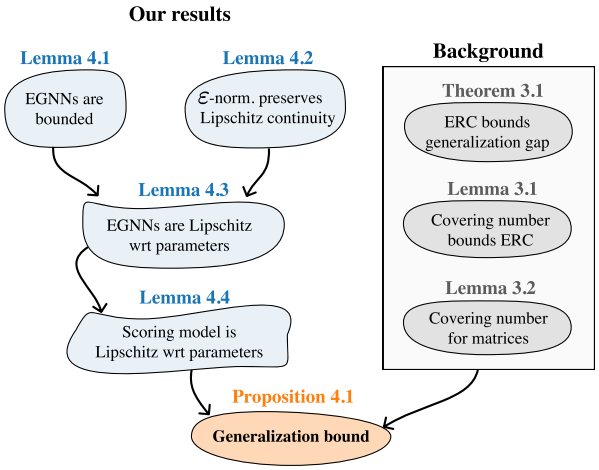

Figure 2. Overview of our results and their dependencies.

Lipschitz continuous w.r.t model parameters; (ii) Show that the Cartesian product of coverings of weight matrices defines a covering of function class of scoring models; and (iii) Use Lemmas 3.2, 3.1 and Theorem 3.1 to establish the generalization bound. We present a diagram of our theoretical contributions and their dependencies in Figure 2.

4.1. Lemmata

We begin by proving that the outputs of the EGNN model remain bounded as long as inputs are bounded:

Lemma 4.1 (Boundedness of EGNN embeddings)**.** *Consider an EGNN model as described in Equations (1)-(4). For any graph G, we have that $\forall v \in \mathcal{V}$, and $u \in \mathcal{J}(v)$:*

$$\max\{\|h_v^{(\ell)}\|_2, \|z_v^{(\ell)} - z_u^{(\ell)}\|_2, D\|\mu_{u\to v}^{(\ell)}\|_2\} \leq C\beta^{(\ell)}, \quad (11)$$

where D is the maximum degree in the graph, $C = 8D\beta$, and $\beta^{(\ell)} = (20D)^\ell \left(\prod_{i=0}^{\ell} M^{(i)}\right)^2$ with

$$M^{(\ell)} = \max_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^{\ell}\}} \prod_{i=1}^{L_\phi} K_\psi \beta_{i,\phi} \quad (12)$$

where $\beta_{i,\phi}$ is the bound of $\|W_\phi^i\|_2$, and L_ϕ, the number of layers of ϕ, and K_ψ the Lipschitz constant of the activation function.

The proof consists of deriving a recursive relation for the bound as a function of the number of layers, determining its growth rate, and an induction argument (Appendix C).

Next, we show that ε-normalization preserves Lipschitz continuity:

Lemma 4.2 (Lipschitz continuity preserved under ε-normalization)**.** *Let f be a K_f-Lipschitz function and $\|\cdot\|$ any norm. Then the ε-normalized function*

$$f_\varepsilon(\cdot) := \frac{f(\cdot)}{\|f(\cdot)\| + \varepsilon}$$

is K_{f_ε}-Lipschitz with $K_{f_\varepsilon} = {}^{3K_f}/_{\varepsilon}$.

The proof involves a telescoping trick and two forms of triangle inequality. See Appendix D for the detailed proof. Now, we move on to the key result that will be used to derive the generalization bounds for the EGNN model, namely Lipschitz continuity of the EGNN node embeddings with respect to model parameters:

Lemma 4.3 (Lipschitz continuity of EGNN wrt params)**.** *Consider EGNNs as defined in Equations (1)-(4). Let $h_v^{(\ell)}(\mathcal{W})$ denote the embedding of node v at layer ℓ produced by an EGNN with parameters $\mathcal{W} = (\mathcal{W}_{\phi_h}, \mathcal{W}_{\phi_z}, \mathcal{W}_{\phi_\mu})$, where $\mathcal{W}_\phi = \{W_i^\phi\}_{i=1}^{L_\phi}$ — recall that W_i^ϕ denotes the weight matrix of the i-th (linear) layer of the MLP ϕ.*

For any two EGNNs with parameters $\mathcal{W}$ and $\tilde{\mathcal{W}}$, we have

$$\|h_v^{(\ell)}(\mathcal{W}) - h_v^{(\ell)}(\tilde{\mathcal{W}})\|_2 \leq CQ^\ell B^{(\ell)} \mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}), \quad (13)$$

where

$$\mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}) = \sum_{\ell'=1}^{\ell} \sum_{\phi \in \{\phi_z^{\ell'-1}, \phi_h^{\ell'-1}, \phi_\mu^{\ell'}\}} \sum_{i=1}^{L_\phi} \|W_i^\phi - \tilde{W}_i^\phi\|_2,$$

D, C, $\beta^{(\ell)}$, $M^{(\ell)}$ are as defined in Lemma 4.1, and

$$Q = 224D^2 \max\left\{\frac{C}{\varepsilon}, 1\right\}$$
$$B^{(\ell)} = \left(\prod_{i=0}^{\ell} M^{(i)}\right)^3 \prod_{i=0}^{\ell-1} \beta^{(i)}, \quad (14)$$

with the convention that an empty product is equal to 1.

The proof is in Appendix E.

Finally, we can show that the scoring model defined in Equation 6 is also Lipschitz continuous w.r.t. its parameters.

Lemma 4.4 (Lipschitz continuity w.r.t. parameters of the scoring model)**.** *Let $\mathcal{W}^{(g)} = (\mathcal{W}, \mathcal{W}_{\phi_{out}})$ denote the parameters of the scoring model g as defined in Equation (6), with $\mathcal{W}$ as defined in Lemma 4.3, $\mathcal{W}_{\phi_{out}} = \{W_i^{\phi_{out}}\}_{i=1}^{L_{out}}$ and $g(G; \mathcal{W}^{(g)})$ the output of the scoring model with parameters $\mathcal{W}^{(g)}$. Then, g is Lipschitz continuous w.r.t. the model parameters:*

$$\|g(G; \mathcal{W}^{(g)}) - g(G; \tilde{\mathcal{W}}^{(g)})\| \leq K_g \mathrm{dist}(\mathcal{W}^{(g)}, \tilde{\mathcal{W}}^{(g)}),$$

where

$$\mathrm{dist}(\mathcal{W}^{(g)}, \tilde{\mathcal{W}}^{(g)}) = \mathrm{dist}_{L_{egnn}}(\mathcal{W}, \tilde{\mathcal{W}}) + \sum_{i=1}^{L_{out}} \|W_i^{\phi_{out}} - \tilde{W}_i^{\phi_{out}}\|_2$$

C, Q, $B^{(\cdot)}$ are as defined in Lemma 4.3 and

$$K_g = 2\left(\prod_{i=1}^{L_{out}} K_\psi \beta_{i,\phi_{out}}\right) CQ^{L_{egnn}} B^{(L_{egnn})}.$$

The proof can be found in Appendix F. We now proceed to our main result on the generalization of the scoring model.

4.2. Generalization bounds

We have developed all the necessary machinery to derive the generalization bound. To simplify notation, we introduce the *stretching factor* of a weight matrix:

$$\kappa(W) \coloneqq \max\{1, \|W\|_2\}, \tag{15}$$

which we will see plays an important role in the bound.

Proposition 4.1 (Generalization bound of the scoring model)**.** *Let $\mathcal{G}$ be the space of geometric graphs as defined in Section 3.1, $\mathcal{Y} = [0,1]$ the space of labels, $g : \mathcal{G} \to \mathbb{R}$ the scoring model as defined in (6) and $\mathcal{L}(\widehat{y}, y) = \min\left\{(\widehat{y} - y)^2, 1\right\}$ the loss function. Then for any $\delta > 0$, with probability at least $1 - \delta$ over choosing a sample $S \sim \mathcal{D}^m$ from a distribution $\mathcal{D}$ over $\mathcal{G} \times \mathcal{Y}$ of size m, the following holds:*

$$\mathcal{R}_{S,\mathcal{L}}(g) = \mathcal{O}\left(\frac{d\sqrt{L}}{\sqrt{m}}\sqrt{\Delta} + \frac{\sqrt{\log\frac{2}{\delta}}}{\sqrt{m}}\right), \tag{16}$$

where $L = 3L_\phi L_{egnn} + L_{out}$ is the total number of weight matrices, d is the maximum width across all layers, and

$$\Delta = \sum_{\ell=0}^{L_{egnn}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^{\ell}\}} \sum_{i=1}^{L_\phi} \gamma_{i,\phi} + \sum_{i=1}^{L_{out}} \gamma_{i,\phi_{out}}$$

$$\gamma_{i,\phi} = w_{i,\phi} \log(\Gamma \cdot \kappa(W_i^\phi))$$

$$w_{i,\phi} = \begin{cases} 1 & \text{for } \phi = \phi_{out} \\ L_{egnn} - \ell + 1 & \text{for } \phi \in \{\phi_h^\ell, \phi_z^\ell, \phi_\mu^\ell\} \end{cases}$$

$$\Gamma = \frac{dmLD\beta K_\psi}{\hat{\varepsilon}}.$$

To prove it, we (i) use the fact that g is Lipschitz continuous w.r.t. weight matrices to show that an appropriately chosen matrix covering yields the covering of the class of scoring models (Lemma G.1); (ii) Use Lemma 3.2 to bound the covering number; (iii) Leverage Lemma 3.1 to bound the ERC and; (iv) Establish the bound via Theorem 3.1. The exact bound together with the detailed proof can be found in Appendix G.

5. Discussion

In this section, we discuss the derived generalization bound and some of its implications.

Table 1. Comparison with existing generalization bounds. We compare with bounds for MLPs (Bartlett et al., 2017; Neyshabur et al., 2018) and Equivariant EGNNs (𝔊-EGNNs) (Behboodi et al., 2022). All values are given in Big-$\mathcal{O}$ notation.

	d (width)	$\mathcal{W}$ (weights)	L (depth)
MLPs (2017)	$\log(d)$	$\prod_{i=1}^{L} \|W_i\|_2 \left(\sum_{i=1}^{L} \left(\frac{\|W_i\|_{2,1}}{\|W_i\|_2}\right)^{2/3}\right)^{3/2}$	1
MLPs (2018)	$\sqrt{d\log(d)}$	$\prod_{i=1}^{L} \|W_i\|_2 \sqrt{\sum_{i=1}^{L} \frac{\|W_i\|_F^2}{\|W_i\|_2^2}}$	$L\sqrt{\log(L)}$
𝔊-EGNNs (2022)	$d\sqrt{\log(d)}$	$\prod_{i=1}^{L} \|W_i\|_2 \sqrt{\sum_{i=1}^{L} \frac{\|W_i\|_F^2}{\|W_i\|_2^2}}$	$L\sqrt{\log(L)}$
EGNNs (ours)	$d\sqrt{\log(d)}$	$\sum_{i=1}^{L} \log(\max\{1, \|W_i\|_2\})$	$L\sqrt{\log(L)}$

Comparison with existing bounds. In Table 1 we compare the derived bound with existing bounds for MLPs (Bartlett et al., 2017; Neyshabur et al., 2018). Most notably, our bound depends on the sum of logarithms of the spectral norms of weight matrices as opposed to their product, which makes it less sensitive to high spectral norms. On the other hand, since our bound ignores weight matrices with a norm lower than 1, it cannot get arbitrarily small. We provide more details on the comparison in Appendix H. Note that it is difficult to directly compare different bounds as they make different assumptions (Xu & Wang, 2018).

ε-normalization. In our proofs and experiments we assumed the EGNN layer to be defined by equations (1)-(4) with $\gamma(z_u^{(\ell)}, z_v^{(\ell)}, \varepsilon) \coloneqq \|z_u^{(\ell)} - z_v^{(\ell)}\|_2 + \varepsilon$. The original formulation (Satorras et al., 2021) includes this variant in the implementation, but the manuscript and experiments consider only the unnormalized model corresponding to $\gamma(z_u^{(\ell)}, z_v^{(\ell)}, \varepsilon) \equiv 1$. In subsequent work (Garcia Satorras et al., 2021) it has been found empirically that using the normalized variant with $\varepsilon = 1$ yields more *stable* results.

Note that ε-normalization does not affect the equivariance properties nor the expressivity. The former follows since $\gamma(z_u^{(\ell)}, z_v^{(\ell)}, \varepsilon)$ is $E(n)$-invariant while the latter from $\mu_{u\to v}^{(\ell)}$ depending on $\|z_u^{(\ell)} - z_v^{(\ell)}\|_2$. We now argue theoretically why normalization plays a crucial role in generalization.

Concretely, following the same reasoning as in the proof of Proposition 4.1, we obtain the bound, which is exponential in the number of layers as opposed to polynomial in the case of the normalized model. See Appendix I for more details. We provide additional empirical evidence in Section 6.

Spectral norm. Proposition 4.1 shows that the generalization gap depends on the spectral norm of weight matrices, which is generally well known (Bartlett et al., 2017; Neyshabur et al., 2018; Behboodi et al., 2022). However, our derivation of the bound for EGNNs reveals new insights. First, not all layers contribute equally. From Equation 16 it can be seen that the bottom layers have a higher weight than the top layers. This aligns with the intuition that bottom

layers learn more abstract, high-level information, which makes them generalize better as opposed to top layers learning more fine-grained, task-specific information.

Second, the dependence on the spectral norms is logarithmic. This implies that the bound is less sensitive to outliers than e.g., the sum of spectral norms. Furthermore, from the definition of κ (Equation 15), we see that weight matrices with spectral norms lower than 1 are ignored, implying that decreasing the norms further yields no additional gains.

Dependence on the training set. We note that the spectral norms of the weight matrices depend on the (size of the) training set, and in general, one cannot decouple this dependence for an already trained network, e.g., we cannot assume the spectral norm would not change as we vary the number of examples while evaluating the effect of sample complexity (i.e., size of the training set size) on generalization. The dependence of the spectral norm on the training set, learning algorithm, etc. was not considered in this study.

Spectral regularization. Multiple regularization methods relying on the spectral norm have been proposed, including the sum of squared spectral norms (Yoshida & Miyato, 2017) or explicitly enforcing the spectral norm to be 1 (Miyato et al., 2018). We propose a new regularization method that leverages the findings unique to our study, i.e. explicit minimization of the generalization gap:

$$R(\mathcal{W}) = \sum_{i,\phi} w_{i,\phi} \log(\kappa(W_i^{\phi})). \tag{17}$$

Since $\log$ is concave, this can intuitively be considered an example of *concave regularization* (Zhang & Zhang, 2012), which generally favors sparse solutions. To see this, note that the derivative of $\log$ is $\frac{1}{x}$, and therefore gradient-based optimization will prioritize decreasing norms of weight matrices, whose norm is the lowest. However, note that in our context, instead of $\log$ we use $\log \circ \kappa$, which is no longer concave in the entire domain. We empirically evaluate R in Section 6.

Impact of equivariance. Since the EGNN network without the equivariant update (Equation 2) reduces to a regular message passing GNN, we can compare our bound to the Rademacher bound for message passing GNNs (Garg et al., 2020). To make that comparison possible, we need to make the following simplifications to our setting: we assume that (1) each MLP has only a single hidden layer; (2) different GNN layers share the same weight matrices; and (3) there is no output MLP ϕ_{out}.

We include three variants of the MP-GNN bound: MP-GNN$_a$, MP-GNN$_b$, and MP-GNN$_c$, because Garg et al. (2020) report a different dependence on various parameters, depending on the value of the *percolation compexity*, which depends on the Lipschitz constants of non-linearities and the spectral norm. See Table 2 for the comparison.

We observe that the EGNN bound enjoys a better dependence on the node degree D, the spectral norm of the weights W, and possibly better (but certainly not worse) on the node depth d. However, it has a worse dependence on the depth L. We leave further in-depth analysis of the impact of $E(n)$-equivariance for future work.

Table 2. Comparison with message-passing GNNs (MP-GNNs) (Garg et al., 2020). The a, b, c subscripts denote different model variants. All values are given in Big-$\mathcal{O}$ notation.

	d (width)	D (node degree)	$\mathcal{W}$ (weights)	L (depth)
MP-GNN$_a$	$d\sqrt{\log d}$	$D\sqrt{\log D}$	$\kappa(W)^3\sqrt{\log \kappa(W)}$	$\sqrt{\log L}$
MP-GNN$_b$	$d\sqrt{\log d}$	$D\sqrt{\log D}$	$\kappa(W)^3\sqrt{\log \kappa(W)}$	$L\sqrt{\log L}$
MP-GNN$_c$	$d\sqrt{d \log d}$	$D\sqrt{\log D}$	$\kappa(W)^3\sqrt{\log \kappa(W)}$	$\sqrt{L \log L}$
EGNNs (ours)	$d\sqrt{\log d}$	$\sqrt{\log D}$	$\sqrt{\log \kappa(W)}$	$L\sqrt{L \log L}$

6. Experiments

This section substantiates our theoretical analysis with experiments on real-world regression tasks. In particular, we first consider the generalization behavior of EGNNs as a function of the model variables (e.g., number of layers). We demonstrate that our bounds highly correlate with empirical generalization gaps. We also empirically assess the beneficial impact of the ε-normalization on generalization, validating our findings. Lastly, we show the efficacy of regularized EGNNs obtained from our analysis, highlighting performance gains compared to another regularization method and empirical risk minimizers (no regularization). Our code is available at `https://github.com/Aalto-QuML/GeneralizationEGNNs`.

Evaluation setup. We consider four molecular property prediction tasks from the QM9 dataset (Ramakrishnan et al., 2014). From the data split available in (Satorras et al., 2021; Anderson et al., 2019), we select a subset of 2K molecules for training and use the entire val/test partitions with approximately 17K and 13K molecules, respectively. We train all models for 1000 epochs using the Adam optimizer. We run five independent trials with different seeds. We provide further implementation details in Appendix J.

Empirical vs. theoretical generalization gaps. To demonstrate the practical relevance of the bounds extracted from our analysis, we contrast the empirical and theoretical generalization gaps in terms of: i) the spectral norms across epochs; ii) the maximum number of hidden units (hidden dim, d); and iii) the number of EGNN layers (L_{egnn}). We also report Pearson correlation coefficients between the generalization curves. Figure 3 shows the results (mean and standard deviation) over five runs.

Regarding the spectral norm (top row in Figure 3), we ob-

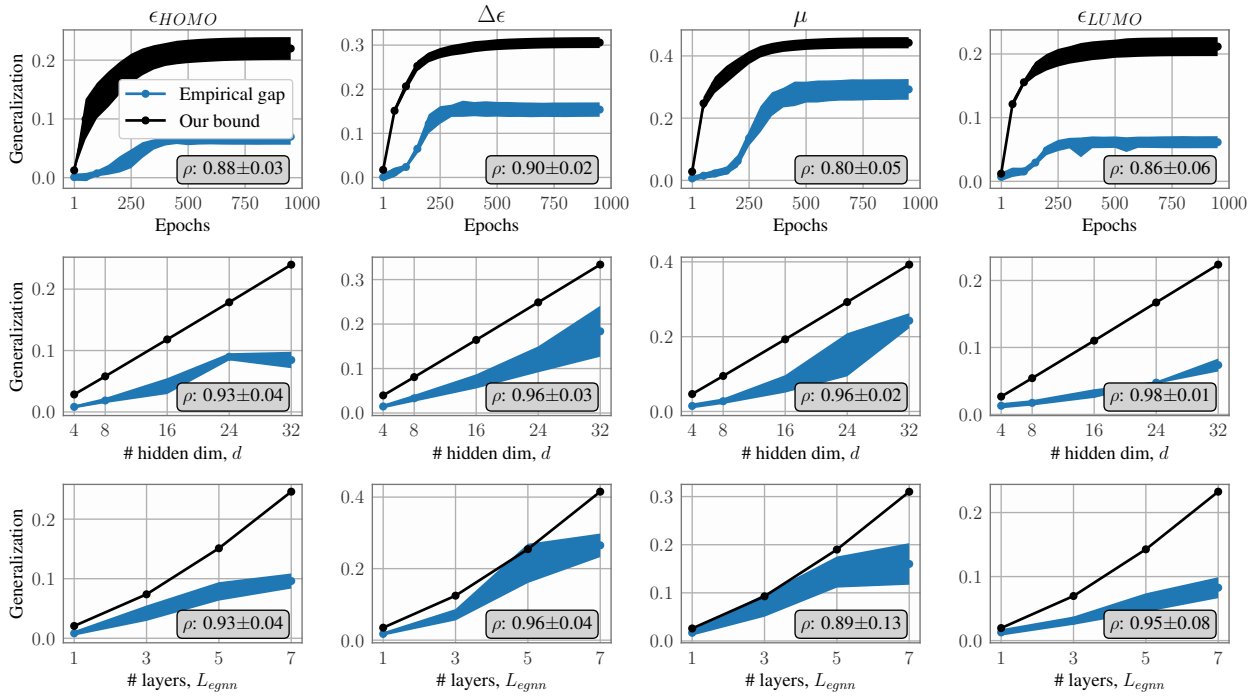

Figure 3. Empirical vs. theoretical generalization gaps as a function of the spectral norms over epochs (top), width (middle), and number of EGNN layers (bottom). Overall, our bounds (black curves) highly correlate with the empirical generalization gap (blue curves). The variables $\varepsilon_{\text{HOMO}}$, $\Delta\varepsilon$, μ, and $\varepsilon_{\text{LUMO}}$ denote molecular properties on the QM9 dataset.

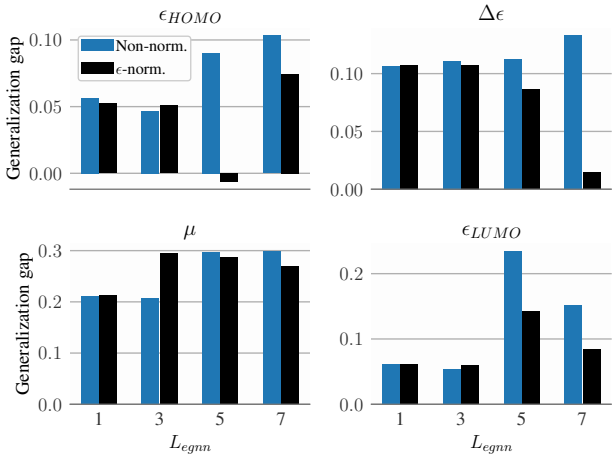

Figure 4. Impact of ε-normalization in terms of the number of layers. Using ε-normalization yields smaller generalization gaps as the depth (L_{egnn}) increases.

serve that our bound can capture the trend of the empirical gap, which settles down at epochs 500 ($\varepsilon_{\text{HOMO}}$ and μ) or 250 ($\Delta\varepsilon$ and $\varepsilon_{\text{LUMO}}$). In 3 out of 4 tasks, the average correlation is greater than 0.8. We can observe an even better correlation when looking at the dependence on hidden dimensionality and number of layers — in 7 out of 8 cases, the correlation is over 0.9. These outcomes support our theory.

To evaluate the impact of the ε-normalization, we run experiments with and without the normalization for $L_{\text{egnn}} \in \{1, 3, 5, 7\}$. Figure 4 reports bar plots for each property. As our theory suggests, ε-normalization positively affects generalization, especially as we increase the number of layers.

Spectral norm regularization. To assess the effectiveness of our bound on the spectral norm as a regularizer (Equation 17), we compare it against a baseline called SpecAvg that takes the average of the spectral norm over all layers as the regularization term. Using the average (or sum) of spectral norms (or their square) is a common practice (Yoshida & Miyato, 2017). We also report results without regularization. Again, we consider regression tasks from QM9. For both regularizers, we select the optimal penalization factor $\lambda \in \{1, 1\text{e-}3, 1\text{e-}5, 1\text{e-}7\}$ using the validation set. Details are given in Appendix J.

Table 3 shows the results in terms of test error (MSE) and generalization gap for different numbers of layers and tasks. Remarkably, our method significantly outperforms SpecAvg in virtually all cases. Also, our approach was able to produce both smaller test errors and generalization gaps than the baselines. Overall, SpecAvg achieves better results than models without regularization.

Additional results. In Appendix K, we provide additional visualizations and experiments. More specifically, we show

Table 3. Test mean squared error (MSE) and generalization gap on QM9 for different regularization methods. 'None' denotes EGNNs without regularization. We denote the best-performing methods in bold. In almost all cases, employing the method derived from our theoretical analysis leads to the smallest test errors and generalization gaps. Results are multiplied by 100 for better visualization.

		Test MSE			**Generalization gap**		
Task	L_{egnn}	None	SPECAVG	Ours	None	SPECAVG	Ours
$\varepsilon_{\text{HOMO}}$	3	5.72 ± 0.48	**5.43** ± 0.37	5.76 ± 0.51	3.61 ± 0.50	**3.33** ± 0.49	3.71 ± 0.48
	5	9.66 ± 1.32	8.45 ± 1.28	**6.71** ± 0.48	8.77 ± 1.40	7.29 ± 1.42	**3.71** ± 1.56
	7	11.13 ± 0.73	9.25 ± 0.70	**7.90** ± 0.50	10.45 ± 0.90	8.33 ± 1.02	**4.43** ± 1.12
$\Delta\varepsilon$	3	12.40 ± 1.34	12.36 ± 1.34	**11.61** ± 1.09	8.02 ± 1.52	7.94 ± 1.46	**6.00** ± 1.90
	5	25.42 ± 7.53	25.38 ± 7.77	**17.39** ± 1.37	22.92 ± 6.87	22.91 ± 7.04	**10.22** ± 5.17
	7	28.49 ± 4.35	26.28 ± 5.00	**23.93** ± 2.73	27.46 ± 3.83	25.14 ± 4.34	**18.16** ± 7.39
μ	3	30.56 ± 1.18	30.03 ± 0.98	**29.56** ± 0.47	8.11 ± 1.96	7.78 ± 1.47	**5.76** ± 1.76
	5	30.94 ± 1.95	30.48 ± 2.26	**28.43** ± 0.26	12.91 ± 2.27	11.80 ± 1.38	**8.47** ± 2.85
	7	31.73 ± 2.46	31.35 ± 2.59	**28.86** ± 0.73	18.42 ± 1.71	17.40 ± 1.66	**9.79** ± 4.45
$\varepsilon_{\text{LUMO}}$	3	5.68 ± 0.45	5.68 ± 0.45	**5.11** ± 0.36	3.07 ± 0.53	3.07 ± 0.53	**2.46** ± 0.41
	5	7.51 ± 1.76	6.74 ± 1.07	**5.85** ± 0.56	6.01 ± 1.68	5.17 ± 0.90	**3.75** ± 1.07
	7	7.87 ± 1.03	7.45 ± 0.73	**7.16** ± 0.90	7.03 ± 1.17	6.71 ± 0.88	**5.58** ± 2.29

the behavior of generalization gaps over training to illustrate that gaps from our method decrease with the values of λ, as expected. In contrast, SPECAVG has little effect in reducing the gap compared to non-regularized models for most penalization values. We also compare our regularizer with SPECAVG on other eight QM9 regression tasks. Overall, our method achieves smaller generalization gaps and competitive MSE values. Lastly, we report the average time per epoch of our regularizer for different tasks to show its computational overhead. For the largest model, the regularized vs. non-regularized time ratio is approximately 1.5.

7. Conclusion

In this work, we analyzed the generalization capabilities of E(n)-Equivariant Graph Neural Networks (EGNNs) from a theoretical perspective. We provided high probability bounds of the generalization error and discussed in detail its implications. Specifically, its logarithmic dependence on the spectral norm of weight matrices, the uneven impact of different layers, and the importance of ε-normalization. We performed extensive experiments to validate our theory.

Additionally, we proposed a novel regularization method inspired by our theoretical results. We show experimentally that it helps reduce both the test loss and the generalization gap and performs better than the commonly used sum of spectral norms of weight matrices.

Acknowledgements

This work has been supported by the Jane and Aatos Erkko Foundation project (grant 7001703) on "Biodesign: Use of artificial intelligence in enzyme design for synthetic biology", and the Finnish Center for Artificial Intelligence FCAI (Flagship program). CSC – IT Center for Science, Finland, provided computational support for this work. We thank the anonymous reviewers for their constructive feedback.

Impact Statement

Incorporating symmetries such as equivariance in deep learning models is crucial for advancing fields like physics-inspired learning, drug and material discovery, protein design, and molecular dynamics simulation. Good generalization performance is a key goal of machine learning models. This work investigates the generalization of equivariant graph neural networks, providing several insights that we hope would foster the design of better models. We do not foresee any particularly negative societal impact of this work.

References

Anderson, B., Hy, T. S., and Kondor, R. Cormorant: Covariant molecular neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Arts, M., Garcia Satorras, V., Huang, C.-W., Zugner, D., Federici, M., Clementi, C., Noé, F., Pinsler, R., and van den Berg, R. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. *Journal of Chemical Theory and Computation*, 19(18):6151–6159, 2023.

Barceló, P., Kostylev, E. V., Monet, M., Pérez, J., Reutter, J., and Silva, J.-P. The logical expressiveness of graph neural networks. In *8th International Conference on Learning Representations (ICLR 2020)*, 2020.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-

normalized margin bounds for neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Behboodi, A., Cesa, G., and Cohen, T. S. A pac-bayesian generalization bound for equivariant networks. *Advances in Neural Information Processing Systems*, 35:5654–5668, 2022.

Bogatskiy, A., Anderson, B., Offermann, J., Roussi, M., Miller, D., and Kondor, R. Lorentz group equivariant neural network for particle physics. In *International Conference on Machine Learning*, pp. 992–1002, 2020.

Bronstein, M. M., Bruna, J., Cohen, T., and Velickovic, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *ArXiv e-prints: 2104.13478*, 2021.

Bulusu, S., Favoni, M., Ipp, A., Müller, D. I., and Schuh, D. Generalization capabilities of translationally equivariant neural networks. *Physical Review D*, 104(7):074504, 2021.

Chen, M., Li, X., and Zhao, T. On generalization bounds of a family of recurrent neural networks. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1233–1243, 2020.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. Spherical cnns. In *International Conference on Learning Representations (ICLR)*, 2018.

Derrow-Pinion, A., She, J., Wong, D., Lange, O., Hester, T., Perez, L., Nunkesser, M., Lee, S., Guo, X., Wiltshire, B., Battaglia, P. W., Gupta, V., Li, A., Xu, Z., Sanchez-Gonzalez, A., Li, Y., and Velickovic, P. Eta prediction with graph neural networks in google maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 3767–3776, 2021.

Elesedy, B. and Zaidi, S. Provably strict generalisation benefit for equivariant models. In *International Conference on Machine Learning*, pp. 2959–2969. PMLR, 2021.

Finzi, M., Stanton, S., Izmailov, P., and Wilson, A. G. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pp. 3165–3176, 2020.

Fu, T., Gao, W., Coley, C., and Sun, J. Reinforced genetic algorithm for structure-based drug design. *Advances in Neural Information Processing Systems*, 35:12325–12338, 2022.

Fuchs, F., Worrall, D., Fischer, V., and Welling, M. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4), 1980.

Garcia Satorras, V., Hoogeboom, E., Fuchs, F., Posner, I., and Welling, M. E(n) equivariant normalizing flows. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4181–4192. Curran Associates, Inc., 2021.

Garg, V. K., Jegelka, S., and Jaakkola, T. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, 2020.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML)*, 2017.

Gori, M., Monfardini, G., and Scarselli, F. A new model for learning in graph domains. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2005.

Joshi, C. K., Bodnar, C., Mathis, S. V., Cohen, T., and Liò, P. On the expressive power of geometric graph neural networks. In *International Conference on Machine Learning*, 2023.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Kondor, R. and Trivedi, S. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pp. 2747–2755, 2018.

Lang, L. and Weiler, M. A wigner-eckart theorem for group equivariant convolution kernels. In *International Conference on Learning Representations*, 2021.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems (NIPS)*, 1990.

Liu, Y., Cheng, J., Zhao, H., Xu, T., Zhao, P., Tsung, F., Li, J., and Rong, Y. Improving generalization in equivariant graph neural networks with physical inductive biases. In *The Twelfth International Conference on Learning Representations*, 2023.

Loukas, A. What graph neural networks cannot learn: depth vs width. In *International Conference on Learning Representations*, 2020.

Maron, H., Ben-Hamu, H., Serviansky, H., and Lipman, Y. Provably powerful graph networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN 9780262018258. URL http://www.jstor.org/stable/j.ctt5hhcw1.

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. Weisfeiler and leman go neural: Higher-order graph neural networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems (NeurIPS - Workshop)*, 2017.

Petrache, M. and Trivedi, S. Approximation-generalization trade-offs under (approximate) group equivariance, 2023.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1), 2014.

Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., and Battaglia, P. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning (ICML)*, 2020.

Sannai, A., Imaizumi, M., and Kawano, M. Improved generalization bounds of group invariant/equivariant deep networks via quotient feature spaces. In *Uncertainty in Artificial Intelligence*, pp. 771–780, 2021.

Sato, R., Yamada, M., and Kashima, H. Approximation ratios of graph neural networks for combinatorial problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, 2021.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

Stärk, H., Ganea, O., Pattanaik, L., Barzilay, D., and Jaakkola, T. EquiBind: Geometric deep learning for drug binding structure prediction. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 20503–20521, 2022.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., and Collins, J. J. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688 – 702, 2020.

Tang, H. and Liu, Y. Towards understanding generalization of graph neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 33674–33719, 23–29 Jul 2023.

Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., and Riley, P. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Valiant, L. G. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Verma, Y., Heinonen, M., and Garg, V. ClimODE: Climate and weather forecasting with physics-informed neural ODEs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xuY33XhEGR.

Wang, R., Walters, R., and Yu, R. Approximately equivariant networks for imperfectly symmetric dynamics. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 23078–23091. PMLR, 2022.

Wang, Y., Wang, T., Li, S., He, X., Li, M., Wang, Z., Zheng, N., Shao, B., and Liu, T.-Y. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15 (1):313, 2024.

Weiler, M. and Cesa, G. General e (2)-equivariant steerable cnns. *Advances in neural information processing systems*, 32, 2019.

Weisfeiler, B. and Lehman, A. A. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsia*, 2(9): 12–16, 1968.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Xu, Y. and Wang, X. Understanding weight normalized deep neural networks with rectified linear units. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Yang, C., Wu, Q., Wang, J., and Yan, J. Graph neural networks are inherently good generalizers: Insights by bridging gnns and mlps. In *International Conference on Learning Representations (ICLR)*, 2023.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. Graph convolutional neural networks for web-scale recommender systems. In *International Conference on Knowledge Discovery & Data Mining (KDD)*, 2018.

Yoshida, Y. and Miyato, T. Spectral norm regularization for improving the generalizability of deep learning. *ArXiv*, abs/1705.10941, 2017.

Zhang, C.-H. and Zhang, T. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

A. Notation

Table 4: Summary of notation and abbreviations.

Notation	Description
$\mathcal{L}$	loss function
$G = (\mathcal{V}, \mathcal{E}, \mathcal{C}, \mathcal{Z})$	Geometric graph
$\mathcal{V}$	Set of nodes in the graph
$\mathcal{E} \subseteq \mathcal{V}^2$	Neighbourhood structure
$\mathcal{C}$	Set of node features
$\mathcal{Z}$	Set of node coordinates
$f \in \mathcal{F}$	f is a function from the function class $\mathcal{F}$
S, m	S is a training set consisting of m input-output samples, i.e., $S = \{(x_i, y_i)\}_{i=1}^m$
$\|W\|_2$	spectral norm of a matrix $W \in \mathbb{R}^{d_1 \times d_2}$
$\|x\|_2$	Euclidean norm of a vector $x \in \mathbb{R}^d$
D	Maximal degree in the graph G
ϕ	Multilayer Perceptron (MLP)
ψ	Activation function
$K_\psi \geq 1$	Lipschitz constant of ψ
W_i^ϕ	weight matrix of i-th layer of ϕ
$\beta_{i,\phi} \geq 1$	bound of the spectral norm of W_i^ϕ, i.e., $\|W_i^\phi\|_2 \leq \beta_{i,\phi}$
L_ϕ, L_{out}, L_{egnn}	Number of layers of ϕ, # layers of the scoring MLP ϕ_{out}, # layers of the EGNN
$\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell$	MLPs in the ℓ-th EGNN layer
$h_v^{(\ell)}, z_v^{(\ell)}, \mu_v^{(\ell)}, \mu_{u\to v}^{(\ell)}$	EGNN embeddings and messages after ℓ layers (Equations (1-4))
$\beta \geq 1$	Bound of the input - Assumption 4.1

B. Lipschitz Continuity of Multilayer Perceptrons

In this section, we show that multilayer perceptrons (MLPs) are Lipschitz continuous, both with respect to their inputs as well as their parameter matrices. We begin with defining the MLP.

Definition B.1 (MLP)**.** A Multi-layer perceptron ϕ with L_ϕ layers and activation function ψ is given by

$$\phi = f_{W_{L_\phi}^\phi} \circ \cdots \circ f_{W_1^\phi},$$

where W_i^ϕ is the i-th weight matrix and f is given by:

$$f_W(x) = \psi(Wx),$$

where ψ is the activation function. Additionally, we will write

$$\phi_k = f_{W_k^\phi} \circ \cdots \circ f_{W_1^\phi}$$

to denote the output of the k-th layer ($\phi \equiv \phi_{L_\phi}$) and

$$\phi(x) = \phi(x; \mathcal{W}_\phi)$$

for $\mathcal{W}_\phi = \{W_i^\phi\}_{i=1}^{L_\phi}$ whenever we want to emphasize which weight matrices were used.

Consider now a function given by

$$f_{(W,b)}(x) = \psi(Wx + b).$$

One can augment the input space by appending a constant 1 to x to obtain $\tilde{x}$ and redefining the weight matrix as $\tilde{W} = \text{concat}(W, b)$. Thus the function can be expressed as

$$f_{(W,b)}(x) = f_{\tilde{W}}(\tilde{x}).$$

Given this representation, we can assume WLOG that $b = 0$.

Lemma B.1 (Boundedness of MLP)**.** *Consider an MLP ϕ with L_ϕ layers, where ψ is the activation function, W_i^ϕ is the i-the weight matrix. Suppose that ψ is K_ψ-Lipschitz and satisfying $\psi(0) = 0$. Then for all x:*

$$\|\phi(x)\|_2 \leq \left(\prod_{i=1}^{L_\phi} K_\psi \|W_i^\phi\|_2\right) \|x\|_2.$$

Proof.

$$\begin{aligned}\|\phi(x)\|_2 = \|\phi_{L_\phi}(x)\|_2 = \|\psi(W_{L_\phi}^\phi \phi_{L_\phi - 1}(x))\|_2 = \|\psi(W_{L_\phi}^\phi \phi_{L_\phi-1}(x)) - \psi(0)\|_2 \leq K_\psi \|W_{L_\phi}^\phi \phi_{L_\phi-1}(x)\|_2 \\ \leq K_\psi \|W_{L_\phi}^\phi\|_2 \|\phi_{L_\phi-1}(x)\|_2 \leq \cdots \leq \|x\|_2 \prod_{i=1}^{L_\phi} K_\psi \|W_i^\phi\|_2.\end{aligned}$$

□

Lemma B.2 (Lipschitz continuity of MLP w.r.t. model parameters)**.** *Consider two L_ϕ-layer MLPs given by two sets of parameters: $\mathcal{W}_\phi$ and $\tilde{\mathcal{W}}_\phi$ sharing the same activation function ψ with Lipschitz constant K_ψ and $\psi(0) = 0$. Assume further that all weight matrices have bounded norm, i.e. for all i, $\max\{\|W_i^\phi\|_2, \|\tilde{W}_i^\phi\|_2\} \leq \beta_{i,\phi}$ for some $\beta_{i,\phi} \geq 1$. Then for all $x, \mathcal{W}_\phi, \tilde{\mathcal{W}}_\phi$:*

$$\|\phi(x; \mathcal{W}_\phi) - \phi(x; \tilde{\mathcal{W}}_\phi)\|_2 \leq \|x\|_2 \left(\prod_{i=1}^{L_\phi} K_\psi \beta_{i,\phi}\right) \text{dist}(\mathcal{W}_\phi, \tilde{\mathcal{W}}_\phi), \tag{18}$$

where

$$\text{dist}(\mathcal{W}_\phi, \tilde{\mathcal{W}}_\phi) \coloneqq \sum_{i=1}^{L_\phi} \|W_i^\phi - \tilde{W}_i^\phi\|_2.$$

Proof. For presentation clarity, we assume throughout this proof $\|\| \equiv \|\|_2$ - spectral norm for matrices and Euclidean norm for vectors.

$$\begin{aligned}\|\phi(x; \mathcal{W}_\phi) - \phi(x; \tilde{\mathcal{W}}_\phi)\| &= \|\psi(W_{L_\phi}^\phi \phi_{L_\phi-1}(x; \mathcal{W}_\phi)) - \psi(\tilde{W}_{L_\phi}^\phi \phi_{L_\phi-1}(x; \tilde{\mathcal{W}}_\phi))\| \\ &\leq K_\psi \|W_{L_\phi}^\phi \phi_{L_\phi-1}(x; \mathcal{W}_\phi) - \tilde{W}_{L_\phi}^\phi \phi_{L_\phi-1}(x; \tilde{\mathcal{W}}_\phi)\|.\end{aligned}$$

Now using the telescoping technique:

$$\begin{aligned}&\|W_{L_\phi}^\phi \phi_{L_\phi-1}(x; \mathcal{W}_\phi) - \tilde{W}_{L_\phi}^\phi \phi_{L_\phi-1}(x; \tilde{\mathcal{W}}_\phi)\| \\ &\leq \|W_{L_\phi}^\phi \phi_{L_\phi-1}(x; \mathcal{W}_\phi) - W_{L_\phi}^\phi \phi_{L_\phi-1}(x; \tilde{\mathcal{W}}_\phi)\| + \|W_{L_\phi}^\phi \phi_{L_\phi-1}(x; \tilde{\mathcal{W}}_\phi) - \tilde{W}_{L_\phi}^\phi \phi_{L_\phi-1}(x; \tilde{\mathcal{W}}_\phi)\| \\ &\leq \|W_{L_\phi}^\phi\| \cdot \|\phi_{L_\phi-1}(x; \mathcal{W}_\phi) - \phi_{L_\phi-1}(x; \tilde{\mathcal{W}}_\phi)\| + \|\phi_{L_\phi-1}(x; \tilde{\mathcal{W}}_\phi)\| \cdot \|W_{L_\phi}^\phi - \tilde{W}_{L_\phi}^\phi\| \\ &\leq \|W_{L_\phi}^\phi\| \cdot \|\phi_{L_\phi-1}(x; \mathcal{W}_\phi) - \phi_{L_\phi-1}(x; \tilde{\mathcal{W}}_\phi)\| + \|x\| \left(\prod_{i=1}^{L_\phi-1} K_\psi \|\tilde{W}_i^\phi\|\right) \|W_{L_\phi}^\phi - \tilde{W}_{L_\phi}^\phi\|.\end{aligned}$$

Therefore

$$\|\phi(x;\mathcal{W}_\phi)-\phi(x;\tilde{\mathcal{W}}_\phi)\| \le \|x\|K_\psi^{L_\phi}\left(\prod_{i=1}^{L_\phi-1}\|\tilde{W}_i^\phi\|\right)\|W_{L_\phi}^\phi-\tilde{W}_{L_\phi}^\phi\| + K_\psi\|W_{L_\phi}^\phi\|\cdot\|\phi_{L_\phi-1}(x;\mathcal{W}_\phi)-\phi_{L_\phi-1}(x;\tilde{\mathcal{W}}_\phi)\|$$

and by induction we get

$$\|\phi(x;\mathcal{W}_\phi)-\phi(x;\tilde{\mathcal{W}}_\phi)\| \le \|x\|K_\psi^{L_\phi}\sum_{i=1}^{L_\phi} a_i\|W_i^\phi-\tilde{W}_i^\phi\|,$$

where

$$a_i = \left(\prod_{j=1}^{i-1}\|\tilde{W}_j^\phi\|\right)\left(\prod_{j'=i+1}^{L_\phi}\|W_{j'}^\phi\|\right) \le \prod_{j=1}^{L_\phi}\beta_{j,\phi}.$$

□

Lemma B.3 (Lipschitz continuity of MLP w.r.t. model input)**.** *Let ϕ be a L_ϕ-layer MLP with a K_ψ-Lipschitz activation function ψ. Then ϕ is Lipschitz continuous with respect to the input:*

$$\|\phi(x)-\phi(y)\|_2 \le \left(\prod_{i=1}^{L_\phi}K_\psi\|W_i^\phi\|_2\right)\|x-y\|_2.$$

Proof. Below we assume $\|\cdot\| \equiv \|\cdot\|_2$.

$$\begin{aligned}\|\phi(x)-\phi(y)\| &= \|\phi_{L_\phi}(x)-\phi_{L_\phi}(y)\| \le K_\psi\|W_{L_\phi}^\phi\|\cdot\|\phi_{L_\phi-1}(x)-\phi_{L_\phi-1}(y)\| \\ &\le \cdots \le \left(\prod_{i=1}^{L_\phi}K_\psi\|W_i^\phi\|\right)\|x-y\|.\end{aligned}$$

□

Lemma B.4 (Multiple inputs MLP)**.** *Suppose ϕ, a L_ϕ-layer MLP takes multiple inputs $x_1,\ldots,x_K$ by concatenating them into a single vector $x = CONCAT([x_1,\ldots,x_K])$. Then the output is bounded:*

$$\|\phi(x_1,\ldots,x_K)\|_2 \le \sqrt{K}\left(\prod_{i=1}^{L_\phi}K_\psi\|W_i^\phi\|_2\right)\sum_{i=1}^{K}\|x_i\|_2.$$

Proof. Claim follows from Lemma B.1 and the inequality:

$$\|x\|_2 = \sqrt{\|x_1\|_2^2+\cdots+\|x_K\|_2^2} \le \sqrt{K}\max_i\|x_i\|_2 \le \sqrt{K}\sum_{i=1}^{K}\|x_i\|_2.$$

□

C. Proof of Lemma 4.1

In this section, we prove Lemma 4.1. We recall it for completeness:

Lemma 4.1 (Boundedness of EGNN embeddings)**.** *Consider an EGNN model as described in Equations (1)-(4). For any graph G, we have that $\forall v \in \mathcal{V}$, and $u \in \mathcal{J}(v)$:*

$$\max\{\|h_v^{(\ell)}\|_2, \|z_v^{(\ell)}-z_u^{(\ell)}\|_2, D\|\mu_{u\to v}^{(\ell)}\|_2\} \le C\beta^{(\ell)}, \tag{11}$$

where D is the maximum degree in the graph, $C = 8D\beta$, and $\beta^{(\ell)} = (20D)^\ell \left(\prod_{i=0}^{\ell} M^{(i)}\right)^2$ with

$$M^{(\ell)} = \max_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell\}} \prod_{i=1}^{L_\phi} K_\psi \beta_{i,\phi} \tag{12}$$

where $\beta_{i,\phi}$ is the bound of $\|W_\phi^i\|_2$, and L_ϕ, the number of layers of ϕ, and K_ψ the Lipschitz constant of the activation function.

Proof. In the proof we assume $\|\cdot\| \equiv \|\cdot\|_2$. First note that (11) implies that for all v and ℓ: $\|\mu_v^{(\ell)}\| \leq C\beta^{(\ell)}$ due to the definition of $\mu_v^{(\ell)}$ as a sum aggregation of $\mu_{u\to v}^{(\ell)}$. We will show the result by induction. Note that, from Lemmas B.3 and B.4:

$$\|\mu_{u\to v}^{(0)}\| = \|\phi_\mu^0\left(h_u^{(0)}, h_v^{(0)}, \|z_u^{(0)} - z_v^{(0)}\|, a_{uv}\right)\| \leq 2(2\beta + \beta + \beta)M^{(0)} \leq \frac{1}{D}CK^{(0)}$$

and thus establishing the base case of $\ell = 0$. Now assume that (11) holds for $\ell' < \ell$.

From Lemmas B.3, B.4 and the inductive hypothesis:

$$\|h_v^{(\ell)}\| \leq \sqrt{2}M^{(\ell)}(\|h_v^{(\ell-1)}\| + \|\mu_v^{(\ell-1)}\|) \leq 2\sqrt{2}M^{(\ell)}C\beta^{(\ell-1)} \leq C\beta^{(\ell)},$$

where the last inequality holds because $2\sqrt{2} \leq 20D$ and $M^{(\ell)} \leq \left(M^{(\ell)}\right)^2$. Similarly for z:

$$\begin{aligned} z_u^{(\ell)} - z_v^{(\ell)} = z_u^{(\ell-1)} - z_v^{(\ell-1)} \\ + \frac{1}{|\mathcal{J}(u)|}\sum_{u' \in \mathcal{J}(u)} \frac{z_u^{(\ell-1)} - z_{u'}^{(\ell-1)}}{\|z_u^{(\ell-1)} - z_{u'}^{(\ell-1)}\| + \varepsilon} \phi_z^{\ell-1}(\mu_{u'\to u}^{(\ell-1)}) \\ - \frac{1}{|\mathcal{J}(v)|}\sum_{v' \in \mathcal{J}(v)} \frac{z_v^{(\ell-1)} - z_{v'}^{(\ell-1)}}{\|z_v^{(\ell-1)} - z_{v'}^{(\ell-1)}\| + \varepsilon} \phi_z^{\ell-1}(\mu_{v'\to v}^{(\ell-1)}) \end{aligned} \tag{19}$$

and therefore:

$$\|z_u^{(\ell)} - z_v^{(\ell)}\| \leq C\beta^{(\ell-1)} + 2M^{(\ell)}C\beta^{(\ell-1)} \leq 3M^{(\ell)}C\beta^{(\ell-1)} \leq C\beta^{(\ell)},$$

because $3 \leq 20D$. Finally for μ:

$$\begin{aligned} \|\mu_{u\to v}^{(\ell)}\| &\leq 2M^{(\ell)}\left(\|h_v^{(\ell)}\| + \|h_u^{(\ell)}\| + \|z_v^{(\ell)} - z_u^{(\ell)}\| + \|a_{uv}\|\right) \\ &\leq 2M^{(\ell)}\left(4\sqrt{2}M^{(\ell)} + 1 + 2M^{(\ell)} + 1\right)C\beta^{(\ell-1)} \\ &\leq 20\left(M^{(\ell)}\right)^2 C\beta^{(\ell-1)} = \frac{1}{D}C\beta^{(\ell)}, \end{aligned}$$

where the last inequality holds because $4 + 4\sqrt{2} \leq 10$.

□

D. Proof of Lemma 4.2

In this section we prove Lemma 4.2. We recall it for completeness:

Lemma 4.2 (Lipschitz continuity preserved under ε-normalization)**.** *Let f be a K_f-Lipschitz function and $\|\cdot\|$ any norm. Then the ε-normalized function*

$$f_\varepsilon(\cdot) \coloneqq \frac{f(\cdot)}{\|f(\cdot)\| + \varepsilon}$$

is K_{f_ε}-Lipschitz with $K_{f_\varepsilon} = {}^{3K_f}/_{\varepsilon}$.

Proof.

$$\begin{aligned}\|f_\varepsilon(x)-f_\varepsilon(y)\| &= \left\|\frac{f(x)}{\|f(x)\|+\varepsilon}-\frac{f(y)}{\|f(y)\|+\varepsilon}\right\| = \left\|\frac{(\|f(y)\|+\varepsilon)f(x)-(\|f(x)\|+\varepsilon)f(y)}{(\|f(x)\|+\varepsilon)(\|f(y)\|+\varepsilon)}\right\| \\ &\le \frac{\varepsilon}{(\|f(x)\|+\varepsilon)(\|f(y)\|+\varepsilon)}\|f(x)-f(y)\| + \left\|\frac{\|f(y)\|f(x)-\|f(x)\|f(y)}{(\|f(x)\|+\varepsilon)(\|f(y)\|+\varepsilon)}\right\| \\ &\le \frac{L_f}{\varepsilon}\|x-y\| + \left\|\frac{\|f(y)\|f(x)-\|f(x)\|f(y)}{(\|f(x)\|+\varepsilon)(\|f(y)\|+\varepsilon)}\right\|.\end{aligned}$$

Focusing on the numerator of the last term:

$$\begin{aligned}\|\|f(y)\|f(x)-\|f(x)\|f(y)\| &= \|\|f(y)\|f(x)-\|f(x)\|f(x)+\|f(x)\|f(x)-\|f(x)\|f(y)\| \\ &\le |\|f(y)\|-\|f(x)\|| \cdot \|f(x)\| + \|f(x)\| \cdot \|f(x)-f(y)\| \\ &\le \|f(x)\|\left(L_f\|x-y\| + |\|f(x)\|-\|f(y)\||\right) \\ &\le 2\|f(x)\|L_f\|x-y\|,\end{aligned}$$

where in the last step we used the fact that $|\|a\|-\|b\|| \le \|a-b\|$. Therefore:

$$\left\|\frac{\|f(y)\|f(x)-\|f(x)\|f(y)}{(\|f(x)\|+\varepsilon)(\|f(y)\|+\varepsilon)}\right\| \le \frac{2\|f(x)\|L_f\|x-y\|}{(\|f(x)\|+\varepsilon)(\|f(y)\|+\varepsilon)} \le \frac{2L_f}{\varepsilon}\|x-y\|$$

and

$$\|f_\varepsilon(x)-f_\varepsilon(y)\| \le \frac{3L_f}{\varepsilon}\|x-y\|.$$

□

E. Proof of Lemma 4.3

In this section we prove Lemma 4.3. We recall it for completeness:

Lemma 4.3 (Lipschitz continuity of EGNN wrt params)**.** *Consider EGNNs as defined in Equations (1)-(4). Let $h_v^{(\ell)}(\mathcal{W})$ denote the embedding of node v at layer ℓ produced by an EGNN with parameters $\mathcal{W} = (\mathcal{W}_{\phi_h}, \mathcal{W}_{\phi_z}, \mathcal{W}_{\phi_\mu})$, where $\mathcal{W}_\phi = \{W_i^\phi\}_{i=1}^{L_\phi}$ — recall that W_i^ϕ denotes the weight matrix of the i-th (linear) layer of the MLP ϕ.*

For any two EGNNs with parameters $\mathcal{W}$ and $\tilde{\mathcal{W}}$, we have

$$\|h_v^{(\ell)}(\mathcal{W}) - h_v^{(\ell)}(\tilde{\mathcal{W}})\|_2 \le CQ^\ell B^{(\ell)} \mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}), \tag{13}$$

where

$$\mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}) = \sum_{\ell'=1}^{\ell} \sum_{\phi \in \{\phi_z^{\ell'-1}, \phi_h^{\ell'-1}, \phi_\mu^{\ell'}\}} \sum_{i=1}^{L_\phi} \|W_i^\phi - \tilde{W}_i^\phi\|_2,$$

D, C, $\beta^{(\ell)}$, $M^{(\ell)}$ are as defined in Lemma 4.1, and

$$\begin{aligned} Q &= 224D^2 \max\left\{\frac{C}{\varepsilon}, 1\right\} \\ B^{(\ell)} &= \left(\prod_{i=0}^{\ell} M^{(i)}\right)^3 \prod_{i=0}^{\ell-1} \beta^{(i)}, \end{aligned} \tag{14}$$

with the convention that an empty product is equal to 1.

Proof. Throughout the proof we assume $\|\cdot\| \equiv \|\cdot\|_2$. We will show the result by induction. We will show a stronger statement, namely that all $h_v^{(\ell)}$, $z_v^{(\ell)}$ and $\mu_{u\to v}^{(\ell)}$ are Lipschitz continuous. The base case obviously holds for $h_v^{(0)}$ and $z_v^{(0)}$ as they do not depend on model parameters.

We see from Lemma B.2 and our assumption that inputs are bounded:

$$\|\mu^{(0)}_{u\to v}(\mathcal{W}) - \mu^{(0)}_{u\to v}(\tilde{\mathcal{W}})\| \leq M^{(0)} 8\beta \mathrm{dist}(\mathcal{W}_{\phi^0_\mu}, \tilde{\mathcal{W}}_{\phi^0_\mu}) \leq \frac{1}{D} C M^{(0)} \mathrm{dist}(\mathcal{W}_{\phi^0_\mu}, \tilde{\mathcal{W}}_{\phi^0_\mu}) \leq \frac{1}{D} C M^{(0)} \mathrm{dist}_0(\mathcal{W}, \tilde{\mathcal{W}}) \quad (20)$$

and thus establishing the base case for induction. Assume now that for all $\ell' < \ell$ there exist $K_h^{(\ell')}, K_z^{(\ell')}, K_\mu^{(\ell')}$ such that for all u, v and all $\mathcal{W}, \tilde{\mathcal{W}}$:

$$\begin{aligned}
\|h_v^{(\ell')}(\mathcal{W}) - h_v^{(\ell')}(\tilde{\mathcal{W}})\| &\leq K_h^{(\ell')} \mathrm{dist}_{\ell'}(\mathcal{W}, \tilde{\mathcal{W}}) \\
\|z_v^{(\ell')}(\mathcal{W}) - z_v^{(\ell')}(\tilde{\mathcal{W}})\| &\leq K_z^{(\ell')} \mathrm{dist}_{\ell'}(\mathcal{W}, \tilde{\mathcal{W}}) \\
\|\mu^{(\ell')}_{u\to v}(\mathcal{W}) - \mu^{(\ell')}_{u\to v}(\tilde{\mathcal{W}})\| &\leq \frac{1}{D} K_\mu^{(\ell')} \mathrm{dist}_{\ell'}(\mathcal{W}, \tilde{\mathcal{W}}).
\end{aligned} \quad (21)$$

Note that (21) implies that $\|\mu_v^{(\ell')}(\mathcal{W}) - \mu_v^{(\ell')}(\tilde{\mathcal{W}})\| \leq K_\mu^{(\ell')} \mathrm{dist}_{\ell'}(\mathcal{W}, \tilde{\mathcal{W}})$. We start with h and proceed with a telescoping argument:

$$\begin{aligned}
&\|h_v^{(\ell)}(\mathcal{W}) - h_v^{(\ell)}(\tilde{\mathcal{W}})\| \\
&= \left\| \phi_h^{\ell-1}(h_v^{(\ell-1)}(\mathcal{W}), \mu_v^{(\ell-1)}(\mathcal{W}); \mathcal{W}_{\phi_h^{\ell-1}}) - \phi_h^{\ell-1}(h_v^{(\ell-1)}(\tilde{\mathcal{W}}), \mu_v^{(\ell-1)}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_h^{\ell-1}}) \right\| \\
&\leq \left\| \phi_h^{\ell-1}(h_v^{(\ell-1)}(\mathcal{W}), \mu_v^{(\ell-1)}(\mathcal{W}); \mathcal{W}_{\phi_h^{\ell-1}}) - \phi_h^{\ell-1}(h_v^{(\ell-1)}(\mathcal{W}), \mu_v^{(\ell-1)}(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_h^{\ell-1}}) \right\| \\
&+ \left\| \phi_h^{\ell-1}(h_v^{(\ell-1)}(\mathcal{W}), \mu_v^{(\ell-1)}(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_h^{\ell-1}}) - \phi_h^{\ell-1}(h_v^{(\ell-1)}(\tilde{\mathcal{W}}), \mu_v^{(\ell-1)}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_h^{\ell-1}}) \right\|
\end{aligned}$$

Now we use Lemmas B.2, B.4 and 4.1 to bound the first term and Lemma B.3 to bound the second

$$\begin{aligned}
&\|h_v^{(\ell)}(\mathcal{W}) - h_v^{(\ell)}(\tilde{\mathcal{W}})\| \\
&\leq \sqrt{2} M^{(\ell)} (\|h_v^{(\ell-1)}(\mathcal{W})\| + \|\mu_v^{(\ell-1)}(\mathcal{W})\|) \mathrm{dist}(\mathcal{W}_{\phi_h^{\ell-1}}, \tilde{\mathcal{W}}_{\phi_h^{\ell-1}}) \\
&+ \sqrt{2} M^{(\ell)} \left(\|h_v^{(\ell-1)}(\mathcal{W}) - h_v^{(\ell-1)}(\tilde{\mathcal{W}})\| + \|\mu_v^{(\ell-1)}(\mathcal{W}) - \mu_v^{(\ell-1)}(\tilde{\mathcal{W}})\| \right) \\
&\leq 2\sqrt{2} M^{(\ell)} C \beta^{(\ell-1)} \mathrm{dist}(\mathcal{W}_{\phi_h^{\ell-1}}, \tilde{\mathcal{W}}_{\phi_h^{\ell-1}}) + \sqrt{2} M^{(\ell)} (K_h^{(\ell-1)} + K_\mu^{(\ell-1)}) \mathrm{dist}_{\ell-1}(\mathcal{W}, \tilde{\mathcal{W}}) \\
&\leq \sqrt{2} M^{(\ell)} (2C\beta^{(\ell-1)} + K_h^{(\ell-1)} + K_\mu^{(\ell-1)}) \mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}).
\end{aligned} \quad (22)$$

Moving on to z:

$$\begin{aligned}
&\|z_v^{(\ell)}(\mathcal{W}) - z_v^{(\ell)}(\tilde{\mathcal{W}})\| \\
&= \|z_v^{(\ell-1)}(\mathcal{W}) - z_v^{(\ell-1)}(\tilde{\mathcal{W}})\| \\
&+ \frac{1}{|\mathcal{J}(v)|} \sum_{u \in \mathcal{J}(v)} \left\| \frac{z_v^{(\ell-1)}(\mathcal{W}) - z_u^{(\ell-1)}(\mathcal{W})}{\|z_v^{(\ell-1)}(\mathcal{W}) - z_u^{(\ell-1)}(\mathcal{W})\| + \varepsilon} \phi_z^{\ell-1}(\mu^{(\ell-1)}_{u\to v}(\mathcal{W}); \mathcal{W}_{\phi_z^{\ell-1}}) \right. \\
&\left. - \frac{z_v^{(\ell-1)}(\tilde{\mathcal{W}}) - z_u^{(\ell-1)}(\tilde{\mathcal{W}})}{\|z_v^{(\ell-1)}(\tilde{\mathcal{W}}) - z_u^{(\ell-1)}(\tilde{\mathcal{W}})\| + \varepsilon} \phi_z^{\ell-1}(\mu^{(\ell-1)}_{u\to v}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_z^{\ell-1}}) \right\| \\
&\leq K_z^{(\ell-1)} \mathrm{dist}_{\ell-1}(\mathcal{W}, \tilde{\mathcal{W}}) \\
&+ \frac{1}{|\mathcal{J}(v)|} \sum_{u \in \mathcal{J}(v)} \left\| \frac{z_v^{(\ell-1)}(\mathcal{W}) - z_u^{(\ell-1)}(\mathcal{W})}{\|z_v^{(\ell-1)}(\mathcal{W}) - z_u^{(\ell-1)}(\mathcal{W})\| + \varepsilon} \right\| \|\phi_z^{\ell-1}(\mu^{(\ell-1)}_{u\to v}(\mathcal{W}); \mathcal{W}_{\phi_z^{\ell-1}}) - \phi_z^{\ell-1}(\mu^{(\ell-1)}_{u\to v}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_z^{\ell-1}})\| \\
&+ \frac{1}{|\mathcal{J}(v)|} \sum_{u \in \mathcal{J}(v)} \|\phi_z^{\ell-1}(\mu^{(\ell-1)}_{u\to v}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_z^{\ell-1}})\| \left\| \frac{z_v^{(\ell-1)}(\mathcal{W}) - z_u^{(\ell-1)}(\mathcal{W})}{\|z_v^{(\ell-1)}(\mathcal{W}) - z_u^{(\ell-1)}(\mathcal{W})\| + \varepsilon} - \frac{z_v^{(\ell-1)}(\tilde{\mathcal{W}}) - z_u^{(\ell-1)}(\tilde{\mathcal{W}})}{\|z_v^{(\ell-1)}(\tilde{\mathcal{W}}) - z_u^{(\ell-1)}(\tilde{\mathcal{W}})\| + \varepsilon} \right\|.
\end{aligned}$$

Now, since (from the inductive assumption) $z_v^{(\ell-1)}, z_u^{(\ell-1)}$ are both Lipschitz continuous with a constant $K_z^{(\ell-1)}$, it follows that $z_v^{(\ell-1)} - z_u^{(\ell-1)}$ is Lipschitz continuous with a constant $2K_z^{(\ell-1)}$. Therefore, using Lemma 4.2, we know that

$\frac{z_v^{(\ell-1)}-z_u^{(\ell-1)}}{\|z_v^{(\ell-1)}-z_u^{(\ell-1)}\|+\varepsilon}$ is Lipschitz continuous with a constant $\frac{6K_z^{(\ell-1)}}{\varepsilon}$ and we can bound:

$$\left\| \frac{z_v^{(\ell-1)}(\mathcal{W}) - z_u^{(\ell-1)}(\mathcal{W})}{\|z_v^{(\ell-1)}(\mathcal{W}) - z_u^{(\ell-1)}(\mathcal{W})\| + \varepsilon} - \frac{z_v^{(\ell-1)}(\tilde{\mathcal{W}}) - z_u^{(\ell-1)}(\tilde{\mathcal{W}})}{\|z_v^{(\ell-1)}(\tilde{\mathcal{W}}) - z_u^{(\ell-1)}(\tilde{\mathcal{W}})\| + \varepsilon} \right\| \leq \frac{6K_z^{(\ell-1)}}{\varepsilon} \mathrm{dist}_{\ell-1}(\mathcal{W}, \tilde{\mathcal{W}}).$$

and consequently:

$$\begin{aligned} \|z_v^{(\ell)}(\mathcal{W}) - z_v^{(\ell)}(\tilde{\mathcal{W}})\| \leq & \left(K_z^{(\ell-1)} + \frac{6K_z^{(\ell-1)}}{\varepsilon} M^{(\ell)} \|\mu_{u\to v}^{(\ell-1)}(\tilde{\mathcal{W}})\| \right) \mathrm{dist}_{\ell-1}(\mathcal{W}, \tilde{\mathcal{W}}) \\ & + \frac{1}{|\mathcal{J}(v)|} \sum_{u \in \mathcal{J}(v)} \|\phi_z^{\ell-1}(\mu_{u\to v}^{(\ell-1)}(\mathcal{W}); \mathcal{W}_{\phi_z^{\ell-1}}) - \phi_z^{\ell-1}(\mu_{u\to v}^{(\ell-1)}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_z^{\ell-1}})\|. \end{aligned} \tag{23}$$

To bound the last term we perform additional telescoping step:

$$\begin{aligned} \|\phi_z^{\ell-1}(\mu_{u\to v}^{(\ell-1)}(\mathcal{W}); \mathcal{W}_{\phi_z^{\ell-1}}) - \phi_z^{\ell-1}(\mu_{u\to v}^{(\ell-1)}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_z^{\ell-1}})\| & \\ \leq \|\phi_z^{\ell-1}(\mu_{u\to v}^{(\ell-1)}(\mathcal{W}); \mathcal{W}_{\phi_z^{\ell-1}}) - \phi_z^{\ell-1}(\mu_{u\to v}^{(\ell-1)}(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_z^{\ell-1}})\| & \\ + \|\phi_z^{\ell-1}(\mu_{u\to v}^{(\ell-1)}(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_z^{\ell-1}}) - \phi_z^{\ell-1}(\mu_{u\to v}^{(\ell-1)}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_z^{\ell-1}})\| & \\ \leq M^{(\ell)} \|\mu_{u\to v}^{(\ell-1)}(\mathcal{W})\| \mathrm{dist}(\mathcal{W}_{\phi_z^{\ell-1}}, \tilde{\mathcal{W}}_{\phi_z^{\ell-1}}) + M^{(\ell)} \|\mu_{u\to v}^{(\ell-1)}(\mathcal{W}) - \mu_{u\to v}^{(\ell-1)}(\tilde{\mathcal{W}})\|. & \end{aligned} \tag{24}$$

Now from Lemma 4.1 we have $\|\mu_{u\to v}^{(\ell-1)}(\mathcal{W})\| \leq C\beta^{(\ell-1)}$ and by applying (24) to (23), we obtain:

$$\begin{aligned} \|z_v^{(\ell)}(\mathcal{W}) - z_v^{(\ell)}(\tilde{\mathcal{W}})\| &\leq \left(K_z^{(\ell-1)} + \frac{6K_z^{(\ell-1)}}{\varepsilon} M^{(\ell)} C\beta^{(\ell-1)} + M^{(\ell)} C\beta^{(\ell-1)} + M^{(\ell)} K_\mu^{(\ell-1)} \right) \mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}) \\ &\leq M^{(\ell)} \left(C\beta^{(\ell-1)} + K_\mu^{(\ell-1)} + K_z^{(\ell-1)} + \frac{6K_z^{(\ell-1)}}{\varepsilon} C\beta^{(\ell-1)} \right) \mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}) \end{aligned}$$

Finally, for μ we will simplifiy notation with $\omega_{uv}^{(\ell)}(\mathcal{W}) \coloneqq (h_u^{(\ell)}(\mathcal{W}), h_v^{(\ell)}(\mathcal{W}), \|z_u^{(\ell)}(\mathcal{W}) - z_v^{(\ell)}(\mathcal{W})\|, a_{uv})$:

$$\begin{aligned} \left\| \mu_{u\to v}^{(\ell)}(\mathcal{W}) - \mu_{u\to v}^{(\ell)}(\tilde{\mathcal{W}}) \right\| &= \left\| \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\mathcal{W}); \mathcal{W}_{\phi_\mu^\ell} \right) - \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_\mu^\ell} \right) \right\| \\ &= \left\| \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\mathcal{W}); \mathcal{W}_{\phi_\mu^\ell} \right) - \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_\mu^\ell} \right) + \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_\mu^\ell} \right) - \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_\mu^\ell} \right) \right\| \\ &\leq \left\| \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\mathcal{W}); \mathcal{W}_{\phi_\mu^\ell} \right) - \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_\mu^\ell} \right) \right\| + \left\| \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_\mu^\ell} \right) - \phi_\mu^\ell \left(\omega_{uv}^{(\ell)}(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_\mu^\ell} \right) \right\| \\ &\leq M^{(\ell)} \|\omega_{uv}^{(\ell)}(\mathcal{W})\| \mathrm{dist}(\mathcal{W}_{\phi_\mu^\ell}, \tilde{\mathcal{W}}_{\phi_\mu^\ell}) + M^{(\ell)} \|\omega_{uv}^{(\ell)}(\mathcal{W}) - \omega_{uv}^{(\ell)}(\tilde{\mathcal{W}})\|. \end{aligned}$$

We will now determine bounds for $\|\omega_{uv}^{(\ell)}(\mathcal{W})\|$ and $\|\omega_{uv}^{(\ell)}(\mathcal{W}) - \omega_{uv}^{(\ell)}(\tilde{\mathcal{W}})\|$. Starting with the first:

$$\begin{aligned} \|\omega_{uv}^{(\ell)}(\mathcal{W})\| &\leq 2 \left(\|h_u^{(\ell)}(\mathcal{W})\| + \|h_v^{(\ell)}(\mathcal{W})\| + \|z_u^{(\ell)}(\mathcal{W}) - z_v^{(\ell)}(\mathcal{W})\| + \|a_{uv}\| \right) \\ &\leq 8C\beta^{(\ell)} = 160D \left(M^{(\ell)} \right)^2 C\beta^{(\ell-1)} \end{aligned} \tag{25}$$

and the second:

$$
\begin{aligned}
&\|\omega_{uv}^{(\ell)}(\mathcal{W}) - \omega_{uv}^{(\ell)}(\tilde{\mathcal{W}})\| \\
&\leq 2\left(\|h_u^{(\ell)}(\mathcal{W}) - h_u^{(\ell)}(\tilde{\mathcal{W}})\| + \|h_v^{(\ell)}(\mathcal{W}) - h_v^{(\ell)}(\tilde{\mathcal{W}})\|\right) \\
&+ 2\left(\left|\|z_u^{(\ell)}(\mathcal{W}) - z_v^{(\ell)}(\mathcal{W})\| - \|z_u^{(\ell)}(\tilde{\mathcal{W}}) - z_v^{(\ell)}(\tilde{\mathcal{W}})\|\right|\right) \\
&\leq 2\left(\|h_u^{(\ell)}(\mathcal{W}) - h_u^{(\ell)}(\tilde{\mathcal{W}})\| + \|h_v^{(\ell)}(\mathcal{W}) - h_v^{(\ell)}(\tilde{\mathcal{W}})\| + \|z_u^{(\ell)}(\mathcal{W}) - z_u^{(\ell)}(\tilde{\mathcal{W}})\| + \|z_v^{(\ell)}(\mathcal{W}) - z_v^{(\ell)}(\tilde{\mathcal{W}})\|\right) \\
&\leq 4\sqrt{2}M^{(\ell)}(2C\beta^{(\ell-1)} + K_h^{(\ell-1)} + K_\mu^{(\ell-1)})\mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}) \\
&+ 4M^{(\ell)}\left(C\beta^{(\ell-1)} + K_\mu^{(\ell-1)} + K_z^{(\ell-1)} + \frac{6K_z^{(\ell-1)}}{\varepsilon}C\beta^{(\ell-1)}\right)\mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}) \\
&\leq 4M^{(\ell)}\left(4C\beta^{(\ell-1)} + 2K_h^{(\ell-1)} + 3K_\mu^{(\ell-1)} + K_z^{(\ell-1)} + \frac{6K_z^{(\ell-1)}}{\varepsilon}C\beta^{(\ell-1)}\right)\mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}})
\end{aligned}
$$

and therefore:

$$
\begin{aligned}
\left\|\mu_{u\to v}^{(\ell)}(\mathcal{W}) - \mu_{u\to v}^{(\ell)}(\tilde{\mathcal{W}})\right\| &\leq 160D\left(M^{(\ell)}\right)^3 C\beta^{(\ell-1)}\mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}) \\
+4\left(M^{(\ell)}\right)^2&\left(4C\beta^{(\ell-1)} + 2K_h^{(\ell-1)} + 3K_\mu^{(\ell-1)} + K_z^{(\ell-1)} + \frac{6K_z^{(\ell-1)}}{\varepsilon}C\beta^{(\ell-1)}\right)\mathrm{dist}_\ell(\mathcal{W}, \tilde{\mathcal{W}}).
\end{aligned} \tag{26}
$$

In summary, we have the following recursive relationships:

$$
\begin{aligned}
K_h^{(\ell)} &= \sqrt{2}M^{(\ell)}(2C\beta^{(\ell-1)} + K_h^{(\ell-1)} + K_\mu^{(\ell-1)}) \\
K_z^{(\ell)} &= M^{(\ell)}\left(C\beta^{(\ell-1)} + K_\mu^{(\ell-1)} + K_z^{(\ell-1)} + \frac{6K_z^{(\ell-1)}}{\varepsilon}C\beta^{(\ell-1)}\right) \\
\frac{1}{D}K_\mu^{(\ell)} &= 160D\left(M^{(\ell)}\right)^3 C\beta^{(\ell-1)} \\
&\quad + 4\left(M^{(\ell)}\right)^2\left(4C\beta^{(\ell-1)} + 2K_h^{(\ell-1)} + 3K_\mu^{(\ell-1)} + K_z^{(\ell-1)} + \frac{6K_z^{(\ell-1)}}{\varepsilon}C\beta^{(\ell-1)}\right)
\end{aligned} \tag{27}
$$

We now determine the growth rates of the constants. Specifically, we will now show by induction that

$$
\begin{aligned}
K_h^{(\ell)} &\leq CQ^\ell B^{(\ell)} \\
K_\mu^{(\ell)} &\leq CQ^\ell B^{(\ell)} \\
K_z^{(\ell)} &\leq CQ^\ell B^{(\ell)},
\end{aligned} \tag{28}
$$

where $Q = 224D^2\max\{\frac{C}{\varepsilon}, 1\}$ and $B^{(\ell)} = \left(\prod_{i=0}^{\ell} M^{(i)}\right)^3 \prod_{i=0}^{\ell-1}\beta^{(i)}$ with the convention that empty product is equal to 1. To show the base case, we recall that $K_h^{(0)} = K_z^{(0)} = 0$ and $K_\mu^{(0)} = CM^{(0)}$. Now, assuming that (28) holds for all $\ell' < \ell$,

we will show that it also holds for ℓ. Using the fact that $M^{(\ell)} \geq 1$, $\beta^{(\ell)} \geq 1$ and $B^{(\ell)} \geq 1$ for all ℓ, we obtain:

$$
\begin{aligned}
K_h^{(\ell)} &\leq \sqrt{2}M^{(l)}\left(2C\beta^{(\ell-1)} + 2CQ^{\ell-1}B^{(\ell-1)}\right) \leq 4\sqrt{2}CQ^{\ell-1}\left(M^{(\ell)}\right)^3\beta^{(\ell-1)}B^{(\ell-1)} \\
&\leq 4\sqrt{2}\max\left\{\frac{C}{\varepsilon}, 1\right\}CQ^{\ell-1}B^{(\ell)} \leq CQ^{\ell}B^{(\ell)} \\
K_z^{(\ell)} &\leq M^{(\ell)}\left(C\beta^{(\ell-1)} + 2CQ^{\ell-1}B^{(\ell-1)} + \frac{6}{\varepsilon}CQ^{\ell-1}B^{(\ell-1)}C\beta^{(\ell-1)}\right) \\
&\leq \left(1 + 2 + \frac{6C}{\varepsilon}\right)CQ^{\ell-1}\left(M^{(\ell)}\right)^3\beta^{(\ell-1)}B^{(\ell-1)} \leq 9\max\left\{\frac{C}{\varepsilon}, 1\right\}CQ^{\ell-1}B^{(\ell)} \leq CQ^{\ell}B^{(\ell)} \\
K_\mu^{(\ell)} &\leq 160D^2\left(M^{(\ell)}\right)^3C\beta^{(\ell-1)} + 4D\left(M^{(\ell)}\right)^2\left(4C\beta^{(\ell-1)} + 6CQ^{\ell-1}B^{(\ell-1)} + \frac{6}{\varepsilon}C\beta^{(\ell-1)}CQ^{\ell-1}B^{(\ell-1)}\right) \\
&\leq \left(160D^2 + 16D + 24D + \frac{24DC}{\varepsilon}\right)CQ^{\ell-1}\left(M^{(\ell)}\right)^3\beta^{(\ell-1)}B^{(\ell-1)} \\
&\leq \left(160D^2 + 64D\right)\max\left\{\frac{C}{\varepsilon}, 1\right\}CQ^{\ell-1}B^{(\ell)} \leq CQ^{\ell}B^{(\ell)}
\end{aligned}
\tag{29}
$$

□

F. Proof of Lemma 4.4

In this section, we prove Lemma 4.4. We recall it for completeness:

Lemma 4.4 (Lipschitz continuity w.r.t. parameters of the scoring model)**.** *Let $\mathcal{W}^{(g)} = (\mathcal{W}, \mathcal{W}_{\phi_{out}})$ denote the parameters of the scoring model g as defined in Equation (6), with $\mathcal{W}$ as defined in Lemma 4.3, $\mathcal{W}_{\phi_{out}} = \{W_i^{\phi_{out}}\}_{i=1}^{L_{out}}$ and $g(G; \mathcal{W}^{(g)})$ the output of the scoring model with parameters $\mathcal{W}^{(g)}$. Then, g is Lipschitz continuous w.r.t. the model parameters:*

$$\|g(G; \mathcal{W}^{(g)}) - g(G; \tilde{\mathcal{W}}^{(g)})\| \leq K_g \mathrm{dist}(\mathcal{W}^{(g)}, \tilde{\mathcal{W}}^{(g)}),$$

where

$$\mathrm{dist}(\mathcal{W}^{(g)}, \tilde{\mathcal{W}}^{(g)}) = \mathrm{dist}_{L_{egnn}}(\mathcal{W}, \tilde{\mathcal{W}}) + \sum_{i=1}^{L_{out}} \|W_i^{\phi_{out}} - \tilde{W}_i^{\phi_{out}}\|_2$$

C, Q, $B^{(\cdot)}$ are as defined in Lemma 4.3 and

$$K_g = 2\left(\prod_{i=1}^{L_{out}} K_\psi \beta_{i,\phi_{out}}\right) CQ^{L_{egnn}} B^{(L_{egnn})}.$$

Proof. In the proof we assume $\|\cdot\| \equiv \|\cdot\|_2$. From Equation 5: $h_G = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} h_v^{(L_{\text{egnn}})}$ and therefore h_G retain Lipschitz continuity of each individual $h_v^{(L_{\text{egnn}})}$ with the same constant. It holds that:

$$
\begin{aligned}
&\|g(G; \mathcal{W}^{(g)}) - g(G; \tilde{\mathcal{W}}^{(g)})\| = \|\phi_{\text{out}}(h_G(\mathcal{W}); \mathcal{W}_{\phi_{\text{out}}}) - \phi_{\text{out}}(h_G(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_{\text{out}}})\| \\
&\leq \|\phi_{\text{out}}(h_G(\mathcal{W}); \mathcal{W}_{\phi_{\text{out}}}) - \phi_{\text{out}}(h_G(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_{\text{out}}})\| + \|\phi_{\text{out}}(h_G(\mathcal{W}); \tilde{\mathcal{W}}_{\phi_{\text{out}}}) - \phi_{\text{out}}(h_G(\tilde{\mathcal{W}}); \tilde{\mathcal{W}}_{\phi_{\text{out}}})\| \\
&\leq \left(\prod_{i=1}^{L_{\text{out}}} K_\psi \beta_{i,\phi_{\text{out}}}\right)\|h_G(\mathcal{W})\|\mathrm{dist}(\mathcal{W}_{\phi_{\text{out}}}, \tilde{\mathcal{W}}_{\phi_{\text{out}}}) + \left(\prod_{i=1}^{L_{\text{out}}} K_\psi \beta_{i,\phi_{\text{out}}}\right)\|h_G(\mathcal{W}) - h_G(\tilde{\mathcal{W}})\| \\
&\leq \left(\prod_{i=1}^{L_{\text{out}}} K_\psi \beta_{i,\phi_{\text{out}}}\right) C(\beta^{(L_{\text{egnn}})} + Q^{L_{\text{egnn}}} B^{(L_{\text{egnn}})})\mathrm{dist}(\mathcal{W}^{(g)}, \tilde{\mathcal{W}}^{(g)}) \\
&\leq 2\left(\prod_{i=1}^{L_{\text{out}}} K_\psi \beta_{i,\phi_{\text{out}}}\right) CQ^{L_{\text{egnn}}} B^{(L_{\text{egnn}})}\mathrm{dist}(\mathcal{W}^{(g)}, \tilde{\mathcal{W}}^{(g)}).
\end{aligned}
$$

□

G. Proof of Proposition 4.1

In this section, we prove our main result, i.e. Proposition 4.1. We begin with a useful lemma:

Lemma G.1 (Weight matrices' coverings yield a function class covering)**.** *Let $\mathcal{F} = \{f_{W_1,\ldots,W_n} : W_i \in \mathcal{W}_i\}$ be a class of functions parametrized with n weight matrices. Suppose further that for all $W_i, \tilde{W}_i$:*

$$\|f_{W_1,\ldots,W_n} - f_{\tilde{W}_1,\ldots,\tilde{W}_n}\| = \sup_{x\in\mathcal{X}} \|f_{W_1,\ldots,W_n}(x) - f_{\tilde{W}_1,\ldots,\tilde{W}_n}(x)\| \leq K\left(\|W_1 - \tilde{W}_1\|_F + \cdots + \leq \|W_n - \tilde{W}_n\|_F\right).$$

Then

$$\mathcal{N}\left(\mathcal{F}, \varepsilon, \|\cdot\|_\infty\right) \leq \prod_{i=1}^{n} \mathcal{N}\left(\mathcal{W}_i, \frac{\varepsilon}{nK}, \|\cdot\|_F\right).$$

Proof. Let $\mathcal{C}_i$ be an $\left(\frac{\varepsilon}{nK}\right)$-covering of $\mathcal{W}_i$ such that $|\mathcal{C}_i| = \mathcal{N}\left(\mathcal{W}_i, \frac{\varepsilon}{nK}, \|\cdot\|_F\right)$ and define

$$\mathcal{C} = \{f_{W_1,\ldots,W_n} : W_i \in \mathcal{C}_i\}.$$

Let $f_{W_1,\ldots,W_n} \in \mathcal{F}$. For all i we can choose $\hat{W}_i \in \mathcal{C}_i$ such that $\|W_i - \hat{W}_i\|_F \leq \frac{\varepsilon}{nK}$. Therefore

$$\|f_{W_1,\ldots,W_n} - f_{\hat{W}_1,\ldots,\hat{W}_n}\|_\infty \leq K\left(\|W_1 - \hat{W}_1\|_F + \cdots + \|W_n - \hat{W}_n\|_F\right) \leq \varepsilon$$

and $f_{\hat{W}_1,\ldots,\hat{W}_n} \in \mathcal{C}$. Hence, $\mathcal{C}$ is an ε-covering of $\mathcal{F}$ w.r.t. $\|\cdot\|_\infty$ and

$$\mathcal{N}\left(\mathcal{F}, \varepsilon, \|\cdot\|_\infty\right) \leq |\mathcal{C}| = \prod_{i=1}^{n} |\mathcal{C}_i| = \prod_{i=1}^{n} \mathcal{N}\left(\mathcal{W}_i, \frac{\varepsilon}{nK}, \|\cdot\|_F\right).$$

□

We now move on to proving Proposition 4.1:

Proposition 4.1 (Generalization bound of the scoring model)**.** *Let $\mathcal{G}$ be the space of geometric graphs as defined in Section 3.1, $\mathcal{Y} = [0,1]$ the space of labels, $g : \mathcal{G} \to \mathbb{R}$ the scoring model as defined in (6) and $\mathcal{L}(\widehat{y}, y) = \min\left\{(\widehat{y} - y)^2, 1\right\}$ the loss function. Then for any $\delta > 0$, with probability at least $1 - \delta$ over choosing a sample $S \sim \mathcal{D}^m$ from a distribution $\mathcal{D}$ over $\mathcal{G} \times \mathcal{Y}$ of size m, the following holds:*

$$\mathcal{R}_{S,\mathcal{L}}(g) = \mathcal{O}\left(\frac{d\sqrt{L}}{\sqrt{m}}\sqrt{\Delta} + \frac{\sqrt{\log\frac{2}{\delta}}}{\sqrt{m}}\right), \tag{16}$$

where $L = 3L_\phi L_{egnn} + L_{out}$ is the total number of weight matrices, d is the maximum width across all layers, and

$$\Delta = \sum_{\ell=0}^{L_{egnn}} \sum_{\phi\in\{\phi_h^{\ell-1},\phi_z^{\ell-1},\phi_\mu^{\ell}\}} \sum_{i=1}^{L_\phi} \gamma_{i,\phi} + \sum_{i=1}^{L_{out}} \gamma_{i,\phi_{out}}$$

$$\gamma_{i,\phi} = w_{i,\phi}\log(\Gamma\cdot\kappa(W_i^\phi))$$

$$w_{i,\phi} = \begin{cases} 1 & \text{for } \phi = \phi_{out} \\ L_{egnn} - \ell + 1 & \text{for } \phi \in \{\phi_h^\ell, \phi_z^\ell, \phi_\mu^\ell\} \end{cases}$$

$$\Gamma = \frac{dmLD\beta K_\psi}{\hat{\varepsilon}}.$$

Proof. From Theorem 3.1 we can see that in order to bound $\mathcal{R}_{S,\mathcal{L}}$ it suffices to find a bound for $\widehat{\mathfrak{R}}_S(\mathcal{F}_\mathcal{L})$, where

$$\mathcal{F}_\mathcal{L} = \{(x,y) \mapsto \mathcal{L}(f(x),y) : f \in \mathcal{F}\}$$

and

$$\mathcal{F} = \{g(\cdot, \mathcal{W}^{(g)}) : \forall \ell \, \forall \phi \in \{\phi_h^\ell, \phi_z^\ell, \phi_\mu^\ell, \phi_{\text{out}}\} \, \forall i \, \|W_i^\phi\| \leq \beta_{i,\phi}\}$$

is the family of scoring models with bounded spectral norm of weight matrices. The primary tool that we will use for bounding $\widehat{\mathfrak{R}}_S(\mathcal{F}_\mathcal{L})$ is Lemma 3.1. However, to use it directly, the assumption $f_0 \in \mathcal{F}_\mathcal{L}$ would need to be satisfied, which is very restrictive as it assumes the existence of a perfect classifier in the model class. Instead, we observe that

$$\mathcal{F}_\mathcal{L} = \{(x, y) \mapsto 1 - \hat{f}(x, y) : \hat{f} \in \hat{\mathcal{F}}_\mathcal{L}\}, \tag{30}$$

where

$$\hat{\mathcal{F}}_\mathcal{L} = \{(x, y) \mapsto 1 - \mathcal{L}(f(x), y) : f \in \mathcal{F}\}.$$

It may seem tautological, but now there exists $\hat{f}_0 \in \hat{\mathcal{F}}_\mathcal{L}$, because we can take $f_0 \equiv -1$ and since $y \in [0, 1]$, it holds that $\forall (x, y) \; \hat{f}_0(x, y) \coloneqq 1 - \mathcal{L}(f_0(x), y) = 0$. We can thus use Lemma 3.1 for $\hat{\mathcal{F}}_\mathcal{L}$ and obtain

$$\widehat{\mathfrak{R}}_S(\hat{\mathcal{F}}_\mathcal{L}) \leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m} \int_\alpha^{2\sqrt{m}} \sqrt{\log \mathcal{N}(\hat{\mathcal{F}}_\mathcal{L}, r, \|\cdot\|_\infty)} \mathrm{d}r \right).$$

Now since $(x, y) \mapsto 1 - \mathcal{L}(x, y)$ is 2-Lipschitz in its first argument, an $r/2$ covering of $\mathcal{F}$ yields an r covering of $\hat{\mathcal{F}}_\mathcal{L}$ and therefore

$$\log \mathcal{N}(\hat{\mathcal{F}}_\mathcal{L}, r, \|\cdot\|) \leq \log \mathcal{N}(\mathcal{F}, \frac{r}{2}, \|\cdot\|_\infty).$$

Furthermore, from Equation (30), one can easily check that $\widehat{\mathfrak{R}}_S(\mathcal{F}_\mathcal{L}) = \widehat{\mathfrak{R}}_S(\hat{\mathcal{F}}_\mathcal{L})$ and in summary it holds that

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_\mathcal{L}) \leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m} \int_\alpha^{2\sqrt{m}} \sqrt{\log \mathcal{N}(\mathcal{F}, \frac{r}{2}, \|\cdot\|_\infty)} \mathrm{d}r \right).$$

However, since $\mathcal{N}(\mathcal{F}, r, \|\cdot\|_\infty)$ is a decreasing function of r we see that

$$\begin{aligned}\int_\alpha^{2\sqrt{m}} \sqrt{\log \mathcal{N}\left(\mathcal{F}, \frac{r}{2}, \|\cdot\|_\infty\right)} \mathrm{d}r &\leq \int_\alpha^{2\sqrt{m}} \sqrt{\log \mathcal{N}\left(\mathcal{F}, \frac{\alpha}{2}, \|\cdot\|_\infty\right)} \mathrm{d}r \leq \int_0^{2\sqrt{m}} \sqrt{\log \mathcal{N}\left(\mathcal{F}, \frac{\alpha}{2}, \|\cdot\|_\infty\right)} \mathrm{d}r \\ &= 2\sqrt{m} \sqrt{\log \mathcal{N}\left(\mathcal{F}, \frac{\alpha}{2}, \|\cdot\|_\infty\right)}.\end{aligned}$$

By choosing $\alpha = \frac{1}{\sqrt{m}}$, we obtain a simplified bound

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_\mathcal{L}) \leq \frac{4}{m} + \frac{24}{\sqrt{m}} \sqrt{\log \mathcal{N}\left(\mathcal{F}, \frac{1}{2\sqrt{m}}, \|\cdot\|_\infty\right)}. \tag{31}$$

We now move on to finding a bound for $\log \mathcal{N}\left(\mathcal{F}, \frac{1}{2\sqrt{m}}, \|\cdot\|_\infty\right)$. From Lemma 4.4:

$$\begin{aligned}\|g(\cdot; \mathcal{W}^{(g)}) - g(\cdot; \tilde{\mathcal{W}}^{(g)})\|_\infty &= \sup_G \|g(G; \mathcal{W}^{(g)}) - g(G; \tilde{\mathcal{W}}^{(g)})\|_2 \\ &\leq K_g \mathrm{dist}(\mathcal{W}^{(g)}, \tilde{\mathcal{W}}^{(g)}) \\ &= K_g \left(\sum_{\ell=0}^{L_{\text{egnn}}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell\}} \sum_{i=1}^{L_\phi} \|W_i^\phi - \tilde{W}_i^\phi\|_2 + \sum_{i=1}^{L_{\text{out}}} \|W_i^{\phi_{\text{out}}} - \tilde{W}_i^{\phi_{\text{out}}}\|_2 \right) \\ &\leq K_g \left(\sum_{\ell=0}^{L_{\text{egnn}}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell\}} \sum_{i=1}^{L_\phi} \|W_i^\phi - \tilde{W}_i^\phi\|_F + \sum_{i=1}^{L_{\text{out}}} \|W_i^{\phi_{\text{out}}} - \tilde{W}_i^{\phi_{\text{out}}}\|_F \right),\end{aligned}$$

where $K_g = 2\left(\prod_{i=1}^{L_{\text{out}}} K_\psi \beta_{i,\phi_{\text{out}}}\right) CQ^{L_{\text{egnn}}} B^{(L_{\text{egnn}})}$ and the last inequality comes from the relationship between the spectral and Frobenius norms. Thus, for $L = 3L_{\text{egnn}}L_\phi + L_{\text{out}}$ (total number of weight matrices), we can use Lemma G.1, i.e. that it suffices to find an $\left(\frac{r}{LK_g}\right)$-covering for each weight matrix W_i^ϕ and their cartesian product yields an r-covering of $\mathcal{F}$:

$$\log \mathcal{N}\left(\mathcal{F}, r, \|\cdot\|_\infty\right) \leq \sum_{\ell=0}^{L_{\text{egnn}}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell\}} \sum_{i=1}^{L_\phi} \log \mathcal{N}\left(W_i^\phi, \frac{r}{LK_g}, \|\cdot\|_F\right) + \sum_{i=1}^{L_{\text{out}}} \log \mathcal{N}\left(W_i^{\phi_{\text{out}}}, \frac{r}{LK_g}, \|\cdot\|_F\right).$$

Now, since the spectral norm of weight matrices is bounded, we can use Lemma 3.2 and obtain

$$\log \mathcal{N}\left(W_i^\phi, \frac{r}{LK_g}, \|\cdot\|_F\right) \leq d^2 \log\left(1 + 2\frac{\sqrt{d}LK_g\beta_{i,\phi}}{r}\right),$$

where $d = \max_{\ell=0}^{L_{\text{egnn}}} \max_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell, \phi_{\text{out}}\}} \max_{i=1}^{L_\phi} \dim(W_i^\phi)$ is the *width* of the scoring model and $\dim(W) = \max(d_1, d_2)$ for $W \in \mathbb{R}^{d_1 \times d_2}$. Choosing $r = \frac{1}{2\sqrt{m}}$ yields

$$\begin{aligned}
\log \mathcal{N}\left(\mathcal{F}, \frac{1}{2\sqrt{m}}, \|\cdot\|_\infty\right) &\leq d^2 \sum_{\ell=0}^{L_{\text{egnn}}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell\}} \sum_{i=1}^{L_\phi} \log\left(1 + 4\sqrt{dm}LK_g\beta_{i,\phi}\right) \\
&\quad + d^2 \sum_{i=1}^{L_{\text{out}}} \log\left(1 + 4\sqrt{dm}LK_g\beta_{i,\phi_{\text{out}}}\right) \\
&\leq d^2 \left(\sum_{\ell=0}^{L_{\text{egnn}}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell\}} \sum_{i=1}^{L_\phi} \log\left(5\sqrt{dm}LK_g\beta_{i,\phi}\right) + \sum_{i=1}^{L_{\text{out}}} \log\left(5\sqrt{dm}LK_g\beta_{i,\phi_{\text{out}}}\right) \right)
\end{aligned} \tag{32}$$

and plugging into (31) gives

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_\mathcal{L}) \leq \frac{4}{m} + \frac{24d}{\sqrt{m}} \sqrt{\underbrace{\sum_{\ell=0}^{L_{\text{egnn}}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell\}} \sum_{i=1}^{L_\phi} \log\left(5\sqrt{dm}LK_g\beta_{i,\phi}\right) + \sum_{i=1}^{L_{\text{out}}} \log\left(5\sqrt{dm}LK_g\beta_{i,\phi_{\text{out}}}\right)}_{=:\Sigma}}. \tag{33}$$

We now proceed to simplify the bound. Note that

$$\Sigma = L\log(5\sqrt{dm}L) + L\log(K_g) + \sum_{\ell=0}^{L_{\text{egnn}}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^\ell\}} \sum_{i=1}^{L_\phi} \log\left(\beta_{i,\phi}\right) + \sum_{i=1}^{L_{\text{out}}} \log\left(\beta_{i,\phi_{\text{out}}}\right)$$

and

$$\begin{aligned}
\log(K_g) &= \log(2) + \sum_{i=1}^{L_{\text{out}}} \log(K_\psi \beta_{i,\phi_{\text{out}}}) + \log(C) + L_{\text{egnn}}\log(Q) + \log(B^{(L_{\text{egnn}})}) \\
&= \log(16) + \log(D\beta) + \sum_{i=1}^{L_{\text{out}}} \log(K_\psi \beta_{i,\phi_{\text{out}}}) \\
&\quad + L_{\text{egnn}} \log\left(224D^2 \max\left\{\frac{8D\beta}{\varepsilon}, 1\right\}\right) \\
&\quad + 3\sum_{\ell=0}^{L_{\text{egnn}}} \log(M^{(\ell)}) + \sum_{\ell=0}^{L_{\text{egnn}}-1} \log\left((20D)^\ell \left(\prod_{i=0}^{\ell} M^{(i)}\right)^2\right) \\
&\leq C_1 \log\left(\frac{D\beta}{\hat{\varepsilon}}\right) \frac{L_{\text{egnn}}(L_{\text{egnn}}-1)}{2} + \sum_{i=1}^{L_{\text{out}}} \log(K_\psi \beta_{i,\phi_{\text{out}}}) + 3\sum_{\ell=0}^{L_{\text{egnn}}} \sum_{i=0}^{\ell} \log(M^{(i)}),
\end{aligned}$$

where C_1 does not depend on the parameters of the model and $\hat{\varepsilon} = \min\{1, \varepsilon\}$.
Since $M^{(\ell)} = \max_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^{\ell}\}} \prod_{i=1}^{L_\phi} K_\psi \beta_{i,\phi}$, we have

$$\log(M^{(i)}) \leq \sum_{\phi \in \{\phi_h^{i-1}, \phi_z^{i-1}, \phi_\mu^{i}\}} \sum_{j=1}^{L_\phi} \log(K_\psi \beta_{j,\phi})$$

and

$$\begin{aligned}\log(K_g) \leq C_2 \log\left(\frac{D\beta}{\hat{\varepsilon}}\right) \frac{L_{\text{egnn}}(L_{\text{egnn}}-1)}{2} + \sum_{i=1}^{L_{\text{out}}} \log(K_\psi \beta_{i,\phi_{\text{out}}}) \\ + 3 \sum_{\ell=0}^{L_{\text{egnn}}} (L_{\text{egnn}} - \ell + 1) \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^{\ell}\}} \sum_{i=1}^{L_\phi} \log(K_\psi \beta_{i,\phi}).\end{aligned}$$

In summary:

$$\Sigma \leq C_3 L \left(\sum_{\ell=0}^{L_{\text{egnn}}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^{\ell}\}} \sum_{i=1}^{L_\phi} w_{i,\phi} \log\left(\Gamma \beta_{i,\phi}\right) + \sum_{i=1}^{L_{\text{out}}} w_{i,\phi_{\text{out}}} \log\left(\Gamma \beta_{i,\phi_{\text{out}}}\right) \right),$$

where

$$\begin{aligned} w_{i,\phi} &= \begin{cases} 1 & \text{for } \phi = \phi_{\text{out}} \\ L_{\text{egnn}} - \ell + 1 & \text{for } \phi \in \{\phi_h^\ell, \phi_z^\ell, \phi_\mu^\ell\} \end{cases} \\ \Gamma &= \frac{dmLD\beta K_\psi}{\hat{\varepsilon}}. \end{aligned}$$

We can now use Equation 33 and Theorem 3.1 to obtain the bound for the generalization error of the scoring model g with probability at least $1 - \delta$:

$$\mathcal{R}_{S,\mathcal{L}}(g) \leq \frac{8}{m} + \frac{48d}{\sqrt{m}} \sqrt{\Sigma} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

Furthermore, for a given $g(\cdot; \mathcal{W})$, we observe that the *stretching factor* as defined in Equation 15 can serve as the bound, i.e. we can set $\beta_{i,\phi} = \kappa(W_i^\phi)$ and get

$$\mathcal{R}_{S,\mathcal{L}}(g) \leq \frac{8}{m} + C_3 \frac{48d\sqrt{L}}{\sqrt{m}} \sqrt{\sum_{\ell=0}^{L_{\text{egnn}}} \sum_{\phi \in \{\phi_h^{\ell-1}, \phi_z^{\ell-1}, \phi_\mu^{\ell}\}} \sum_{i=1}^{L_\phi} \gamma_{i,\phi} + \sum_{i=1}^{L_{\text{out}}} \gamma_{i,\phi_{\text{out}}}} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

where

$$\begin{aligned} L &= 3L_\phi L_{\text{egnn}} + L_{\text{out}} \\ \gamma_{i,\phi} &= w_{i,\phi} \log(\Gamma \kappa(W_i^\phi)) \\ \Gamma &= \frac{dmLD\beta K_\psi}{\hat{\varepsilon}} \\ w_{i,\phi} &= \begin{cases} 1 & \text{for } \phi = \phi_{\text{out}} \\ L_{\text{egnn}} - \ell + 1 & \text{for } \phi \in \{\phi_h^\ell, \phi_z^\ell, \phi_\mu^\ell\} \end{cases} \\ \kappa(W_i^\phi) &= \max\{1, \|W_i^\phi\|_2\} \\ d &= \max_{i,\phi} \dim(W_i^\phi) \\ \dim(W) &= \max(d_1, d_2) \text{ for } W \in \mathbb{R}^{d_1 \times d_2} \end{aligned}$$

□

H. Logarithmic dependence on spectral norms

In section 5 we discuss how our bound differs from existing ones, specifically the logarithmic dependence on the norm of weight matrices. Here we provide more details on where this difference stems from. The logarithmic dependence comes from Lemma 3.2 (Lemma 8 in (Chen et al., 2020)), which states:

$$\log \mathcal{N}\left(\left\{A \in \mathbb{R}^{d_1 \times d_2} : \|A\|_2 \leq \lambda\right\}, \varepsilon, \|\cdot\|_F\right) \leq d_1 d_2 \log\left(1 + 2\frac{\min\{d_1, d_2\}\lambda}{\varepsilon}\right).$$

This is in contrast with perhaps a more commonly used result (Lemma 10 in (Chen et al., 2020), special case of Lemma 3.2 in (Bartlett et al., 2017)), which states:

$$\log \mathcal{N}\left(\left\{A \in \mathbb{R}^{d_1 \times d_2} : \|A\|_{2,1} \leq \lambda\right\}, \varepsilon, \|\cdot\|_2\right) \leq \frac{\lambda^2}{\varepsilon^2} \log(2d_1 d_2).$$

The former yields a logarithmic dependence on the norm, while the latter produces a multiplicative one. However, these are not directly comparable. Firstly, there is a tradeoff between the dependence on the norm and the width of the network $d = \max\{d_1, d_2\}$. The log of the covering norm in the former yields logarithmic dependence on the norm but quadratic on the width, and it is the other way around in the latter. In particular, for large values of λ and low values of d, the former may be preferred, whereas the latter would be better for low values of λ and high values of d.

Furthermore, these two lemmas provide bounds on the covering numbers of sets of matrices. Still, the former one assumes that the spectral norm is bounded and gives a bound w.r.t. Frobenius norm, while the latter assumes that $\|\cdot\|_{2,1}$ norm is bounded and gives a bound w.r.t. spectral norm, which makes the comparison difficult.

I. Impact of ε-normalization

In section 5 we argue that ε-normalization plays an important role in the derivation of the generalization bound. Here, we provide more details to support this claim. Specifically, suppose we follow the same reasoning as we did in Lemmas 4.1-4.4, for the unnormalized EGNN model, i.e. $\gamma(z_v^{(\ell)}, z_u^{(\ell)}, \varepsilon) \equiv 1$ in Equation (2). Assume further that $\forall i, \phi$, it holds that $\|M_{i,\phi}\|_2 \leq \beta_\phi$.

Consider the bound of the ℓ-layer EGNN embeddings $\beta^{(\ell)}$, i.e. $\|h_v^{(l)}\|_2 \leq \beta^{(\ell)}$. From Lemma 4.1, we know that for $\beta^{(\ell)} = \mathcal{O}(C_1^\ell)$ for some $C_1 > 1$ in the $\varepsilon-$normalized EGNN model. However, from Equation (19), we can see that $\beta^{(\ell)}$ would satisfy the following recursive relationship:

$$\beta^{(\ell)} = 20D(M^{(\ell)}\beta^{(\ell-1)})^2,$$

which implies

$$\beta^{(\ell)} = \mathcal{O}(C_2^{2^\ell})$$

for some $C_2 > 1$. The super-exponential growth of $\beta^{(\ell)}$ leads to a super-exponential Lipshitz constant of the unnormalized EGNN and therefore an exponential dependence of the generalization gap on the number of EGNN layers.

J. Implementation details

The QM9 dataset (Ramakrishnan et al. (2014)) comprises small molecules with a maximum of 29 atoms in 3D space. Each atom is defined by a 3D position and a five-dimensional one-hot node embedding that represents the atom type, indicated as $(\mathrm{H}, \mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{F})$. The dataset has several different chemical properties (outputs), from which we arbitrarily chose 4 to assess the generalization of EGNNs. We report the results for the remaining properties in Appendix K.

We implement all models using the PyTorch (Paszke et al., 2017) and train them for 1000 epochs using the Adam optimizer and MSE loss function. We use batch size equal to 96 and cosine decay for the learning rate starting at 10^{-3} except for the Mu (μ) property, where the initial value was set to 5×10^{-4}. We run five independent trials with different seeds.

For the experiments in Figure 3, we use width $d = 64$ (for all layers) and $L_{\text{egnn}} = 5$ for the experiments regarding the spectral norm, $L_{\text{egnn}} = 3$ for the one regarding the width, and width $d = 16$ for assessing the generalization gap in terms

of the number of layers. We apply ε-normalization with $\varepsilon = 1$. All internal MLPs ($\phi_h, \phi_z, \phi_\mu, \phi_{out}$) have two layers with SiLU activation function. Following the original repo, we use the identity activation function at the last layer of ϕ_z. For the experiments on the impact ε-normalization, we use $\varepsilon = 1$ and $d = 128$ (width).

We note that in their original, Satorras et al. (2021) do not update the positional features z over the layers on the experiments using QM9. In our experiments, we consider full EGNN models.

For the experiments in Table 3, we use width $d = 16$, ε-normalization ($\varepsilon = 1$), and internal MLPs with two layers (i.e., $L_\phi = L_{\text{out}} = 2$) with SiLU activation functions. We consider 2K samples for training, and full validation and test sets. The statistics are obtained from three independent runs, with different seeds. We perform model selection for the regularization factor $\lambda \in \{1, 1\text{e-}3, 1\text{e-}5, 1\text{e-}7\}$ using the validation set.

K. Additional visualizations and experiments

Figure 5 and Figure 6 show the learning curves (for a single run) obtained using our spectral regularizer and SpecAvg, respectively. Here, we use $L_{\text{egnn}} = 5$. Notably, our regularizer produces generalization gaps that decrease with the regularization factor λ. In contrast, SpecAvg shows a hard-threshold behavior: for $\lambda \neq 1$, it has little effect; for $\lambda = 1$, it produces a very small generalization gap but high test error.

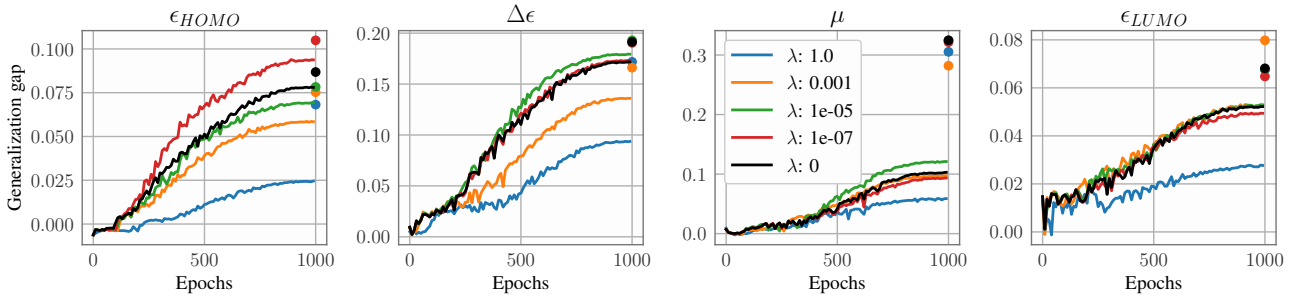

Figure 5. Generalization gap over training epochs for the regularization method proposed in this work. Circles at the end of training denote the final test MSE error.

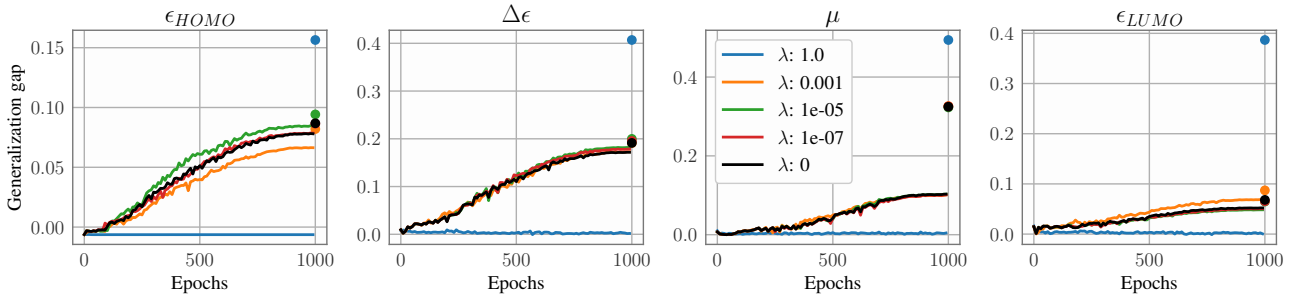

Figure 6. Generalization gap over training epochs for SpecAvg. Circles at the end of training denote the final test MSE error.

Time comparison. We note that regularization does not affect inference time as it only applies during training. The impact of the proposed method on training is reported in the Table 5. For the largest model (7 layers), the regularized version incurs a computational overhead of approximately 50% — i.e., the regularized vs. non-regularized time ratio is around 1.5.

Additional QM9 properties. Table 6 reports results (from three independent runs) for the remaining eight QM9 tasks. For these experiments, we used width $d = 32$, cosine decay with an initial learning rate of 10^{-3}, batch size of 96, and 2000 epochs. Overall, we observe that SpecAvg and our regularizer achieve similar Test MSE values, while our method obtains smaller generalization gaps.

Table 5. **Time per epoch (in seconds)**.

Task	L_{egnn}	None	Our Regularizer
alpha	3	1.37 ± 0.04	1.76 ± 0.29
	5	1.52 ± 0.04	2.08 ± 0.21
	7	1.64 ± 0.04	2.42 ± 0.06
Cv	3	1.38 ± 0.04	1.75 ± 0.32
	5	1.44 ± 0.04	2.08 ± 0.11
	7	1.58 ± 0.05	2.44 ± 0.08
G	3	1.37 ± 0.29	1.73 ± 0.05
	5	1.46 ± 0.05	2.07 ± 0.10
	7	1.57 ± 0.13	2.38 ± 0.10
H	3	1.32 ± 0.05	3.54 ± 0.09
	5	1.44 ± 0.04	2.07 ± 0.04
	7	1.59 ± 0.27	2.36 ± 0.05

Table 6. Test mean squared error (MSE) and generalization gap on additional QM9 tasks for different regularization methods. We denote the best-performing methods in bold. Results are multiplied by 100 for better visualization.

		Test MSE			**Generalization gap**		
Task	L_{egnn}	None	SpecAvg	Ours	None	SpecAvg	Ours
α	3	42.87 ± 0.99	42.72 ± 0.88	**41.76** ± 1.11	29.29 ± 17.10	29.05 ± 16.87	**16.46** ± 5.65
	5	41.47 ± 0.92	41.31 ± 0.40	**41.24** ± 1.30	34.79 ± 10.96	**34.34** ± 11.37	34.41 ± 11.19
	7	44.11 ± 1.97	43.10 ± 0.93	**41.91** ± 2.27	38.98 ± 5.76	28.67 ± 14.88	**18.78** ± 22.20
Cv	3	7.59 ± 1.38	7.10 ± 1.71	**7.09** ± 1.63	7.04 ± 1.64	**6.32** ± 2.36	6.49 ± 1.91
	5	12.11 ± 1.00	11.32 ± 1.03	**11.28** ± 1.21	12.09 ± 1.01	11.30 ± 1.02	**10.03** ± 1.57
	7	10.10 ± 2.28	10.24 ± 2.45	**9.08** ± 0.51	10.08 ± 2.28	10.23 ± 2.45	**9.04** ± 0.48
G	3	16.60 ± 7.35	15.71 ± 6.62	**16.00** ± 6.08	15.17 ± 6.91	13.91 ± 5.81	**10.74** ± 3.43
	5	21.88 ± 9.19	**20.63** ± 7.84	21.55 ± 5.24	21.57 ± 9.00	19.93 ± 7.18	**17.26** ± 4.95
	7	28.92 ± 1.89	26.89 ± 3.39	**26.11** ± 4.67	28.83 ± 1.94	26.29 ± 4.01	**22.84** ± 9.23
H	3	15.49 ± 8.10	**14.91** ± 6.92	15.19 ± 4.33	13.23 ± 7.48	**12.36** ± 5.74	12.97 ± 3.85
	5	22.60 ± 8.35	**22.39** ± 7.58	23.00 ± 4.53	22.20 ± 8.11	21.95 ± 7.28	**11.67** ± 5.77
	7	**25.75** ± 0.59	26.81 ± 2.47	26.91 ± 5.16	25.64 ± 0.60	26.67 ± 2.61	**14.67** ± 9.72
r2	3	56.63 ± 4.99	56.62 ± 4.99	56.65 ± 4.94	3.80 ± 2.29	3.81 ± 2.31	3.79 ± 2.31
	5	55.31 ± 1.07	55.22 ± 1.24	55.53 ± 1.17	5.03 ± 1.45	5.02 ± 1.45	4.95 ± 1.63
	7	58.28 ± 3.76	58.25 ± 3.78	**57.61** ± 2.50	6.86 ± 3.36	7.20 ± 3.02	**5.22** ± 0.54
U	3	15.80 ± 7.34	**15.46** ± 6.59	16.44 ± 3.67	13.59 ± 6.58	**12.82** ± 5.15	14.03 ± 2.68
	5	22.22 ± 6.94	**19.66** ± 7.46	23.34 ± 4.77	21.84 ± 6.68	19.39 ± 7.16	**17.75** ± 7.28
	7	27.43 ± 1.49	**25.63** ± 0.90	26.18 ± 2.99	27.31 ± 1.52	25.09 ± 1.77	**22.98** ± 8.18
U0	3	15.70 ± 7.18	**15.10** ± 5.99	15.87 ± 6.51	13.62 ± 6.32	**12.98** ± 5.08	13.80 ± 5.80
	5	22.03 ± 8.80	**21.66** ± 8.44	22.79 ± 4.57	21.74 ± 8.63	21.30 ± 8.22	**12.67** ± 5.19
	7	26.81 ± 3.12	21.71 ± 9.39	**21.33** ± 9.74	26.71 ± 3.20	21.64 ± 9.46	**21.13** ± 9.86
zpve	3	**94.14** ± 0.26	94.25 ± 0.29	94.21 ± 0.45	2.17 ± 0.81	2.13 ± 0.92	**1.81** ± 1.06
	5	95.29 ± 0.41	**94.82** ± 0.25	95.23 ± 0.29	7.18 ± 2.95	**1.67** ± 0.50	3.82 ± 2.67
	7	95.74 ± 0.32	**95.03** ± 0.29	95.14 ± 0.22	19.62 ± 13.81	4.95 ± 3.97	**2.03** ± 0.97