ON MULTI-SCALE GRAPH REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

1 INTRODUTION

Graph neural networks (GNNs) play a fundamental role in modern computational biology, where they often form the backbone for both subcellular tasks such as protein structure prediction (Jumper et al., 2021) as well as cell type annotation (Shao et al., 2021) and modeling of protein protein interactions (Liu et al., 2019) at the multi-cellular and organism wide levels. An underexplored drawback of common GNN methods, however, is that they are not inherently multiscale consistent: Two graphs describing the same object or situation at different resolution scales are assigned vastly different latent representations. This prevents graph networks from generating data representations that are consistent across scales. It also complicates the integration of representations at the molecular scale with those generated at the biological scale. Here we discuss why existing GNNs struggle with multiscale consistency and show how to overcome this problem by modifying the message passing paradigm within GNNs.

2 STANDARD GNNS ARE NOT MULTI-SCALE CONSISTENT

To illustrate that standard GNNs are unable to consistently integrate multiple scales, we utilize the QM7 dataset (Rupp et al., 2012). Here, molecular atomization energies of organic molecules (containing both hydrogen and heavy atoms) are to be predicted. Each molecule is represented by an adjacency matrix with entries $A_{ij} = Z_i Z_j |\vec{x}_i - \vec{x}_j|^{-1}$ given as Coulomb energies of atoms i, j.

From a physical perspective, describing a molecule at the level of interacting atoms corresponds to a specific choice of resolution scale, where interactions of individual protons and neutrons inside individual atoms are discarded. To test the multi-scale consistency of GNNs we additionally also consider a version of QM7 where we further lower the resolution scale: Here we aggregate each heavy atomic core together with its surrounding (single-proton) hydrogen atoms into super-nodes.

To showcase the failure of GNNs to consistently incorporate multiple scales, we confront models during inference with a version of QM7 on a scale different from the one they were trained on. As Table 1 details, mean-absolute-errors (MAEs) increase significantly when going from a same-resolution setting to a cross-resolution setting. None of the considered standard architectures (including multi-scale methods (SAG-M – PushNet)) consistently handles multiple scales. We can trace this back to the latent embeddings F and $\underline{F}$ that are being generated for original- $\{G\}$ and coarsified graphs $\{\underline{G}\}$: For models of Table 1 on average $10 \lesssim \|F - \underline{F}\| \lesssim 10^4$ (c.f. also Fig. 2): Latent representations of graphs describing the same object at different resolutions differ significantly.

To understand this behavior, we interpolate between fine and coarse resolution: Original graphs $\{G\}$ of QM7 are modified ($\{G_\omega\}$) by moving hydrogen atoms towards their corresponding heavy atom by a factor of $\omega \geqslant 1$ (i.e. $\text{dist}_{\text{new}} = \text{dist}_{\text{equilib.}}/\omega$). For $\omega \to \infty$, they arrive at the respective heavy atom ($\{\underline{G}\}$). In Fig. 2, we compare the latent distance between the coarse embeddings $\underline{F}$ and the embeddings for the intermediate graph F_ω. Embeddings F_ω do *not* converge to the coarse embeddings $\underline{F}$. Since the convergence of graph-sequence G_ω to the limit graph $\underline{G}$ is not turned into a convergence of latent embeddings $F_\omega \nrightarrow \underline{F}$ we conclude: **GNNs are not continuous**. This discontinuity explains why GNNs can map similar graphs (describing the same object at different resolutions) to different latent representations.

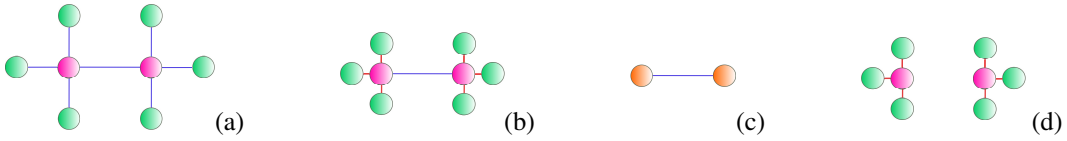

Figure 1: (a) G of QM7 (b) Modified G_ω (c) Coarsified $\underline{G}$ (d) Effective propagation in GCN

Table 1: Regression using high- and low-resolution QM7

Mean Absolute Error (↓) on QM7 [kcal/mol]				
Training	**High Resolution**		**Low Resolution**	
Inference	**Low Resolution**	High Resolution	Low Resolution	**High Resolution**
GCN	$125.34_{\pm 2.47}$	$63.17_{\pm 0.92}$	$67.75_{\pm 3.73}$	$380.51_{\pm 30.33}$
GATv2	$415.09_{\pm 96.57}$	$48.41_{\pm 19.20}$	$60.01_{\pm 3.34}$	$245.03_{\pm 90.97}$
ChebNet	$568.47_{\pm 37.70}$	$64.63_{\pm 1.21}$	$64.90_{\pm 4.55}$	$339.64_{\pm 101.30}$
SAG	$542.16_{\pm 27.33}$	$68.43_{\pm 1.93}$	$104.20_{\pm 3.92}$	$506.75_{\pm 60.57}$
BernNet	$765.22_{\pm 495.28}$	$83.76_{\pm 21.75}$	$90.52_{\pm 37.17}$	$594.62_{\pm 341.55}$
SAG-M	$285.53_{\pm 95.54}$	$66.22_{\pm 4.51}$	$73.57_{\pm 14.57}$	$307.67_{\pm 77.24}$
UFGNet	$620.21_{\pm 4.80}$	$13.71_{\pm 1.05}$	$24.53_{\pm 4.80}$	$156.44_{\pm 156.44}$
Lanczos	$939.87_{\pm 16.35}$	$10.55_{\pm 3.22}$	$83.11_{\pm 5.27}$	$654.61_{\pm 529.13}$
PushNet	$2442.59_{\pm 303.27}$	$60.94_{\pm 1.83}$	$69.25_{\pm 3.11}$	$124.08_{\pm 3.94}$
Resolvent	$16.54_{\pm 3.01}$	$16.53_{\pm 3.03}$	$15.79_{\pm 0.98}$	$13.80_{\pm 1.34}$
Exponential	$16.37_{\pm 1.71}$	$16.36_{\pm 2.16}$	$16.25_{\pm 1.41}$	$16.25_{\pm 1.41}$

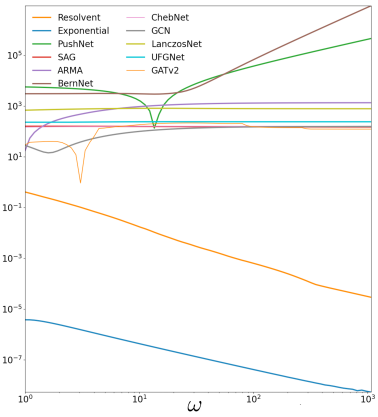

Figure 2: Latent distance $\|F_\omega - \underline{F}\|$

To understand this discontinuity, we exemplarily investigate (GCNs) (Kipf & Welling, 2017). There the layer-wise update acts as $X \mapsto \hat{A}XW$, with the feature matrix $X \in \mathbb{R}^{N\times F}$ (N nodes; latent dimension F), the weight matrix $W \in \mathbb{R}^{F\times F}$ and the *renormalized* adjacency matrix $\hat{A} \in \mathbb{R}^{N\times N}$. As hydrogen atoms move closer to the heavy atoms, the entries $\hat{A}_{\text{heavy,heavy}}$ in $\hat{A}_{ij} \sim A_{ij}/\sqrt{d_i d_j}$ tend to zero (as degrees d_{heavy} tend to infinity). Thus communication between heavy atoms becomes severely disrupted. Information only propagates along a increasingly disconnected effective graph (Fig.1(d)).

3 Global Laplace Propagation facilitates scale-consistency

To avoid a disconnected effective propagation graph as in Fig. 1 (d) we modify the message passing paradigm in GCN: To connect the information flows over G_ω and $\underline{G}$ we observe that features in G_ω should equalize faster between nodes connected by large edge weights. When such a large weight tends to infinity, features between strongly connected nodes are then equalized immediately, so that entire strongly connected clusters exactly behave as the single nodes in $\underline{G}$.

Noting that this is exactly the behavior that heat dissipating over a graph exhibits, we make use of the the heat diffusion equation $dX(t)/dt = -L \cdot X(t)$ (with Graph Laplacian L and time t) and the structure $X(t) = e^{-Lt} \cdot X(0)$ of its solutions, when designing our graph networks:

Definition 3.1. *Let $\hat{\psi}$ be a bounded (generalized) function defined on $[0, \infty)$. A **Global Laplacian Propagation Matrix** $\psi(L)$ is any matrix arising as $\psi(L) := \int_0^\infty e^{-tL}\hat{\psi}(t)dt$.*

Thus $\psi(L)$ represent a weighted sum of diffusion flows that have progressed to various times. Specifically, if we choose the Dirac distribution $\hat{\psi}_{\delta_{t_k}}(t) := \delta(t - t_k)$ as the weightinf function $\hat{\psi}_k$, we obtain **exponential** matrices $\psi_k(L) = \int_0^\infty \delta(t - t_k)e^{-tL}dt = e^{-t_k L}$ and $\hat{\psi}_k := (-t)^{k-1}e^{-\lambda t}$ to get powers of **resolvents** $\psi_k(L) = [(zId + L)^{-1}]^k$. The propagation matrix is then used instead of the adjacency matrix in each layer of the GNN leading to the update rule $X \mapsto \sum_k \psi_k(L)XW_k$.

As we prove in Appendix C.3, we indeed have $\|F_\omega - \underline{F}\| \to 0$ as $\omega \to \infty$, for such networks based on global Laplacian propagation matrices. This behaviour can clearly also be observed for the examples of exponential- and resolvent propagation matrices in Fig. 2. As ω increases, the distance between latent embeddings tends to zero. Thus these networks are indeed continuous.

In the previous Section 2, we had identified the discontinuity of standard GNNs as the obstruction to consistently incorporating multiple scales. This explained their sub-par performance in Table 1. Since networks based on global Laplacian propagation schemes are continuous, we thus expect a consistent incorporation of scales, as well as a good performance in cross resolution setting. This is exactly what we observe in Table 1: MAEs of GNNs based on global Laplacian propagation schemes (using either exponential or resolvent matrices) do not increase when going from a same- to a cross-resolution setting; MAEs of such methods are lower than those of standard graph learning methods by factors of order 10^1 to 10^2. Hence these methods indeed do consistently incorporate varying scales.

MEANINGFULNESS STATEMENT

A model generating meaningful representations of life should be able to consistently represent all aspects of life across all relevant scales, starting from the molecular level all the way up to the biological level. While graph neural networks have emerged as a popular network architecture for biological problems at any individual scale, we show that they are not directly suitable to facilitate connections *between* respective scales. To remedy this and facilitate progress in eventually continuously traversing between the molecular and the biological scale, we propose a new propagation scheme that allows graph neural networks to indeed incorporate multiple scales.

REFERENCES

Wolfgang Arendt. APPROXIMATION OF DEGENERATE SEMIGROUPS. *Taiwanese Journal of Mathematics*, 5(2):279 – 295, 2001. doi: 10.11650/twjm/1500407337. URL https://doi.org/10.11650/twjm/1500407337.

Filippo Maria Bianchi, Daniele Grattarola, Lorenzo Francesco Livi, and Cesare Alippi. Graph neural networks with convolutional arma filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:3496–3507, 2019.

L. C. Blum and J.-L. Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.*, 131:8732, 2009.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=F72ximsx7C1.

Julian Busch, Jiaxing Pi, and Thomas Seidl. Pushnet: Efficient and adaptive neural message passing. In Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang (eds.), *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pp. 1039–1046. IOS Press, 2020. doi: 10.3233/FAIA200199. URL https://doi.org/10.3233/FAIA200199.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

Mingguo He, Zhewei Wei, Zengfeng Huang, and Hongteng Xu. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 14239–14251, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/76f1cfd7754a6e4fc3281bcccb3d0902-Abstract.html.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Christian Koke. Strong connectivity in graphs: Norm resolvent convergence to effective descriptions, 2024.

Christian Koke and Daniel Cremers. Holonets: Spectral convolutions do extend to directed graphs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EhmEwfavOW.

Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3734–3743. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19c.html.

Renjie Liao, Zhizhen Zhao, Raquel Urtasun, and Richard S. Zemel. Lanczosnet: Multi-scale deep graph convolutional networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BkedznAqKQ.

Zhiyu Liu, Chia-Hsiu Wu, Yujia Zhang, et al. Graph neural network-based prediction of protein-protein interactions in ppi networks. *Nature Communications*, 10(1):1–10, 2019. doi: 10.1038/s41467-019-09308-3.

M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108:058301, 2012.

Xin Shao, Haihong Yang, Xiang Zhuang, Jie Liao, Penghui Yang, Junyun Cheng, Xiaoyan Lu, Huajun Chen, and Xiaohui Fan. scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic acids research*, 49(21):e122–e122, 2021.

T. Tao. *An Introduction to Measure Theory*. Graduate studies in mathematics. American Mathematical Society, 2013. ISBN 9781470409227. URL https://books.google.de/books?id=SPGJjwEACAAJ.

Gerald Teschl. *Mathematical Methods in Quantum Mechanics*. American Mathematical Society, 2014.

Xuebin Zheng, Bingxin Zhou, Junbin Gao, Yuguang Wang, Pietro Lió, Ming Li, and Guido Montúfar. How framelets enhance graph neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12761–12771. PMLR, 2021. URL http://proceedings.mlr.press/v139/zheng21c.html.

A Coarse-graining Graphs

In this Appendix we illustrate:

$$\|(L + Id)^{-1} - J^{\uparrow}(\underline{L} + Id)^{-1}J^{\downarrow}\| \lesssim 1/\lambda_1(\Delta_{\text{high}}).$$

which – as we will see – also implies

$$\|e^{-tL} - J^{\uparrow}e^{-t\underline{L}}J^{\downarrow}\| \lesssim 1/w_{\text{high}}^{\min} \text{ for any } t > 0. \tag{1}$$

after noting the linear relation in scaling behaviour $\lambda_1(L_{\text{cluster}}) \sim w_{\text{high}}^{\min}$.

This will be used to prove the convergence result of Section 3 in Appendix

For convenience, we restate the definitions leading up to this setting again:

Definition A.1. *Denote by $\underline{\mathcal{G}}$ the set of connected components in G_{high}. We give this set a graph structure as follows: Let R and P be elements of $\underline{\mathcal{G}}$ (i.e. connected components in G_{high}). We define the real number*

$$\underline{W}_{RP} = \sum_{r \in R} \sum_{p \in P} W_{rp},$$

with r and p nodes in the original graph G. We define the set of edges $\underline{\mathcal{E}}$ on $\underline{G}$ as

$$\underline{\mathcal{E}} = \{(R, P) \in \underline{\mathcal{G}} \times \underline{\mathcal{G}} : \underline{W}_{RP} > 0\}$$

and assign $\underline{W}_{RP}$ as weight to such edges. Node weights of limit nodes are defined similarly as aggregated weights of all nodes r (in G) contained in the component R as

$$\underline{\mu}_R = \sum_{r \in R} \mu_r.$$

In order to translate signals between the original graph G and the limit description $\underline{G}$, we need translation operators mapping signals from one graph to the other:

Definition A.2. *Denote by $\mathbb{1}_R$ the vector that has 1 as entries on nodes r belonging to the connected (in G_{high}) component R and has entry zero for all nodes not in R. We define the down-projection operator $J^{\downarrow}$ component-wise via evaluating at node R in $\underline{\mathcal{G}}$ as*

$$(J^{\downarrow} x)_R = \langle \mathbb{1}_R, x \rangle / \underline{\mu}_R.$$

The upsampling operator $J^{\uparrow}$ is defined as

$$J^{\uparrow} u = \sum_R u_R \cdot \mathbb{1}_R;$$

where u_R is a scalar value (the component entry of u at $R \in \underline{\mathcal{G}}$) and the sum is taken over all connected components in G_{high}.

As proved in (Koke, 2024), we then have the following:

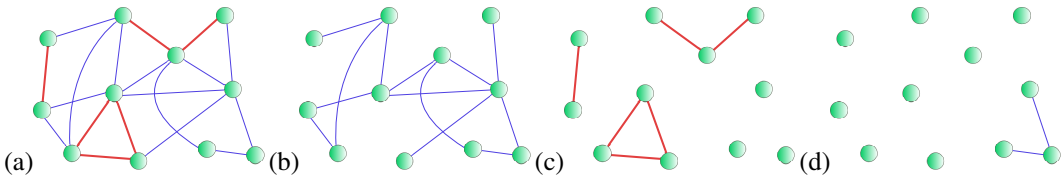

Figure 3: (a) Graph G with $\mathcal{E}_{\text{reg.}}$ (blue) & $\mathcal{E}_{\text{high}}$ (red); (b) $G_{\text{reg.}}$; (c) G_{high}; (d) $G_{\text{reg., exclusive}}$

Theorem A.3. *We have*

$$\left\| R_z(\Delta) - J^{\uparrow} R_z(\underline{\Delta}) J^{\downarrow} \right\| = \mathcal{O}\left(\frac{\|\Delta_{reg.}\|}{\lambda_1(\Delta_{high})} \right)$$

holds; with $\lambda_1(\Delta_{high})$ denoting the first non-zero eigenvalue of Δ_{high}.

We here restate the proof for convenience. We use the notation $\Delta = L$.

Proof. We will split the proof of this result into multiple steps. For $z < 0$ Let us denote by

$$R_z(\Delta) = (\Delta - zId)^{-1},$$
$$R_z(\Delta_{high}) = (\Delta_{high} - zId)^{-1}$$
$$R_z(\Delta_{reg.}) = (\Delta_{reg.} - zId)^{-1}$$

the resolvents correspodning to Δ, Δ_{high} and $\Delta_{reg.}$ respectively.
Our first goal is establishing that we may write

$$R_z(\Delta) = [Id + R_z(\Delta_{high})\Delta_{reg.}]^{-1} \cdot R_z(\Delta_{high})$$

This will follow as a consequence of what is called the second resolvent formula Teschl (2014):

"Given self-adjoint operators A, B, we may write

$$R_z(A+B) - R_z(A) = -R_z(A)BR_z(A+B).\text{"}$$

In our case, this translates to

$$R_z(\Delta) - R_z(\Delta_{high}) = -R_z(\Delta_{high})\Delta_{\text{reg.}}R_z(\Delta)$$

or equivalently

$$\left[Id + R_z(\Delta_{high})\Delta_{\text{reg.}}\right] R_z(\Delta) = R_z(\Delta_{high}).$$

Multiplying with $\left[Id + R_z(\Delta_{high})\Delta_{\text{reg.}}\right]^{-1}$ from the left then yields

$$R_z(\Delta) = \left[Id + R_z(\Delta_{high})\Delta_{reg.}\right]^{-1} \cdot R_z(\Delta_{high})$$

as desired.
Hence we need to establish that $\left[Id + R_z(\Delta_{high})\Delta_{reg.}\right]$ is invertible for $z < 0$.

To establish a contradiction, assume it is not invertible. Then there is a signal x such that

$$\left[Id + R_z(\Delta_{high})\Delta_{reg.}\right] x = 0.$$

Multiplying with $(\Delta_{\text{high}} - zId)$ from the left yields

$$(\Delta_{\text{high}} + \Delta_{\text{reg.}} - zId)x = 0$$

which is precisely to say that

$$(\Delta - zId)x = 0$$

But since Δ is a graph Laplacian, it only has non-negative eigenvalues. Hence we have reached our contradiction and established

$$R_z(\Delta) = \left[Id + R_z(\Delta_{high})\Delta_{reg.}\right]^{-1} R_z(\Delta_{high}).$$

Our next step is to establish that

$$R_z(\Delta_{high}) \to \frac{P_0^{\text{high}}}{-z},$$

where P_0^{high} is the spectral projection onto the eigenspace corresponding to the lowest lying eigenvalue $\lambda_0(\Delta_{high}) = 0$ of Δ_{high}. Indeed, by the spectral theorem for finite dimensional operators (c.f. e.g. Teschl (2014)), we may write

$$R_z(\Delta_{high}) \equiv (\Delta_{high} - zId)^{-1} = \sum_{\lambda \in \sigma(\Delta_{high})} \frac{1}{\lambda - z} \cdot P_\lambda^{high}.$$

Here $\sigma(\Delta_{high})$ denotes the spectrum (i.e. the collection of eigenvalues) of Δ_{high} and the $\{P_\lambda^{high}\}_{\lambda \in \sigma(\Delta_{high})}$ are the corresponding (orthogonal) eigenprojections onto the eigenspaces of the respective eigenvalues. Thus we find

$$\left\| R_z(\Delta_{high}) - \frac{P_0^{high}}{-z} \right\| = \left\| \sum_{0 < \lambda \in \sigma(\Delta_{high})} \frac{1}{\lambda - z} \cdot P_\lambda^{high} \right\|;$$

where the sum on the right hand side now excludes the eigenvalue $\lambda = 0$.

Using orthonormality of the spectral projections, the fact that $z < 0$ and monotonicity of $1/(\cdot + |z|)$ we find

$$\left\| R_z(\Delta_{high}) - \frac{P_0^{high}}{-z} \right\| = \frac{1}{\lambda_1(\Delta_{high}) + |z|}.$$

Here $\lambda_1(\Delta_{high})$ is the firt non-zero eigenvalue of (Δ_{high}).
Non-zero eigenvalues scale linearly with the weight scale since we have

$$\lambda(S \cdot \Delta) = S \cdot \lambda(\Delta)$$

for any graph Laplacian (in fact any matrix) Δ with eigenvalue λ. Thus we have

$$\left\| R_z(\Delta_{high}) - \frac{P_0^{high}}{-z} \right\| = \frac{1}{\lambda_1(\Delta_{high}) + |z|} \leqslant \frac{1}{\lambda_1(\Delta_{high})} \longrightarrow 0$$

as $\lambda_1(\Delta_{high}) \to \infty$.

Our next task is to use this result in order to bound the difference

$$I := \left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} \frac{P_0^{high}}{-z} - \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} R_z(\Delta_{high}) \right\|.$$

To this end we first note that the relation

$$[A + B - zId]^{-1} = [Id + R_z(A)B]^{-1} R_z(A)$$

provided to us by the second resolvent formula, implies

$$[Id + R_z(A)B]^{-1} = Id - B[A + B - zId]^{-1}.$$

Thus we have

$$\begin{aligned} \left\| [Id + R_z(\Delta_{high}) \Delta_{reg.}]^{-1} \right\| &\leqslant 1 + \|\Delta_{\text{reg.}}\| \cdot \|R_z(\Delta)\| \\ &\leqslant 1 + \frac{\|\Delta_{\text{reg.}}\|}{|z|}. \end{aligned}$$

With this, we have

$$\begin{aligned} &\left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} \cdot \frac{P_0^{high}}{-z} - R_z(\Delta) \right\| \\ =& \left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} \cdot \frac{P_0^{high}}{-z} - \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} \cdot R_z(\Delta_{high}) \right\| \\ \leqslant& \left\| \frac{P_0^{high}}{-z} \right\| \cdot \left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} - \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} \right\| + \left\| \frac{P_0^{high}}{-z} - R_z(\Delta_{\text{high}}) \right\| \cdot \left\| [Id + R_z(\Delta_{high}) \Delta_{reg.}]^{-1} \right\| \\ \leqslant& \frac{1}{|z|} \left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} - \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} \right\| + \left(1 + \frac{\|\Delta_{\text{reg.}}\|}{|z|} \right) \cdot \frac{1}{\lambda_1(\Delta_{high})}. \end{aligned}$$

Hence it remains to bound the left hand summand. For this we use the following fact (c.f. Horn & Johnson (2012), Section 5.8. "Condition numbers: inverses and linear systems"):

Given square matrices A, B, C with $C = B - A$ and $\|A^{-1}C\| < 1$, we have

$$\|A^{-1} - B^{-1}\| \leqslant \frac{\|A^{-1}\| \cdot \|A^{-1}C\|}{1 - \|A^{-1}C\|}.$$

In our case, this yields (together with $\|P_0^{high}\| = 1$) that

$$\begin{aligned} &\left\| \left[Id + P_0^{high}/(-z) \cdot \Delta_{reg.} \right]^{-1} - \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} \right\| \\ \leqslant& \frac{(1 + \|\Delta_{\text{reg.}}\|/|z|)^2 \cdot \|\Delta_{\text{reg.}}\| \cdot \| \frac{P_0^{\text{high}}}{-z} - R_z(\Delta_{\text{high}}) \|}{1 - (1 + \|\Delta_{\text{reg.}}\|/|z|) \cdot \|\Delta_{\text{reg.}}\| \cdot \| \frac{P_0^{\text{high}}}{-z} - R_z(\Delta_{\text{high}}) \|} \end{aligned}$$

For S_{high} sufficiently large, we have

$$\| - P_0^{\text{high}}/z - R_z(\Delta_{\text{high}})\| \leqslant \frac{1}{2\left(1 + \|\Delta_{\text{reg.}}\|/|z|\right)}$$

so that we may estimate

$$\begin{aligned}
&\left\| \left[Id + \Delta_{reg.} \frac{P_0^{high}}{-z} \right]^{-1} - \left[Id + \Delta_{reg.} R_z(\Delta_{high}) \right]^{-1} \right\| \\
\leqslant & 2 \cdot (1 + \|\Delta_{\text{reg.}}\|) \cdot \| \frac{P_0^{\text{high}}}{-z} - R_z(\Delta_{\text{high}})\| \\
= & 2 \frac{1 + \|\Delta_{\text{reg.}}\|/|z|}{\lambda_1(\Delta_{\text{high}})}
\end{aligned}$$

Thus we have now established

$$\left| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} \cdot \frac{P_0^{high}}{-z} - R_z(\Delta) \right| = \mathcal{O}\left(\frac{\|\Delta_{\text{reg.}}\|}{\lambda_1(\Delta_{\text{high}})} \right).$$

Hence we are done with the proof, as soon as we can establish

$$\left[-zId + P_0^{high} \Delta_{reg.} \right]^{-1} P_0^{high} = J^{\uparrow} R_z(\underline{\Delta}) J^{\downarrow},$$

with $J^{\uparrow}, \underline{\Delta}, J^{\downarrow}$ as defined above. To this end, we first note that

$$J^{\uparrow} \cdot J^{\downarrow} = P_0^{high} \tag{2}$$

and

$$J^{\downarrow} \cdot J^{\uparrow} = Id_{\underline{G}}. \tag{3}$$

Indeed,the relation (2) follows from the fact that the eigenspace corresponding to the eignvalue zero is spanned by the vectors $\{\mathbb{1}_R\}_R$, with $\{R\}$ the connected components of G_{high}. Equation (3) follows from the fact that

$$\langle \mathbb{1}_R, \mathbb{1}_R \rangle = \underline{\mu}_R.$$

With this we have

$$\left[Id + P_0^{high} \Delta_{reg.} \right]^{-1} P_0^{high} = \left[Id + J^{\uparrow} J^{\downarrow} \Delta_{reg.} \right]^{-1} J^{\uparrow} J^{\downarrow}.$$

To proceed, set

$$\underline{x} := F^{\downarrow} x$$

and

$$\mathscr{X} = \left[P_0^{high} \Delta_{reg.} - zId \right]^{-1} P_0^{high} x.$$

Then

$$\left[P_0^{high} \Delta_{reg.} - zId \right] \mathscr{X} = P_0^{high} x$$

and hence $\mathscr{X} \in \text{Ran}(P_0^{high})$. Thus we have

$$J^{\uparrow} J^{\downarrow} (\Delta_{\text{reg.}} - zId) J^{\uparrow} J^{\downarrow} \mathscr{X} = J^{\uparrow} J^{\downarrow} x.$$

Multiplying with $J^{\downarrow}$ from the left yields

$$J^{\downarrow} (\Delta_{\text{reg.}} - zId) J^{\uparrow} J^{\downarrow} \mathscr{X} = J^{\downarrow} x.$$

Thus we have

$$(J^{\downarrow} \Delta_{\text{reg.}} J^{\uparrow} - zId) J^{\uparrow} J^{\downarrow} \mathscr{X} = J^{\downarrow} x.$$

This – in turn – implies

$$J^{\uparrow} J^{\downarrow} \mathscr{X} = \left[J^{\downarrow} \Delta_{\text{reg.}} J^{\uparrow} - zId \right]^{-1} J^{\downarrow} x.$$

Using

$$P_0^{high}\mathscr{X} = \mathscr{X},$$

we then have

$$\mathscr{X} = J^{\uparrow}\left[J^{\downarrow}\Delta_{\text{reg.}}J^{\uparrow} - zId\right]^{-1}J^{\downarrow}x.$$

We have thus concluded the proof if we can prove that $J^{\downarrow}\Delta_{\text{reg.}}J^{\uparrow}$ is the Laplacian corresponding to the graph $\underline{G}$ defined in Definition A.1. But this is a straightforward calculation. □

As a corollary, we find

Corollary A.4. *We have*

$$R_z(\Delta)^k \to J^{\uparrow}R^k(\underline{\Delta})J^{\downarrow}$$

Proof. This follows directly from the fact that

$$J^{\downarrow}J^{\uparrow} = Id_{\underline{G}}.$$

□

To prove (1), we establish the following theorem:

Theorem A.5. *Consider a graph sequence G_n with $\|(L_n + \lambda Id)^{-1} - \tilde{J}_n(\tilde{L} + \lambda Id)^{-1}J_n\| \to 0$. Then we have $\|\psi(L_n) - \tilde{J}_n\psi(\tilde{L})J_n\| \to 0$ if ψ is complex differentiable and $\lim_{r\to\infty}\psi(r) = 0$.*

Proof. We make use of the holomorphic functional calculus (c.f. e.g. (Koke & Cremers, 2024)) to establish

$$\|\psi(L) - \tilde{J}\psi(\tilde{L})J\| \leqslant \frac{1}{2\pi}\oint_{\Gamma}|\psi(z)|\cdot\|(L - zId)^{-1} - \tilde{J}(\tilde{L} - zId)^{-1}J\|d|z|.$$

Since $\|(L_n + \lambda Id)^{-1} - \tilde{J}_n(\tilde{L} + \lambda Id)^{-1}J_n\| \to 0$ implies $\|(L_n - zId)^{-1} - \tilde{J}_n(\tilde{L} - zId)^{-1}J_n\| \to 0$ uniformly (in z) on compact sets (c.f. e.g. Arendt (2001)), we can apply dominated convergence, if we find an majorizing function that is integrable on Γ. But this is ensured by the decay of ψ. □

Choosing the function ψ to be given as $\psi(z) = e^{-tz}$ then establishes (1).

B Global Laplacian Propagation Matrices, Generalized Functions, Measures and all that

In this section we discuss global Laplacian propagation matrices, generalized functions and measures

B.1 Complex measures on $\mathbb{R}_{\geqslant 0}$ and their Theory of Integration

As reference for this section Tao (2013) might serve.

In mathematics, a measure is a formal generalization of concepts such as length, area and volume.

More specifically, we are here interested in assigning a generalized notion of length (or mass) to subsets of the real half-line

$$\mathbb{R}_{\geqslant 0} = [0, \infty).$$

These sets will turn out to be elements of a so called σ-Algebra; i.e. a set Σ of sets for which

- $\varnothing, \mathbb{R}_{\geqslant 0} \in \Sigma$
- $A, B \in \sigma \Rightarrow A \cap B \in \Sigma$
- $A, B \in \Sigma \Rightarrow A \backslash B \in \Sigma$
- $A, B \in \Sigma \Rightarrow A \cup B \in \Sigma$.

We now take $\Sigma_{\mathbb{R}_{\geqslant 0}}$ to be the smallest such set of sets Σ that contains all open intervals.

A complex measure then is a set-function that assigns to each set in $\Sigma_{\mathbb{R}_{\geqslant 0}}$ a complex number in a certain way:

Definition B.1. *A complex measure μ on $\mathbb{R}_{\geqslant 0}$ is a complex valued function $\mu : \Sigma_{\mathbb{R}_{\geqslant 0}} \to \mathbb{C}$ satisfying*

$$\mu\left(\bigcup_n A_n\right) = \sum_n \mu\left(A_n\right)$$

for any countable (potentially infinite) collection of sets in $\Sigma_{\mathbb{R}_{\geqslant 0}}$ which are pairwise disjoint.

Let us provide some examples:

Example B.2. *The prototypical example of a measure is the standard Lebesgue measure that assigns to any interval (a, b) the length $\mu_{Leb}((a, b)) = |a - b|$ $(a, b \in \mathbb{R}_{\geqslant 0})$.*

Example B.3. *Alternatively, we might consider the Dirac measure $\mu_{\delta_{t_0}}$, which assigns the value $\mu_{\delta_{t_0}}((a, b)) = 1$ to any interval (a, b) containing t_0 (i.e. $t_0 \in (a, b)$). Otherwise it assigns the value $\mu_{\delta_{t_0}}((a, b)) = 0$ if $t_0 \notin (a, b)$.*

Example B.4. *Every integrable function $\hat{\psi} : \mathbb{R}_{\geqslant 0} \to \mathbb{C}$ defines a complex measure via $\mu_{\hat{\psi}}((a, b)) = \int_a^b \hat{\psi}(t)dt$.*

Hence we may think of **measures as generalizations of functions**.

Any given measure on $\mathbb{R}_{\geqslant 0}$ defines a unique way of integrating (known as Lebesgue integration) a function f defined on $\mathbb{R}_{\geqslant 0}$. This proceeds by approximating any function f via a weighted sequence of indicator functions (with $A \in \Sigma_{\mathbb{R}_{\geqslant 0}}$ a set)

$$\chi_A(t) = \begin{cases} 1 & ; t \in A \\ 0 & ; t \notin A \end{cases}.$$

as

$$f(t) \approx f_n(t) := \sum_k a_k^n \chi_{A_k}(t).$$

with $a_k \in \mathbb{C}$. For these functions, one then sets

$$\int_{\mathbb{R}_{\geqslant 0}} f_n d\mu \equiv \sum_k a_k^n \cdot \mu(A_k).$$

Since we have $\lim_{n\to\infty} f_n = f$, one then simply sets

$$\int_{\mathbb{R}_{\geqslant 0}} f d\mu \equiv \lim_{n\to\infty} \int_{\mathbb{R}_{\geqslant 0}} f_n d\mu.$$

Example B.5. *For the prototypical example of the standard Lebesgue measure, this process simply yields*

$$\int_{\mathbb{R}_{\geqslant 0}} f(t) d\mu_{Leb}(t) = \int_0^\infty f(t)dt.$$

Example B.6. *For the Dirac measure $\mu_{\delta_{t_0}}$, the above process yields*

$$\int_{\mathbb{R}_{\geqslant 0}} f(t) d\mu_{\delta_{t_0}}(t) = f(t_0)$$

Example B.7. *For measures arising from integrable functions $\hat{\psi} : \mathbb{R}_{\geqslant 0} \to \mathbb{C}$ as $\mu_{\hat{\psi}}((a, b)) = \int_a^b \hat{\psi}(t)dt$, we find*

$$\int_{\mathbb{R}_{\geqslant 0}} f(t) d\mu_{\hat{\psi}} = \int_0^\infty \hat{\psi}(t) f(t) dt.$$

B.2 LAPLACE TRANSFORMS

We say a complex valued measure μ is finite if we have

$$\int_{\mathbb{R}_{\geqslant 0}} d|\mu|(t) < \infty.$$

Here the measure $|\mu|$ arises from the original measure μ via

$$|\mu|((a,b)) \equiv |\mu((a,b))|.$$

For any such finite measure μ we may define its Laplace transform as

$$\psi_\mu(z) := \int_{\mathbb{R}_{\geqslant 0}} e^{-tz} d\mu(t).$$

This function f_μ is well defined for z in the right hemisphere

$$\mathbb{C}_R := \{z \in \mathbb{C} : \mathrm{Re}(z) \geqslant 0\}.$$

of the complex plane $\mathbb{C}$, since there we have

$$\begin{aligned} |\psi_\mu(z)| &= \left| \int_{\mathbb{R}_{\geqslant 0}} e^{-tz} d\mu(t) \right| \\ &\leqslant \int_{\mathbb{R}_{\geqslant 0}} |e^{-tz}| d|\mu|(t) \\ &\leqslant \int_{\mathbb{R}_{\geqslant 0}} d|\mu|(t) < \infty. \end{aligned}$$

Example B.8. *For the Dirac measure $\mu_{\delta_{t_0}}$, we have*

$$\psi_{\mu_{\delta_{t_0}}}(z) = e^{-t_0 z}.$$

Example B.9. *For any integrable function $\hat{\psi}$, we have*

$$\psi(z) \equiv \int_{\mathbb{R}_{\geqslant 0}} e^{-tz} d\mu_{\hat{\psi}} = \int_0^\infty \hat{\psi}(t) e^{-tz} dt.$$

More specifically, if the integrable function is given as $\hat{\psi}_k := (-t)^{k-1} e^{-\lambda t}$ (with $\mathrm{Re}(\lambda) > 0$), then $\psi_k(z) = (z+\lambda)^{-k}$:

Example B.10. *If $\hat{\psi}_k := (-t)^{k-1} e^{-\lambda t}$ yields $\psi_k(z) = (z+\lambda)^{-k}$, then*

$$\psi_k(z) = (z+\lambda)^{-k}.$$

For $k = 1$, this can be seen from

$$\int_0^\infty e^{-tz} e^{-\lambda t} dt = -\frac{1}{z+\lambda} e^{-(z+\lambda)} \Big|_0^\infty.$$

For $k > 1$, the claim follows from differentiating the above expression with respect to z Note that the functions $\psi_k(z) = (z+\lambda)^{-k}$ are also defined if Re$(z) \leqslant 0$, as long as $z \neq -\lambda$.

Using the function ψ_k of the examples above, a wide class of functions may be parametrized

Theorem B.11. *Let $f : \mathbb{R}_{\geqslant 0} \to 0$ be any function with $\lim_{x\to\infty} f(x) = 0$. Then for any $\epsilon > 0$, there is a function*

$$h(x) = \sum_k \theta_k \psi_k(x)$$

for which

$$\sup_{x \in [0,\infty)} |f(x) - h(x)| < \epsilon.$$

Here the basis functions $\{\psi_k\}$ may either be chosen as $\psi_k(z) = (z+\lambda)^{-k}$ or $\psi_k(x) = e^{-(kt_0)x}$ for any $t_0 > 0$.

Proof. This is a direct consequence of the Weierstrass approximation theorem. □

B.3 Global Laplacian Propagation matrices

A Global Laplacian Propagation matrix is then constructed by applying a function ψ arising as a Laplace transform to a graph Laplacian L. The resulting filter matrix $\psi(L) \in \mathbb{R}^{N\times N}$ acts on scalar graph signals $x \in \mathbb{R}^N$ via matrix multiplication; sending x to $\psi(L)\cdot x$:

$$x \mapsto \psi(L)\cdot x$$

C Proofs related to Generalization Ability

C.1 Generalization Ability of global Laplacian Propagation Matrices

In this section, we establish the generalization ability of global Laplacian propagation matrices.

Theorem C.1. *We have that $\|\psi(L) - J^{\uparrow}\psi(\underline{L})J^{\downarrow}\| \leqslant \int_0^{\infty} |\hat{\psi}(t)|\eta(t)dt$ holds true.*

Proof. We start by proving the first claim. To this end, we note

$$\begin{aligned}\|\psi(L) - J^{\downarrow}\psi(\underline{L})J^{\downarrow}\| &= \left\| \int_{\mathbb{R}_{\geqslant 0}} \left[e^{-tL} - J^{\uparrow}e^{-t\underline{L}}J^{\downarrow}\right] d\mu_{\hat{\psi}} \right\| \\ &\leqslant \int_{\mathbb{R}_{\geqslant 0}} \left\| e^{-tL} - J^{\uparrow}e^{-t\underline{L}}J^{\downarrow} \right\| d|\mu|_{\hat{\psi}}\end{aligned}$$

Using the notation for generalized fundtions, we have $d|\mu|_{\hat{\psi}}(t) = |\hat{\psi}(t)|dt$ and hence

$$\begin{aligned}\|\psi(L) - J^{\downarrow}\psi(\underline{L})J^{\downarrow}\| &= \left\| \int_{\mathbb{R}_{\geqslant 0}} \left[e^{-tL} - J^{\uparrow}e^{-t\underline{L}}J^{\downarrow}\right] d\mu_{\hat{\psi}} \right\| \\ &\leqslant \int_{\mathbb{R}_{\geqslant 0}} \left\| e^{-tL} - J^{\uparrow}e^{-t\underline{L}}J^{\downarrow} \right\| |\hat{\psi}(t)|dt.\end{aligned}$$

□

Thus if $\eta(t) \equiv \left\| e^{-tL} - J^{\uparrow}e^{-t\underline{L}}J^{\downarrow} \right\| \approx 0$ on the support of $\hat{\psi}$, we also have $\|\psi(L) - J^{\uparrow}\psi(\underline{L})J^{\downarrow}\| \approx 0$. In this case, propagation as implemented via $\psi(L)$ is essentially the same as propagation via $J^{\downarrow}\psi(\underline{L})J^{\downarrow}$.

C.2 Generalization and stability when $\|L - \tilde{L}\| \ll 1$

In this section we prove in addition to results in the main body of the paper also stability and generalization ability in the setting where for the Laplacians $L, \tilde{L}$ of two graphs $G, \tilde{G}$ defined on a common node set we have $\|L - \tilde{L}\| \ll 1$ (as opposed to the setting where one graph is a coarser version of another). We denote the collection of weight matrices by $\mathcal{W}$, the collection of biases by $\mathcal{B}$ and the (collection of) utilized global Laplacian propagation matrices used in the update rule "$X \mapsto \sum_k \psi_k(L)XW_k$" as Ψ. We denote the network by $\Phi_{\mathcal{W},\mathcal{B},\Psi}$ and write the generated embeddings for the node feature matrix X as $\Phi_{\mathcal{W},\mathcal{B},\Psi}(X)$. With this, we have:

Theorem C.2. *Let $\Phi_{\mathscr{W},\mathscr{B},\Psi}$ be a K-layer deep graph convolutional architecture. Assume in each layer $1 \leqslant \ell \leqslant K$ that $\sum_i \|W_i^{\ell}\| \leqslant W$ and $\|B^{\ell}\| \leqslant B$. Choose $C \geqslant \|\Psi_i(L)\|$ $(\forall i \in I)$ and w.l.o.g. assume $CW > 1$. With this, we have with $\delta = \max_{i\in I}\{\|\Psi_i(L) - \Psi_i(\widetilde{L})\|\}$ that*

$$\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},X)\| \leqslant \left[K\cdot C^K W^{K-1}\cdot\left(\|X\| + \frac{1}{CW-1}B\right)\right]\cdot\delta.$$

Proof. For simplicity in notation, let us denote the hidden representations in the network corresponding to $\tilde{L}$ by X^ℓ. With this, we note:

$$\begin{aligned}
\|X^K - \tilde{X}^K\| &\leqslant \sum_{i\in I} \|\psi_i(L) - \psi_i(\tilde{L})\| \cdot \|X^{K-1}\| \cdot \|W_i^K\| + \sum_{i\in I} \|\psi_i(\tilde{L})\| \cdot \|\tilde{X}^{K-1} - X^{K-1}\| \cdot \|W_i^K\| \\
&\leqslant \delta W \|X^{K-1}\| + CW\|\tilde{X}^{K-1} - X^{K-1}\| \\
&\leqslant \delta W \|X^{K-1}\| + CW\delta\|X^{K-2}\| + (CW)^2\|\tilde{X}^{K-1} - X^{K-1}\| \\
&\leqslant \frac{\delta}{C} \cdot \left(\sum_{\ell=1}^{K} (CW)^\ell \|X^{K-\ell}\| \right) \\
&= \frac{\delta}{C} \cdot \left(\sum_{j=0}^{K-1} (CW)^{K-j} \|X^j\| \right)
\end{aligned}$$

Hence we need to bound the quantity $\|X^j\|$ in terms of C, W, B and X.

We have

$$\begin{aligned}
\|X^j\| &\leqslant \sum_i \|\psi_i(L)\| \cdot \|X^{j-1}\| \cdot \|W_i^j| + \|B^J\| \\
&\leqslant CW\|X^{j-1}\| + B \\
&\leqslant (CW)^2\|X^{j-2}\| + CWB + B \\
&\leqslant B\left(\sum_{k=0}^{j-1} (CW)^k \right) + (CW)^j\|X\| \\
&= \begin{cases} B\frac{(CW)^j - 1}{CW - 1} + (CW)^j\|X\| & ; CW \neq 1 \\ jB + \|X\| & ; CW = 1 \end{cases}.
\end{aligned}$$

For the case $CW = 1$, we thus find

$$\begin{aligned}
\|X^K - \tilde{X}^K\| &\leqslant \frac{\delta}{C} \cdot \left(\sum_{j=0}^{K-1} (jB + \|X\|) \right) \\
&= \frac{\delta}{C} \cdot \left(K\|X\| + B\frac{K(K-1)}{2} \right).
\end{aligned}$$

For the case $CW \neq 1$, we find

$$\|X^K - \tilde{X}^K\| \leqslant \frac{\delta}{C} \cdot \left(\sum_{j=0}^{K-1} (CW)^{K-j} \left[B\frac{(CW)^j - 1}{CW - 1} + (CW)^j\|X\| \right] \right)$$

For $CW > 1$, we may further estimate this as

$$\begin{aligned}
\|X^K - \tilde{X}^K\| &\leqslant \frac{\delta}{C} \cdot \left(\sum_{j=0}^{K-1} (CW)^{K-j} \left[B\frac{(CW)^j - 1}{CW - 1} + (CW)^j\|X\| \right] \right) \\
&\leqslant \delta \cdot \frac{K(CW)^K}{C} \left[\frac{B}{CW - 1} + \|X\| \right].
\end{aligned}$$

This proves the claim. □

C.3 Proof of Convergence in the sense of Section 3

The result in Section 3 is concerned with the graph-level setting; i.e. the setting where entire graphs are embedded into latent spaces. Before proving this result, we first prove a corresponding result for the node-level, where individual nodes in a graph are embedded. We will then use this node-level

result (Thoerem C.3 below) to prove the graph-level Convergence result in section 3.

In the node-level setting, we start by considering initial node-features X on G. We then fix a graph neural network Φ based on global Laplacian propagation schemes and consider two ways of generating embeddings on the graph G: On the one hand, we may simply generate embeddings with the network Φ on G. On the other hand, we may also project the node feature matrix X to $\underline{G}$ via $J^{\downarrow}$, apply ne the network Φ to the matrix $J^{\downarrow}X$ on $\underline{G}$ and then finally interpolate the generated node embeddings back to G via $J^{\uparrow}$.

The following result bounds the difference between these two respective node embeddings generated on the same graph.

Theorem C.3. *Let $\Phi_{\mathscr{W},\mathscr{B},\Psi}$ be a K-layer deep Global-Laplacian-Propagation-based network. Assume $\sum_{i\in I}\|W_i^{\ell}\| \leqslant W$ and bound bias matrices in layer ℓ as $\|B^{\ell}\| \leqslant B$. Choose $C \geqslant \|\Psi_i(L)\|$ $(i \in I)$ and w.l.o.g. assume $CW > 1$ (which can always be satisfied by choosing C large enough). Assume $\rho(J^{\uparrow}X) = J^{\uparrow}\rho(X)$ and if biases are enabled, assume $J^{\uparrow}\mathbb{1}_{\underline{G}} = \mathbb{1}_G$. Set $\max_{i\in I}\{\|\psi_i(L) - J^{\uparrow}\psi_i(\underline{L})J^{\downarrow}\|\} = \delta_1$ and define $\delta_2 = \max_{i\in I}\{\|\psi_i(L^{\uparrow})[J^{\downarrow}J^{\uparrow} - Id_{\underline{G}}]\|\}$. With this, we have that*

$$\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - J^{\uparrow}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L},J^{\downarrow}X)\| \leqslant \left[K\cdot C^K W^{K-1}\cdot\left(\|X\| + \frac{1}{CW-1}B\right)\right]\cdot(\delta_1+\delta_2).$$

It should be noted that the result above is more general than the setting considered in Section 3. In the setting considered in Section 3 we have $J^{\downarrow}J^{\uparrow} = Id_{\underline{G}}$ (in addition to $\rho(J^{\uparrow}X) = J^{\uparrow}\rho(X)$). There we thus automatically have $\delta_2 = 0$.

Proof. Let us define

$$\underline{X} := J^{\downarrow}X.$$

Let us further use the notation $\underline{\psi}_i := \psi_i(\underline{L})$ and $\psi_i := \psi_i(L)$.

Denote by X^{ℓ} and $\underline{X}^{\ell}$ the (hidden) feature matrices generated in layer ℓ for networks based on ψ_i and $\underline{\psi}_i$ respectively: I.e. we have

$$X^{\ell} = \rho\left(\sum_{i\in I}\psi_i X^{\ell-1}W_i^{\ell} + B^{\ell}\right)$$

and

$$\underline{X}^{\ell} = \rho\left(\sum_{i\in I}\underline{\psi}_i\underline{X}^{\ell-1}W_i^{\ell} + \underline{B}^{\ell}\right).$$

We then have

$$\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - J^{\uparrow}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L},J^{\downarrow}X)\|\\
=&\|X^K - J^{\uparrow}\underline{X}^K\|\\
=&\left\|\rho\left(\sum_{i\in I}\psi_i X^{K-1}W_i^K + B^K\right) - J^{\uparrow}\rho\left(\sum_{i\in I}\underline{\psi}_i\underline{X}^{K-1}W_i^K + \underline{B}^L\right)\right\|\\
=&\left\|\rho\left(\sum_{i\in I}\psi_i X^{K-1}W_i^K + B^K\right) - \rho\left(\underline{J}\sum_{i\in I}\underline{\psi}_i\underline{X}^{K-1}W_i^K + B^L\right)\right\|
\end{aligned}$$

Here we used the assumption that ρ and $\underline{J}$ commute. In fact since $\mathrm{ReLU}(\cdot)$ maps positive entries to positive entries and acts pointwise, it commutes with $J^{\uparrow}$. We also made use of the assumption $J^{\uparrow}\mathbb{1}_{\underline{G}} = \mathbb{1}_G$ when dealing with biases .

Using the fact that $\rho(\cdot)$ is 1-Lipschitz-continuous, we can establish

$$\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - J^{\uparrow}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L},JX)\| \\
\leqslant &\left\|\rho\left(\sum_{i\in I}\psi_i X^{K-1}W_i^K + B^K\right) - \rho\left(J^{\uparrow}\sum_{i\in I}\underline{\psi}_i\underline{X}^{K-1}W_i^K + B^L\right)\right\| \\
\leqslant &\left\|\sum_{i\in I}\psi_i X^{K-1}W_i^K + B^K - J^{\uparrow}\sum_{i\in I}\underline{\psi}_i\underline{X}^{K-1}W_i^K + B^K\right\|.
\end{aligned}$$

Using the assumption that $\|\underline{\psi}[J^{\downarrow}J^{\uparrow} - Id_{\underline{G}}]\| \leqslant \delta_2$, we have

$$\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - J^{\uparrow}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L},JX)\| \\
\leqslant &\left\|\sum_{i\in I}\psi_i X^{K-1}W_i^K - \sum_{i\in I}(J^{\uparrow}\underline{\psi}_i J)J^{\uparrow}\underline{X}^{K-1}W_i^K\right\| + \left\|\sum_{i\in I}J^{\uparrow}\underline{\psi}_i[Id_{\underline{G}} - J^{\downarrow}J^{\uparrow}]\underline{X}^{K-1}W_i^K\right\| \\
\leqslant &\left\|\sum_{i\in I}\psi_i X^{K-1}W_i^K - \sum_{i\in I}(J^{\uparrow}\underline{\psi}_i J)J^{\uparrow}\underline{X}^{K-1}W_i^K\right\| + \delta_2\cdot\left\|\sum_{i\in I}\underline{X}^{K-1}W_i^K\right\| \\
\leqslant &\left\|\sum_{i\in I}\psi_i X^{K-1}W_i^K - \sum_{i\in I}(J^{\uparrow}\underline{\psi}_i J^{\downarrow})J^{\uparrow}\underline{X}^{K-1}W_i^K\right\| + \delta_2\cdot\left\|\underline{X}^{K-1}\right\|\cdot W
\end{aligned}$$

From this, we find (assuming $\|J^{\uparrow}\|, \|J^{\downarrow}\| \leqslant 1$ for notational simplicity (and which is true in the setting of Section 3)), that

$$\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - J^{\uparrow}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L},JX)\| \\
\leqslant &\left\|\sum_{i\in I}\psi_i X^{K-1}W_i^K - \sum_{i\in I}(J^{\uparrow}\underline{\psi}_i J^{\downarrow})J^{\uparrow}\underline{X}^{K-1}W_i^K\right\| + \delta_2\cdot\left\|\underline{X}^{K-1}\right\|\cdot W \\
\leqslant &\left\|\sum_{i\in I}(\psi_i - J^{\uparrow}\underline{\psi}_i J)X^{K-1}W_i^K\right\| + \sum_{i\in I}\|J^{\uparrow}\underline{\psi}_i J\|\cdot\|J^{\uparrow}\underline{X}^{K-1} - X^{K-1}\|\cdot\|W_i^K\| + \delta_2\cdot\left\|\underline{X}^{K-1}\right\|\cdot W \\
\leqslant &\left\|\sum_{i\in I}(\psi_i - J^{\uparrow}\underline{\psi}_i J)X^{K-1}W_i^K\right\| + CW\cdot\|J^{\uparrow}\underline{X}^{K-1} - X^{K-1}\| + \delta_2\cdot\left\|\underline{X}^{K-1}\right\|\cdot W \\
\leqslant &\sum_{i\in I}\left\|(\psi_i - J^{\uparrow}\underline{\psi}_i J)\right\|\cdot\left\|X^{K-1}\right\|\cdot\left\|W_i^K\right\| + CW\cdot\|J^{\uparrow}\underline{X}^{K-1} - X^{K-1}\| + \delta_2\cdot\left\|\underline{X}^{K-1}\right\|\cdot W \\
\leqslant &\delta_1\cdot\left\|X^{K-1}\right\|W + CW\cdot\|J^{\uparrow}\underline{X}^{K-1} - X^{K-1}\| + \delta_2\cdot\left\|\underline{X}^{K-1}\right\|\cdot W
\end{aligned}$$

Arguing as in the proof of Appendix C.2 then yields the claim.

□

Let us move from the node-level to the graph-level. We first specify how graph-level latent embeddings arise:

Definition C.4. *We aggregate embeddings $X \in \mathbb{R}^{N\times F}$ of individual nodes to graph-embeddings $\Omega(X) \in \mathbb{R}^F$ as $\Omega(X)_j = \sum_{i=1}^N |X_{ij}|\cdot\mu_i$. Here $\{\mu_i\}_i$ is the set of node-weights.*

In a social network, a node weight $\mu_i = 1$ might e.g. signify that node i represents a single user. A weight $\mu_j > 1$ would indicate that node j represents a group of users.
Given such an aggregation of node embeddings into latent-embeddings of entire graphs, we may then relegate graph-level transferability back to node-level transferability:

Theorem C.5. *Assuming $\Omega(\underline{X}) = \Omega(J^{\uparrow}\underline{X})$, we have in the setting of Theorem C.3 that $\|\Omega\circ\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Omega\circ\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L},J^{\downarrow}X)\| \leqslant \|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - J^{\uparrow}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L},J^{\downarrow}X)\|$.*

Proof. We note

$$\begin{aligned}&\|\Omega \circ \Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Omega \circ \Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L}, J^{\downarrow}X)\| \\ =&\|\Omega(\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)) - \Omega(\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L}, J^{\downarrow}X))\| \\ =&\|\Omega(\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)) - \Omega(J^{\uparrow}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L}, J^{\downarrow}X))\|.\end{aligned}$$

To prove the claim from here, we only have to note that the aggregation method Ω as defined in Definition C.5 above is 1-Lipschitz (as a consequence of the reverse triangle inequality). The proof for the bidirectional setting proceeds analogously. □

This result then proves the continuity result of Section 3. Indeed: In the notation of Section 3, we have $F_\omega = \Omega(\Phi_{\mathscr{W},\mathscr{B},\Psi}(L_\omega, X))$ and $\underline{F} = \Omega(\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L}, J^{\downarrow}X))$ Thus we have

$$\|F_\omega - \underline{F}\| = \|\Omega \circ \Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Omega \circ \Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L}, J^{\downarrow}X)\| \leqslant \|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L_\omega, X) - J^{\uparrow}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L}, J^{\downarrow}X)\|.$$

By Theorem C.3 and the fact that $[Id_{\underline{G}} - J^{\uparrow}J^{\downarrow}] = 0$, we have

$$\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L_\omega, X) - J^{\uparrow}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\underline{L}, J^{\downarrow}X)\| \lesssim \max_k \{\|\psi_k(L_\omega) - J^{\uparrow}\psi_k(\underline{L})J^{\downarrow}\|\},$$

with "$\lesssim$" as per usual "denoting smaller than, up to a positive multiplicative constant".

Finally Theorem C.1 implies

$$\|\psi_k(L_\omega) - J^{\uparrow}\psi_k(\underline{L})J^{\downarrow}\| \leqslant \int_0^\infty |\hat{\psi}_k(t)|\eta(t)dt = \int_{\mathbb{R}_{\geqslant 0}} \|e^{-tL_\omega} - J^{\uparrow}e^{-t\underline{L}}J^{\downarrow}\|\,|\hat{\psi}_k(t)|dt.$$

Thus upon combining these steps and noting that $\eta_\omega \to 0$ by (1), the convergence result of Section 3 is indeed proved.

D Additional Experimental Considerations

Collapsing strongly connected clusters: Intuition and exact Definitions

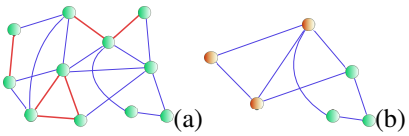

Figure 4: (a) G (stongly connected) clusters in red (b) Coarse grained $\underline{G}$

From a diffusion perspective, information in a graph equalizes faster along edges with large weights. In the limit where edge-weights within certain sub-graphs tend to infinity, information within these clusters equalizes immediately and such sub-graphs thus effectively behave as single nodes. We might thus consider a coarse grained graph $\underline{G}$ where these strongly connected clusters are indeed fused together and represented only via single nodes. The corresponding node set $\underline{\mathcal{G}}$ of $\underline{G}$ is then given by the set of connected components in G_{cluster} (c.f. Fig 5). Edges $\underline{\mathcal{E}}$ are given by elements $(R,P) \in \underline{\mathcal{G}} \times \underline{\mathcal{G}}$ with non-zero accumulated edge weight $\underline{W}_{RP} = \sum_{r\in R}\sum_{p\in P} W_{rp}$. Node weights in $\underline{G}$ are defined accordingly by aggregating as $\underline{\mu}_R = \sum_{r\in R}\mu_r$. To compare signals on these two graphs, we define intertwining operators $J^{\downarrow}, J^{\uparrow}$ transferring information between G and $\underline{G}$: Let x be a scalar graph signal and let $\mathbb{1}_R$ be the vector that has 1 as entry for nodes $r \in R$ and is zero otherwise. Denote by u_R the entry of u at node $R \in \underline{\mathcal{G}}$. Projection $J^{\downarrow}$ is then defined component-wise by evaluation at node $R \in \underline{\mathcal{G}}$ as the average of x over R: $(J^{\downarrow}x)_R = \langle \mathbb{1}_R, x\rangle / \underline{\mu}_R$. Going in the opposite direction, interpolation is defined as $J^{\uparrow}u = \sum_{R\in\underline{\mathcal{G}}} u_R \cdot \mathbb{1}_R$.

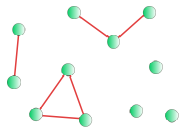

Figure 5: G_{cluster}

In this setting, we have (c.f. Appendix A) that

$$\|e^{-tL} - J^{\uparrow}e^{-t\underline{L}}J^{\downarrow}\| \lesssim 1/w_{\text{high}}^{\min} \text{ for any } t > 0.$$

Here $w_{\text{high}}^{\min} \gg 1$ denotes the minimal edge weight inside the strongly connected clusters in G.

Dataset: The dataset we consider is the **QM7** dataset, introduced in Blum & Reymond (2009); Rupp et al. (2012). This dataset contains descriptions of 7165 organic molecules, each with up to seven heavy atoms, with all non-hydrogen atoms being considered heavy. A molecule is represented by its Coulomb matrix C^{Clmb}, whose off-diagonal elements

$$C_{ij}^{\text{Clmb}} = \frac{Z_i Z_j}{|R_i - R_j|}$$

correspond to the Coulomb-repulsion between atoms i and j. We discard diagonal entries of Coulomb matrices; which would encode a polynomial fit of atomic energies to nuclear charge Rupp et al. (2012).

For each atom in any given molecular graph, the individual Cartesian coordinates R_i and the atomic charge Z_i are (in principle) also accessible individually. To each molecule an atomization energy - calculated via density functional theory - is associated. The objective is to predict this quantity. The performance metric is mean absolute error. Numerically, atomization energies are negative numbers in the range -600 to -2200. The associated unit is $[\textit{kcal/mol}]$.

Details on collapsing procedure as applied to QM7: Again, we make use of the QM7 dataset Rupp et al. (2012) and its Coulomb matrix description

$$C_{ij}^{\text{Clmb}} = \frac{Z_i Z_j}{|R_i - R_j|} \tag{4}$$

of molecules. We modify (all) molecular graphs in QM7 by deflecting hydrogen atoms (H) out of their equilibrium positions towards the respective nearest heavy atom. This is possible since the QM7 dataset also contains the Cartesian coordinates of individual atoms. Edge weights between heavy atoms then remain the same, while Coulomb repulsions between H-atoms and respective nearest heavy atom increasingly diverge; as is evident from (4).

Given an original molecular graph G with node weights $\mu_i = Z_i$, the corresponding limit graph $\underline{G}$ corresponds to a coarse grained description, where heavy atoms and surrounding H-atoms are aggregated into single super-nodes.

Mathematically, $\underline{G}$ is obtained by removing all nodes corresponding to H-atoms from G, while adding the corresponding charges $Z_H = 1$ to the node-weights of the respective nearest heavy atom. Charges in (4) are modified similarly to generate the weight matrix $\underline{W}$.

On original molecular graphs, atomic charges are provided via one-hot encodings. For the graph of methane – consisting of one carbon atom with charge $Z_C = 6$ and four hydrogen atoms of charges $Z_H = 1$ – the corresponding node-feature-matrix is e.g. given as

$$X = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0\cdots \\ 1 & 0 & \cdots & 0 & 0 & 0\cdots \\ 1 & 0 & \cdots & 0 & 0 & 0\cdots \\ 1 & 0 & \cdots & 0 & 0 & 0\cdots \\ 1 & 0 & \cdots & 0 & 0 & 0\cdots \end{pmatrix}$$

with the non-zero entry in the first row being in the 6^{th} column, in order to encode the charge $Z_C = 6$ for carbon.

The feature vector of an aggregated node represents charges of the heavy atom and its neighbouring H-atoms jointly.

Node feature matrices are translated as $\underline{X} = J^{\downarrow} X$. Applying $J^{\downarrow}$ to one-hot encoded atomic charges yields (normalized) bag-of-word embeddings on $\underline{G}$: Individual entries of feature vectors encode how much of the total charge of the super-node is contributed by individual atom-types. In the example of methane, the limit graph $\underline{G}$ consists of a single node with node-weight

$$\mu = 6 + 1 + 1 + 1 + 1 = 10.$$

The feature matrix

$$\underline{X} = J^{\downarrow} X$$

is a single row-vector given as

$$\underline{X} = \left(\frac{4}{10}, 0, \cdots, 0, \frac{6}{10}, 0, \cdots \right).$$

Experimental Setup: We randomly select 1500 molecules for testing and train on the remaining graphs. On QM7 we run experiments for 23 different random random seeds and report mean and standard deviation. All experiments were performed on a single NVIDIA Quadro RTX 8000 graphics card.

Additional details on training and models: Typical GNN models are divided into **standard** architectures (GCN (Kipf & Welling, 2017), ChebNet (Defferrard et al., 2016), ARMA (Bianchi et al., 2019), BernNet (He et al., 2021), GATv2 (Brody et al., 2022)) and **multi- scale** architectures (PushNet (Busch et al., 2020), UFGNet (Zheng et al., 2021), Lanczos (Liao et al., 2019)). Apart from UFGNet (already acting as a **pooling** layer) we also consider self-attention-pooling (Lee et al., 2019); both acting on the final layer (SAG) and as acting on the output of each indivifual layer, with resulting layer-wise features concatenated to produce the final embedding (SAG-M). All considered convolutional layers are incorporated into a two layer deep and fully connected graph convolutional architecture. In each hidden layer, we set the width (i.e. the hidden feature dimension) to

$$F_1 = F_2 = 64.$$

For BernNet, we set the polynomial order to $K = 3$ to combat appearing numerical instabilities. ARMA is set to $K = 2$ and $T = 1$. ChebNet uses $K = 2$. Lnaczos uses 20 Lanczos iterations, as proposed in the original paper (Liao et al., 2019). UFGNet uses Haar wavelets. For all baselines, the standard mean-aggregation scheme is employed after the graph-convolutional layers to generate graph level features. Finally, predictions are generated via an MLP.

For the **resolvent** based global Laplacian propagation architecture, we set $\lambda = 1$ and and build filters using the $k = 1$ and $= 2$ matrices in $\Psi^{\text{Res}} = \{(z + \lambda)^{-k}\}_{k\in\mathbb{N}}$.

For the**based global Laplacian propagation architecture,** based global Laplacian propagation architecture, we set $t_0 = 1$ and and build filters using the $k = 1$ and $= 2$ matrices in $\Psi^{\text{Exp}} = \{e^{-(kt_0)z}\}_{k\in\mathbb{N}}$.

As aggregation, we employ the graph level feature aggregation scheme introduced in Definition C.4 with node weights set to atomic charges of individual atoms. Predictions are then generated via a final MLP with the same specifications as the one used for baselines.