Graph Neural Networks as Gradient Flows

Anonymous Author(s)
Affiliation
Address
email

Abstract

Dynamical systems minimizing an energy are ubiquitous in geometry and physics. We propose a gradient flow framework for GNNs where the equations follow the direction of steepest descent of a learnable energy. This approach allows to analyse the GNN evolution from a multi-particle perspective as learning attractive and repulsive forces in feature space via the positive and negative eigenvalues of a symmetric 'channel-mixing' matrix. We perform spectral analysis of the solutions and conclude that gradient flow graph convolutional models can induce a dynamics dominated by the graph high frequencies, which is desirable for heterophilic datasets. We also describe structural constraints on common GNN architectures allowing to interpret them as gradient flows. We perform thorough ablation studies corroborating our theoretical analysis and show competitive performance of simple and lightweight models on real-world homophilic and heterophilic datasets.

1 Introduction and motivations

Graph neural networks (GNNs) [38, 20, 21, 36, 7, 15, 27] and in particular their Message Passing formulation (MPNN) [19] have become the standard ML tool for dealing with different types of relations and interactions, ranging from social networks to particle physics and drug design. One of the often cited drawbacks of traditional GNN models is their poor 'explainability', making it hard to know why and how they make certain predictions [46, 47], and in which situations they may work and when they would fail. Limitations of GNNs that have attracted attention are over-smoothing [29, 30, 8], over-squashing and bottlenecks [1, 40], and performance on heterophilic data [31, 51, 13, 4, 45] – where adjacent nodes usually have different labels.

Contributions. We propose a *Gradient Flow Framework* (GRAFF) where the GNN equations follow the direction of steepest descent of a *learnable energy*. Thanks to this framework we can (i) interpret GNNs as a multi-particle dynamics where the learned parameters determine pairwise attractive and repulsive potentials in the feature space. This sheds light on how GNNs can adapt to heterophily and explains their performance and the smoothness of the prediction. (ii) GRAFF leads to residual convolutional models where the *channel-mixing* $\mathbf{W}$ is performed by a shared symmetric bilinear form inducing attraction and repulsion via its positive and negative eigenvalues, respectively. We theoretically investigate the interaction of the graph spectrum with the spectrum of the channel-mixing, proving that if there is more mass on the negative eigenvalues of $\mathbf{W}$, then the dynamics is dominated by the graph-high frequencies, which could be desirable on heterophilic graphs. We also extend results of [29, 30, 8] by showing that when we drop the residual connection intrinsic to the gradient flow framework,

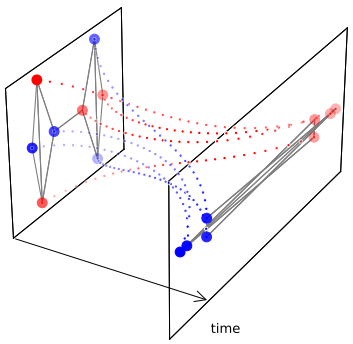

Figure 1: GRAFF dynamics: attractive and repulsive forces lead to a non-smoothing process able to separate labels.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

graph convolutional models always induce a low-frequency dominated dynamics *independent* of the sign and magnitude of the spectrum of the channel-mixing. We also discuss how simple choices make common architectures fit GRAFF and conduct thorough ablation studies to corroborate the theoretical analysis on the role of the spectrum of $\mathbf{W}$. (iii) We crystallize *an instance* of our framework into a linear, residual, convolutional model that achieves competitive performance on homophilic and heterophilic real world graphs whilst being faster than GCN.

Related work. Our analysis is related to studying GNNs as filters on the graph spectrum [15, 24, 2, 25] and over-smoothing [29, 30, 8, 50] and partly adopts techniques similar to [30]. The key difference is that we also consider the spectrum of the 'channel-mixing' matrix. The concept of gradient flows has been a standard tool in physics and geometry [16], from which they were adopted for image processing [26], and recently used in ML [35] for the analysis of Transformers [41] – see also [18] for discussion of loss landscapes. Our continuous-time evolution equations follows the spirit of Neural ODES [22, 12, 3] and the study of GNNs as continuous dynamical systems [44, 10, 17, 9].

Outline. In Section 2, we review the continuous and discrete Dirichlet energy and the associated gradient flow framework. We formalize the notion of over-smoothing and low(high)-frequency-dominated dynamics to investigate GNNs and study the dominant components in their evolution. We extend the graph Dirichlet energy to allow for a non-trivial norm for the feature edge-gradient. This leads to gradient flow equations that diffuse the features and over-smooth in the limit. Accordingly, in Section 3 we introduce a more general energy with a symmetric channel-mixing matrix $\mathbf{W}$ giving rise to attractive and repulsive pairwise terms via its positive and negative eigenvalues and show that the negative spectrum can induce high-frequency-dominant dynamics. In Section 4 we first compare with continuous GNN models and then discretize the equations and provide a 'recipe' for making standard GNN architectures fit a gradient flow framework. We adapt the spectral analysis to discrete-time showing that gradient flow convolutional models *can* generate a dynamics dominated by the high frequencies via the negative eigenvalues of $\mathbf{W}$ while this is impossible if we drop the residual connection. In Section 5 we corroborate our theoretical analysis on the role of the spectrum of $\mathbf{W}$ via ablation studies on graphs with varying homophily. Experiments on real world datasets show a competitive performance of our model despite its simplicity and reduced number of parameters.

2 Gradient-flow formalism

Notations adopted throughout the paper. Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be an *undirected* graph with n nodes. We denote by $\mathbf{F} \in \mathbb{R}^{n\times d}$ the matrix of d-dimensional node features, by $\mathbf{f}_i \in \mathbb{R}^d$ its i-th row (transposed), by $\mathbf{f}^r \in \mathbb{R}^n$ its r-th column, and by $\mathrm{vec}(\mathbf{F}) \in \mathbb{R}^{nd}$ the vectorization of $\mathbf{F}$ obtained by stacking its columns. Given a symmetric matrix $\mathbf{B}$, we let $\lambda_+^{\mathbf{B}}, \lambda_-^{\mathbf{B}}$ denote its most positive and negative eigenvalues, respectively, and $\rho_{\mathbf{B}}$ be its *spectral radius*. If $\mathbf{B} \succeq 0$, then $\mathrm{gap}(\mathbf{B})$ denotes the *positive smallest eigenvalue* of $\mathbf{B}$. $\dot{f}(t)$ denotes the temporal derivative, $\otimes$ is the Kronecker product and 'a.e.' means *almost every* w.r.t. Lebesgue measure and usually refers to data in the complement of some lower dimensional subspace in $\mathbb{R}^{n\times d}$. Proofs and additional results appear in the Appendix.

Starting point: a geometric parallelism. To motivate a gradient-flow approach for GNNs, we start from the continuous case (see Appendix A.1 for details). Consider a smooth map $f : \mathbb{R}^n \to (\mathbb{R}^d, h)$ with h a constant metric represented by $\mathbf{H} \succeq 0$. The *Dirichlet energy* of f is defined by

$$\mathcal{E}(f,h) = \frac{1}{2}\int_{\mathbb{R}^n} \|\nabla f\|_h^2 \, dx = \frac{1}{2}\sum_{q,r=1}^{d}\sum_{j=1}^{n}\int_{\mathbb{R}^n} h_{qr}\partial_j f^q \partial_j f^r(x)dx \tag{1}$$

and measures the 'smoothness' of f. A natural approach to find minimizers of $\mathcal{E}$ - called *harmonic maps* - was introduced in [16] and consists in studying the **gradient flow** of $\mathcal{E}$, wherein a given map $f(0) = f_0$ is evolved according to $\dot{f}(t) = -\nabla_f \mathcal{E}(f(t))$. These type of evolution equations have historically been the core of *variational* and *PDE-based image processing*; in particular, gradient flows of the Dirichlet energy were shown [26] to recover the Perona-Malik nonlinear diffusion [32].

Motivation: GNNs for node-classification. We wish to extend the gradient flow formalism to node classification on graphs. Assume we have a graph G, node-features $\mathbf{F}_0$ and labels $\{y_i\}$ on $\mathsf{V}_{\mathrm{train}} \subset \mathsf{V}$, and that we want to predict the labels on $\mathsf{V}_{\mathrm{test}} \subset \mathsf{V}$. A GNN typically evolves the features via some

parametric rule, $\mathrm{GNN}_\theta(\mathsf{G}, \mathbf{F}_0)$, and uses a decoding map for the prediction $y = \psi_{\mathrm{DE}}(\mathrm{GNN}_\theta(\mathsf{G}, \mathbf{F}_0))$. In graph convolutional models [15, 27], GNN_θ consists of two operations: applying a shared linear transformation to the features (**'channel mixing'**) and propagating them along the edges of the graph (**'diffusion'**). Our **goal** consists in studying when GNN_θ is the *gradient flow* of some parametric class of energies $\mathcal{E}_\theta : \mathbb{R}^{n\times d} \to \mathbb{R}$, which generalize the Dirichlet energy. This means that the parameters can be interpreted as 'finding the right notion of smoothness' for our task. We evolve the features by $\dot{\mathbf{F}}(t) = -\nabla_{\mathbf{F}}\mathcal{E}_\theta(\mathbf{F}(t))$ with prediction $y = \psi_{\mathrm{DE}}(\mathbf{F}(T))$ for some optimal time T.

Why a gradient flow? Since $\dot{\mathcal{E}}_\theta(\mathbf{F}(t)) = -||\nabla_{\mathbf{F}}\mathcal{E}_\theta(\mathbf{F}(t))||^2$, the energy dissipates along the gradient flow. Accordingly, this framework allows to *explain the GNN dynamics* as flowing the node features in the direction of steepest descent of $\mathcal{E}_\theta$. Indeed, we find that parametrizing an energy leads to equations governed by attractive and repulsive forces that can be controlled via the spectrum of symmetric 'channel-mixing' matrices. This shows that by learning to distribute more mass over the negative (positive) eigenvalues of the channel-mixing, gradient flow models can generate dynamics dominated by the higher (respectively, lower) graph frequencies and hence tackle different homophily scenarios. The gradient flow framework also leads to sharing of the weights across layers (since we parametrize the *energy* rather than the *evolution equations*, as usually done in GNNs), allowing us to reduce the number of parameters without compromising performance (see Table 1).

Analysis on graphs: preliminaries. Given a *connected* graph G with self-loops, its adjacency matrix $\mathbf{A}$ is defined as $a_{ij} = 1$ if $(i,j) \in \mathsf{E}$ and zero otherwise. We let $\mathbf{D} = \mathrm{diag}(d_i)$ be the degree matrix and write $\bar{\mathbf{A}} := \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. Let $\mathbf{F} \in \mathbb{R}^{n\times d}$ be the matrix representation of a signal. Its *graph gradient* is $(\nabla\mathbf{F})_{ij} := \mathbf{f}_j/\sqrt{d_j} - \mathbf{f}_i/\sqrt{d_i}$. We define the *Laplacian* as $\Delta := -\frac{1}{2}\mathrm{div}\,\nabla$ (the *divergence* div is the adjoint of ∇), represented by $\mathbf{\Delta} = \mathbf{I} - \bar{\mathbf{A}} \succeq 0$. We refer to the eigenvalues of $\mathbf{\Delta}$ as *frequencies*: the lowest frequency is always 0 while the highest frequency is $\rho_{\mathbf{\Delta}} \leq 2$ [14]. As for the continuum case, the gradient allows to define a *(graph) Dirichlet energy* as [49]

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) := \frac{1}{4}\sum_i \sum_{j:(i,j)\in\mathsf{E}} ||(\nabla\mathbf{F})_{ij}||^2 \equiv \frac{1}{4}\sum_{(i,j)\in\mathsf{E}} ||\frac{\mathbf{f}_i}{\sqrt{d_i}} - \frac{\mathbf{f}_j}{\sqrt{d_j}}||^2 = \frac{1}{2}\mathrm{trace}(\mathbf{F}^\top\mathbf{\Delta}\mathbf{F}), \qquad (2)$$

where the extra $\frac{1}{2}$ is for convenience. As for manifolds, $\mathcal{E}^{\mathrm{Dir}}$ measures smoothness. If we stack the columns of $\mathbf{F}$ into $\mathrm{vec}(\mathbf{F}) \in \mathbb{R}^{nd}$, the gradient flow of $\mathcal{E}^{\mathrm{Dir}}$ yields the *heat equation* on each channel:

$$\mathrm{vec}(\dot{\mathbf{F}}(t)) = -\nabla_{\mathrm{vec}(\mathbf{F})}\mathcal{E}^{\mathrm{Dir}}(\mathrm{vec}(\mathbf{F}(t))) = -(\mathbf{I}_d \otimes \mathbf{\Delta})\mathrm{vec}(\mathbf{F}(t)) \iff \dot{\mathbf{f}}^r(t) = -\mathbf{\Delta}\mathbf{f}^r(t), \quad (3)$$

for $1 \leq r \leq d$. Similarly to [8], we rely on $\mathcal{E}^{\mathrm{Dir}}$ to assess whether a given dynamics $t \mapsto \mathbf{F}(t)$ is a smoothing process. A different choice of Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ with non-normalized adjacency induces the analogous Dirichlet energy $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{L}}(\mathbf{F}) = \frac{1}{2}\mathrm{trace}(\mathbf{F}^\top\mathbf{L}\mathbf{F})$. Throughout this paper, we rely on the following definitions (see Appendix A.3 for further equivalent formulations and justifications):

Definition 2.1. $\dot{\mathbf{F}}(t) = \mathrm{GNN}_\theta(\mathbf{F}(t), t)$ initialized at $\mathbf{F}(0)$ is *smoothing* if $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \leq C + \varphi(t)$, with C a constant only depending on $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(0))$ and $\dot{\varphi}(t) \leq 0$. *Over-smoothing* occurs if either $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \to 0$ or $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{L}}(\mathbf{F}(t)) \to 0$ for $t \to \infty$.

Our notion of 'over-smoothing' is a relaxed version of the definition in [34] – although in the linear case one always finds an *exponential decay* of $\mathcal{E}^{\mathrm{Dir}}$. We note that $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \to 0$ iff $\mathbf{\Delta}\mathbf{f}^r(t) \to \mathbf{0}$ for each column $\mathbf{f}^r$. As in [30], this corresponds to a loss of separation power along the solution where nodes with *equal degree* become indistinguishable since we converge to $\ker(\mathbf{\Delta})$ (if we replaced $\mathbf{\Delta}$ with $\mathbf{L}$ then we would not even be able to separate nodes with different degrees in the limit).

To motivate the next definition, consider $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)$. Despite $||\mathbf{F}(t)||$ being unbounded for a.e. $\mathbf{F}(0)$, the low-frequency components are growing the fastest and indeed $\mathbf{F}(t)/||\mathbf{F}(t)|| \to \mathbf{F}_\infty$ s.t. $\mathbf{\Delta}\mathbf{f}^r_\infty = \mathbf{0}$ for $1 \leq r \leq d$. We formalize this scenario – including the opposite case of high-frequency components being dominant – by studying $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||)$, i.e. the Rayleigh quotient of $\mathbf{I}_d \otimes \mathbf{\Delta}$.

Definition 2.2. $\dot{\mathbf{F}}(t) = \mathrm{GNN}_\theta(\mathbf{F}(t), t)$ initialized at $\mathbf{F}(0)$ is *Low/High-Frequency-Dominant* (L/HFD) if $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \to 0$ (respectively, $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \to \rho_{\mathbf{\Delta}}/2$) for $t \to \infty$.

We report a consequence of Definition 2.2 and refer to Appendix A.3 for additional details and motivations for the characterizations of LFD and HFD.

Lemma 2.3. GNN_θ *is* LFD *(*HFD*) iff for each* $t_j \to \infty$ *there exist* $t_{j_k} \to \infty$ *and* $\mathbf{F}_\infty$ *s.t.* $\mathbf{F}(t_{j_k})/||\mathbf{F}(t_{j_k})|| \to \mathbf{F}_\infty$ *and* $\mathbf{\Delta}\mathbf{f}^r_\infty = \mathbf{0}$ *(* $\mathbf{\Delta}\mathbf{f}^r_\infty = \rho_{\mathbf{\Delta}}\mathbf{f}^r_\infty$*, respectively).*

If a graph is *homophilic*, adjacent nodes are likely to share the same label and we expect a smoothing or LFD dynamics enhancing the low-frequency components to be successful at node classification tasks [43, 28]. In the opposite case of *heterophily*, the high-frequency components might contain more relevant information for separating classes [4, 5] – the prototypical example being the eigenvector of $\mathbf{\Delta}$ associated with largest frequency $\rho_{\mathbf{\Delta}}$ separating a regular bipartite graph. In other words, the class of heterophilic graphs contain instances where signals should be *sharpened* by increasing $\mathcal{E}^{\mathrm{Dir}}$ rather than smoothed out. Accordingly, an ideal framework for learning on graphs must accommodate both of these opposite scenarios by being able to induce either an LFD or a HFD dynamics.

Parametric Dirichlet energy: channel-mixing as metric in feature space. In eq. (1) a constant nontrivial metric h in $\mathbb{R}^d$ leads to the mixing of the feature channels. We adapt this idea by considering a symmetric positive semi-definite $\mathbf{H} = \mathbf{W}^\top \mathbf{W}$ with $\mathbf{W} \in \mathbb{R}^{d\times d}$ and using it to generalize $\mathcal{E}^{\mathrm{Dir}}$ as

$$\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}) := \frac{1}{4}\sum_{q,r=1}^{d}\sum_{i}\sum_{j:(i,j)\in \mathsf{E}} h_{qr}(\nabla \mathbf{f}^q)_{ij}(\nabla \mathbf{f}^r)_{ij} = \frac{1}{4}\sum_{(i,j)\in\mathsf{E}} ||\mathbf{W}(\nabla \mathbf{F})_{ij}||^2. \tag{4}$$

We note the analogy with eq. (1), where the sum over the nodes replaces the integration over the domain and the j-th derivative at some point i is replaced by the gradient along the edge $(i, j) \in \mathsf{E}$. We generally treat $\mathbf{W}$ as *learnable weights* and study the gradient flow of $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}$:

$$\dot{\mathbf{F}}(t) = -\nabla_{\mathbf{F}}\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}(t)) = -\mathbf{\Delta}\mathbf{F}(t)\mathbf{W}^\top\mathbf{W}. \tag{5}$$

We see that eq. (5) generalizes eq. (3). Below 'smoothing' is intended as in Definition 2.1.

Proposition 2.4. *Let* $P^{\mathrm{ker}}_{\mathbf{W}}$ *be the projection onto* $\ker(\mathbf{W}^\top\mathbf{W})$. *Equation* (5) *is smoothing since*

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \leq e^{-2t\mathrm{gap}(\mathbf{W}^\top\mathbf{W})\mathrm{gap}(\mathbf{\Delta})}||\mathbf{F}(0)||^2 + \mathcal{E}^{\mathrm{Dir}}((P^{\mathrm{ker}}_{\mathbf{W}}\otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0))), \quad t \geq 0.$$

In fact $\mathbf{F}(t) \to \mathbf{F}_\infty$ *s.t.* $\exists\, \boldsymbol{\phi}_\infty \in \mathbb{R}^d$*: for each* $i \in \mathsf{V}$ *we have* $(\mathbf{f}_\infty)_i = \sqrt{d_i}\boldsymbol{\phi}_\infty + P^{\mathrm{ker}}_{\mathbf{W}}\mathbf{f}_i(0)$.

Proposition 2.4 implies that *no weight matrix* $\mathbf{W}$ in eq. (5) *can separate the limit embeddings* $\mathbf{F}(\infty)$ *of nodes with same degree and input features*. If $\mathbf{W}$ has a trivial kernel, then nodes with same degrees converge to the same representation and *over-smoothing* occurs as per Definition 2.1. Differently from [29, 30, 8], over-smoothing occurs independently of the spectral radius of the 'channel-mixing' if its eigenvalues are *positive* – even for equations which lead to residual GNNs when discretized [12]. According to Proposition 2.4, we do not expect eq. (5) to succeed on heterophilic graphs where *smoothing* processes are generally harmful – this is confirmed in Figure 2 (see *prod*-curve). To remedy this problem, we generalize eq. (5) to a gradient flow that can be HFD as per Definition 2.2.

3 A general parametric energy for pairwise interactions

We first rewrite the energy $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}$ in eq. (4) as

$$\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}) = \frac{1}{2}\sum_i \langle \mathbf{f}_i, \mathbf{W}^\top\mathbf{W}\mathbf{f}_i\rangle - \frac{1}{2}\sum_{i,j}\bar{a}_{ij}\langle \mathbf{f}_i, \mathbf{W}^\top\mathbf{W}\mathbf{f}_j\rangle. \tag{6}$$

We then define a *new, more general* energy by replacing the occurrences of $\mathbf{W}^\top\mathbf{W}$ with new symmetric matrices $\mathbf{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d}$ since we also want to generate repulsive forces:

$$\mathcal{E}^{\mathrm{tot}}(\mathbf{F}) := \frac{1}{2}\sum_i \langle \mathbf{f}_i, \mathbf{\Omega}\mathbf{f}_i\rangle - \frac{1}{2}\sum_{i,j}\bar{a}_{ij}\langle \mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle \equiv \mathcal{E}^{\mathrm{ext}}_{\mathbf{\Omega}}(\mathbf{F}) + \mathcal{E}^{\mathrm{pair}}_{\mathbf{W}}(\mathbf{F}), \tag{7}$$

with associated gradient flow of the form (see Appendix B)

$$\dot{\mathbf{F}}(t) = -\nabla_{\mathbf{F}}\mathcal{E}^{\mathrm{tot}}(\mathbf{F}(t)) = -\mathbf{F}(t)\mathbf{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}. \tag{8}$$

Note that eq. (8) *is gradient flow of some energy* $\mathbf{F} \mapsto \mathcal{E}^{\mathrm{tot}}(\mathbf{F})$ *iff both* $\mathbf{\Omega}$ *and* $\mathbf{W}$ *are symmetric*.

A multi-particle system point of view: attraction vs repulsion. Consider the d-dimensional node-features as particles in $\mathbb{R}^d$ with energy $\mathcal{E}^{\mathrm{tot}}$. While the term $\mathcal{E}^{\mathrm{ext}}_{\mathbf{\Omega}}$ is *independent of the graph topology* and represents an **external** field in the feature space, the second term $\mathcal{E}^{\mathrm{pair}}_{\mathbf{W}}$ constitutes a potential energy, with $\mathbf{W}$ a *bilinear form* determining the **pairwise interactions** of adjacent node

representations. Given a symmetric $\mathbf{W}$, we write $\mathbf{W} = \mathbf{\Theta}_+^\top \mathbf{\Theta}_+ - \mathbf{\Theta}_-^\top \mathbf{\Theta}_-$, by decomposing the spectrum of $\mathbf{W}$ in positive and negative values.We can rewrite $\mathcal{E}^{\text{tot}} = \mathcal{E}^{\text{ext}}_{\mathbf{\Omega}-\mathbf{W}} + \mathcal{E}^{\text{Dir}}_{\mathbf{\Theta}_+} - \mathcal{E}^{\text{Dir}}_{\mathbf{\Theta}_-}$, i.e.

$$\mathcal{E}^{\text{tot}}(\mathbf{F}) = \frac{1}{2}\sum_i \langle \mathbf{f}_i, (\mathbf{\Omega} - \mathbf{W})\mathbf{f}_i \rangle + \frac{1}{4}\sum_{i,j} ||\mathbf{\Theta}_+(\nabla \mathbf{F})_{ij}||^2 - \frac{1}{4}\sum_{i,j} ||\mathbf{\Theta}_-(\nabla \mathbf{F})_{ij}||^2. \tag{9}$$

The gradient flow of $\mathcal{E}^{\text{tot}}$ *minimizes* $\mathcal{E}^{\text{Dir}}_{\mathbf{\Theta}_+}$ and *maximizes* $\mathcal{E}^{\text{Dir}}_{\mathbf{\Theta}_-}$. The matrix $\mathbf{W}$ encodes *repulsive pairwise interactions* via its negative-definite component $\mathbf{\Theta}_-$ which lead to terms $||\mathbf{\Theta}_-(\nabla \mathbf{F})_{ij}||$ increasing along the solution. The latter affords a 'sharpening' effect desirable on heterophilic graphs where we need to disentangle adjacent node representations and hence 'magnify' the edge-gradient.

Spectral analysis of the channel-mixing. We will now show that eq. (8) can lead to a HFD dynamics. To this end, we assume that $\mathbf{\Omega} = \mathbf{0}$ so that eq. (8) becomes $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$. According to eq. (9) the negative eigenvalues of $\mathbf{W}$ lead to repulsion. We show that the latter can induce HFD dynamics as per Definition 2.2. We let $P^{\rho_-}_{\mathbf{W}}$ be the orthogonal projection into the eigenspace of $\mathbf{W} \otimes \bar{\mathbf{A}}$ associated with the eigenvalue $\rho_- := |\lambda^{\mathbf{W}}_-|(\rho_{\mathbf{\Delta}} - 1)$. We define ϵ_{HFD} explicitly in eq. (24).

Proposition 3.1. *If $\rho_- > \lambda^{\mathbf{W}}_+$, then $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ is* HFD *for a.e. $\mathbf{F}(0)$: there exists ϵ_{HFD} s.t.*

$$\mathcal{E}^{\text{Dir}}(\mathbf{F}(t)) = e^{2t\rho_-}\left(\frac{\rho_{\mathbf{\Delta}}}{2}||P^{\rho_-}_{\mathbf{W}}\mathbf{F}(0)||^2 + \mathcal{O}(e^{-2t\epsilon_{\text{HFD}}})\right), \quad t \geq 0,$$

and $\mathbf{F}(t)/||\mathbf{F}(t)||$ converges to $\mathbf{F}_\infty \in \mathbb{R}^{n\times d}$ such that $\mathbf{\Delta}\mathbf{f}^r_\infty = \rho_{\mathbf{\Delta}}\mathbf{f}^r_\infty$, for $1 \leq r \leq d$.

Proposition 3.1 shows that *if enough mass of the spectrum of the 'channel-mixing' is distributed over the negative eigenvalues, then the evolution is dominated by the graph high frequencies*. This analysis is made possible in our gradient flow framework where $\mathbf{W}$ must be *symmetric*. The HFD dynamics induced by negative eigenvalues of $\mathbf{W}$ is confirmed in Figure 2 (*neg-prod*-curve in the bottom chart).

A more general energy. Equations with a source term may have better expressive power [44, 11, 39]. In our framework this means adding an extra energy term of the form $\mathcal{E}^{\text{source}}_{\tilde{\mathbf{W}}}(\mathbf{F}) := \beta\langle \mathbf{F}, \mathbf{F}(0)\tilde{\mathbf{W}}\rangle$ to eq. (7) with some learnable β and $\tilde{\mathbf{W}}$. This leads to the following gradient flow:

$$\dot{\mathbf{F}}(t) = -\mathbf{F}(t)\mathbf{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} - \beta\mathbf{F}(0)\tilde{\mathbf{W}}. \tag{10}$$

We also observe that one could replace the fixed matrix $\bar{\mathbf{A}}$ with a more general *symmetric graph vector field* $\boldsymbol{\mathcal{A}}$ satisfying $\boldsymbol{\mathcal{A}}_{ij} = 0$ if $(i,j) \notin \mathsf{E}$, although in this work we focus on the case $\boldsymbol{\mathcal{A}} = \bar{\mathbf{A}}$. We also note that when $\mathbf{\Omega} = \mathbf{W}$, then eq. (8) becomes $\dot{\mathbf{F}}(t) = -\mathbf{\Delta}\mathbf{F}(t)\mathbf{W}$. We perform a spectral analysis of this case in Appendix B.2.

Non-linear activations. In Appendix B.3 we discuss non-linear gradient flow equations. Here we study what happens if the gradient flow in eq. (10) is activated *pointwise* by $\sigma : \mathbb{R} \to \mathbb{R}$. We show that although we are no longer a gradient flow, the learnable multi-particle energy $\mathcal{E}^{\text{tot}}$ is still decreasing along the solution, meaning that the interpretation of the channel-mixing $\mathbf{W}$ inducing attraction and repulsion via its positive and negative eigenvalues respectively **is preserved**.

Proposition 3.2. *Consider a non-linear map $\sigma : \mathbb{R} \to \mathbb{R}$ such that the function $x \mapsto x\sigma(x) \geq 0$. If $t \mapsto \mathbf{F}(t)$ solves the equation*

$$\dot{\mathbf{F}}(t) = \sigma\left(-\mathbf{F}(t)\mathbf{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} - \beta\mathbf{F}(0)\tilde{\mathbf{W}}\right),$$

where σ acts elementwise, then

$$\frac{d\mathcal{E}^{\text{tot}}(\mathbf{F}(t))}{dt} \leq 0.$$

A proof of this result and more details and discussion are reported in Appendix E. We emphasize here that differently from previous results about behaviour of ReLU wrt $\mathcal{E}^{\text{Dir}}$ [30, 8], we deal with a much more general energy that can also induce repulsion and a more general family of activation functions (that include ReLU, $\tanh$, $\arctan$ and many others).

4 Comparison with GNNs

In this Section, we study standard GNN models from the perspective of our gradient flow framework.

4.1 Continuous case

Continuous GNN models replace layers with continuous time. In contrast with Proposition 3.1, we show that three main *linearized* continuous GNN models are either *smoothing* or LFD as per Definition 2.2. The linearized PDE-GCN$_D$ model [17] corresponds to choosing $\beta = 0$ and $\boldsymbol{\Omega} = \mathbf{W} = \mathbf{K}(t)^\top \mathbf{K}(t)$ in eq. (10), for some time-dependent family $t \mapsto \mathbf{K}(t) \in \mathbb{R}^{d\times d}$:

$$\dot{\mathbf{F}}_{\mathrm{PDE-GCN_D}}(t) = -\boldsymbol{\Delta}\mathbf{F}(t)\mathbf{K}(t)^\top\mathbf{K}(t).$$

The CGNN model [44] can be derived from eq. (10) by setting $\boldsymbol{\Omega} = \mathbf{I} - \tilde{\boldsymbol{\Omega}}, \mathbf{W} = \tilde{\mathbf{W}} = \mathbf{I}, \beta = 1$:

$$\dot{\mathbf{F}}_{\mathrm{CGNN}}(t) = -\boldsymbol{\Delta}\mathbf{F}(t) + \mathbf{F}(t)\tilde{\boldsymbol{\Omega}} + \mathbf{F}(0).$$

Finally, in linearized GRAND [10] a row-stochastic matrix $\boldsymbol{\mathcal{A}}(\mathbf{F}(0))$ is *learned* from the encoding via an attention mechanism and we have

$$\dot{\mathbf{F}}_{\mathrm{GRAND}}(t) = -\boldsymbol{\Delta}_{\mathrm{RW}}\mathbf{F}(t) = -(\mathbf{I} - \boldsymbol{\mathcal{A}}(\mathbf{F}(0)))\mathbf{F}(t).$$

We note that if $\boldsymbol{\mathcal{A}}$ is not symmetric, then GRAND is *not* a gradient flow.

Proposition 4.1. $\mathrm{PDE-GCN}_D$, CGNN *and* GRAND *satisfy the following:*

(i) $\mathrm{PDE-GCN}_D$ *is a smoothing model:* $\dot{\mathcal{E}}^{\mathrm{Dir}}(\mathbf{F}_{\mathrm{PDE-GCN}_D}(t)) \leq 0$.

(ii) For a.e. $\mathbf{F}(0)$ *it holds:* CGNN *is never* HFD *and if we remove the source term, then* $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}_{\mathrm{CGNN}}(t)/||\mathbf{F}_{\mathrm{CGNN}}(t)||) \leq e^{-\mathrm{gap}(\boldsymbol{\Delta})t}$.

(iii) If G *is connected,* $\mathbf{F}_{\mathrm{GRAND}}(t) \to \boldsymbol{\mu}$ *as* $t \to \infty$*, with* $\boldsymbol{\mu}^r = \mathrm{mean}(\mathbf{f}^r(0))$, $1 \leq r \leq d$.

By (ii) the source-free CGNN-evolution is LFD *independent of* $\tilde{\boldsymbol{\Omega}}$. Moreover, by (iii), over-smoothing occurs for GRAND as per Definition 2.1. On the other hand, Proposition 3.1 shows that the negative eigenvalues of $\mathbf{W}$ can make the source-free gradient flow in eq. (8) HFD. Experiments in Section 5 confirm that the gradient flow model outperforms CGNN and GRAND on heterophilic graphs.

4.2 Discrete case

We now describe a discrete version of our gradient flow model and compare it to 'discrete' GNNs where discrete time steps correspond to different layers. In the spirit of [12], we use explicit Euler scheme with step size $\tau \leq 1$ to solve eq. (10) and set $\tilde{\mathbf{W}} = \mathbf{I}$. In the gradient flow framework we *parametrize the energy* rather than the actual equations, which leads to *symmetric* channel-mixing matrices $\boldsymbol{\Omega}, \mathbf{W} \in \mathbb{R}^{d\times d}$ that are *shared across the layers*. Since the matrices are square, an *encoding* block $\psi_{\mathrm{EN}} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times d}$ is used to process input features $\mathbf{F}_0 \in \mathbb{R}^{n\times p}$ and generally reduce the hidden dimension from p to d. Moreover, the iterations inherently lead to a residual architecture because of the explicit Euler discretization:

$$\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau\left(-\mathbf{F}(t)\boldsymbol{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} + \beta\mathbf{F}(0)\right), \quad \mathbf{F}(0) = \psi_{\mathrm{EN}}(\mathbf{F}_0), \tag{11}$$

with prediction $y = \psi_{\mathrm{DE}}(\mathbf{F}(T))$ produced by a *decoder* $\psi_{\mathrm{DE}} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times k}$, where k is the number of label classes and T *integration time* of the form $T = m\tau$, so that $m \in \mathbb{N}$ represents the number of *layers*. Although eq. (11) is linear, we can include non-linear activations in $\psi_{\mathrm{EN}}, \psi_{\mathrm{DE}}$ making the entire model generally non-linear. We emphasize two important points:

- Since the framework is residual, even if the message-passing is linear, this is *not equivalent* to collapsing the dynamics into a single layer with diffusion matrix $\bar{\mathbf{A}}^m$, with m the number of layers, see eq. (27) in the appendix where we derive the expansion of the solution.
- We could also activate the equations pointwise and maintain the physics interpretation thanks to Proposition 3.2 to gain greater expressive power. In the following though, we mainly stick to the linear discrete gradient flow unless otherwise stated.

Are discrete GNNs gradient flows? Given a (learned) symmetric graph vector field $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n\times n}$ satisfying $\boldsymbol{\mathcal{A}}_{ij} = 0$ if $(i,j) \notin \mathsf{E}$, consider a family of linear GNNs with shared weights of the form

$$\mathbf{F}(t+1) = \mathbf{F}(t)\boldsymbol{\Omega} + \boldsymbol{\mathcal{A}}\mathbf{F}(t)\mathbf{W} + \beta\mathbf{F}(0)\tilde{\mathbf{W}}, \quad 0 \leq t \leq T. \tag{12}$$

Symmetry is the key requirement to interpret GNNs in eq. (12) in a gradient flow framework.

Lemma 4.2. *Equation* (12) *is the unit step size discrete gradient flow of* $\mathcal{E}^{\text{ext}}_{\mathbf{I}-\mathbf{\Omega}} + \mathcal{E}^{\text{pair}}_{\boldsymbol{\mathcal{A}},\mathbf{W}} - \mathcal{E}^{\text{source}}_{\tilde{\mathbf{W}}}$, *with* $\mathcal{E}^{\text{pair}}_{\boldsymbol{\mathcal{A}},\mathbf{W}}$ *defined by replacing* $\bar{\mathbf{A}}$ *with* $\boldsymbol{\mathcal{A}}$ *in eq.* (7)*, iff* $\mathbf{\Omega}$ *and* $\mathbf{W}$ *are symmetric.*

Lemma 4.2 provides a recipe for making standard architectures into a gradient flow, with *symmetry* being the key requirement. When eq. (12) is a gradient flow, the underlying GNN dynamics is equivalent to minimizing a multi-particle energy by learning attractive and repulsive directions in feature space as discussed in Section 3. In Appendix C.2, we show how Lemma 4.2 covers linear versions of GCN [27, 43], GAT [42], GraphSAGE [23] and GCNII [11] to name a few.

Over-smoothing analysis in discrete setting. By Proposition 3.1 we know that the continuous version of eq. (11) can be HFD thanks to the negative eigenvalues of $\mathbf{W}$. The next result represents a discrete counterpart of Proposition 3.1 and shows that *residual, symmetrized graph convolutional models can be* HFD. Below $P^{\rho_-}_{\mathbf{W}}$ is the projection into the eigenspace associated with the eigenvalue $\rho_- := |\lambda^{\mathbf{W}}_-|(\rho_{\boldsymbol{\Delta}} - 1)$ and we report the explicit value of δ_{HFD} in eq. (28) in Appendix C.3. We let:

$$\lambda^{\mathbf{W}}_+(\rho_{\boldsymbol{\Delta}} - 1))^{-1} < |\lambda^{\mathbf{W}}_-| < 2(\tau(2 - \rho_{\boldsymbol{\Delta}}))^{-1}. \tag{13}$$

Theorem 4.3. *Given* $\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau\bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$*, with* $\mathbf{W}$ *symmetric, if eq.* (13) *holds then*

$$\mathcal{E}^{\text{Dir}}(\mathbf{F}(m\tau)) = (1+\tau\rho_-)^{2m}\left(\frac{\rho_{\boldsymbol{\Delta}}}{2}||P^{\rho_-}_{\mathbf{W}}\mathbf{F}(0)||^2 + \mathcal{O}\left(\left(\frac{1+\tau\delta_{\text{HFD}}}{1+\tau\rho_-}\right)^{2m}\right)\right), \quad \delta_{\text{HFD}} < \rho_-,$$

hence the dynamics is HFD *for a.e.* $\mathbf{F}(0)$ *and in fact* $\mathbf{F}(m\tau)/||\mathbf{F}(m\tau)|| \to \mathbf{F}_\infty$ *s.t.* $\boldsymbol{\Delta}\mathbf{f}^r_\infty = \rho_{\boldsymbol{\Delta}}\mathbf{f}^r_\infty$. *Conversely, if* G *is not bipartite, then for a.e.* $\mathbf{F}(0)$ *the system* $\mathbf{F}(t+\tau) = \tau\bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$*, with* $\mathbf{W}$ *symmetric, is* LFD *independent of the spectrum of* $\mathbf{W}$.

Theorem 4.3 shows that linear discrete gradient flows can be HFD due to the negative eigenvalues of $\mathbf{W}$. This differs from statements that standard GCNs act as low-pass filters and thus over-smooth in the limit. Indeed, in these cases the spectrum of $\mathbf{W}$ is generally ignored [43, 11] or required to be sufficiently small in terms of singular value decomposition [29, 30, 8] *when no residual connection is present*. On the other hand, Theorem 4.3 emphasizes that the spectrum of $\mathbf{W}$ plays a key role to enhance the high frequencies when enough mass is distributed over the negative eigenvalues provided that a residual connection exists – this is confirmed by the *neg-prod*-curve in Figure 2.

The residual connection from a spectral perspective. Given a sufficiently small step-size so that the right hand side of inequality 13 is satisfied, $\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau\bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ is HFD for a.e. $\mathbf{F}(0)$ if $|\lambda^{\mathbf{W}}_-|(\rho_{\boldsymbol{\Delta}} - 1) > \lambda^{\mathbf{W}}_+$, i.e. 'there is more mass' in the negative spectrum of $\mathbf{W}$ than in the positive one. This means that differently from [29, 30, 8], there is no requirement on the minimal magnitude of the spectral radius of $\mathbf{W}$ coming from the graph topology as long as $\lambda^{\mathbf{W}}_+$ is small enough. Conversely, without a residual term, the dynamics is LFD for a.e. $\mathbf{F}(0)$ *independently* of the sign and magnitude of the eigenvalues of $\mathbf{W}$. This is also confirmed by the GCN-curve in Figure 2.

Over-smoothing vs LFD**.** We highlight how in general a linear GCN equation as $\mathbf{F}(t+\tau) = \tau\bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$ may avoid over-smoothing in the sense of Definition 2.1, meaning that $\mathcal{E}^{\text{Dir}}(\mathbf{F}(t)) \to \infty$ as soon as there exist $\lambda^{\boldsymbol{\Delta}}_i \in (0,1)$ and the spectral radius of $\mathbf{W}$ is large enough. However, this will not lead to over-separation since the dominating term is the lowest frequency one: in other words, once we re-set the scale right as per the normalization in Theorem 4.3, we encounter loss of separability even with large (and possibly negative) spectrum of $\mathbf{W}$.

5 Experiments

In this section we evaluate the gradient flow framework (GRAFF). We corroborate the spectral analysis using synthetic data with controllable homophily. We confirm that having negative (positive) eigenvalues of the channel-mixing $\mathbf{W}$ are essential in heterophilic (homophilic) scenarios where the gradient flow should align with HFD (LFD) respectively. We show that the gradient flow in eq. (11) – a linear, residual, symmetric graph convolutional model – achieves competitive performance on heterophilic datasets.

Methodology. We crystallize GRAFF in the model presented in eq. (11) with $\psi_{\text{EN}}, \psi_{\text{DE}}$ implemented as single linear layers or MLPs, and we set $\mathbf{\Omega}$ to be diagonal. For the real-world experiments we consider *diagonally-dominant* (DD), *diagonal* (D) and *time-dependent* choices for the structure of $\mathbf{W}$ that offer explicit control over its spectrum. In the (DD)-case, we consider a $\mathbf{W}^0 \in \mathbb{R}^{d\times d}$ *symmetric* with zero diagonal and $\mathbf{w} \in \mathbb{R}^d$ defined by $\mathbf{w}_\alpha = q_\alpha \sum_\beta |\mathbf{W}^0_{\alpha\beta}| + r_\alpha$, and set $\mathbf{W} = \text{diag}(\mathbf{w}) + \mathbf{W}^0$. Due to the Gershgorin Theorem the eigenvalues of $\mathbf{W}$ belong to $[\mathbf{w}_\alpha - \sum_\beta |\mathbf{W}^0_{\alpha\beta}|, \mathbf{w}_\alpha + \sum_\beta |\mathbf{W}^0_{\alpha\beta}|]$, so the model 'can' easily re-distribute mass in the spectrum of $\mathbf{W}$ via q_α, r_α. This generalizes the decomposition of $\mathbf{W}$ in [11] providing a justification in terms of its spectrum and turns out to be more efficient w.r.t. the hidden dimension d as shown in Figure 4 in the Appendix. For (D) we take $\mathbf{W}$ to be diagonal, with entries sampled $\mathcal{U}[-1, 1]$ and fixed – i.e., **we do not train** over $\mathbf{W}$ – and only learn $\psi_{\text{EN}}, \psi_{\text{DE}}$. We also include a *time-dependent* model where $\mathbf{W}_t$ varies across layers. To investigate the role of the spectrum of $\mathbf{W}$ on synthetic graphs, we construct three additional variants: $\mathbf{W} = \mathbf{W}' + \mathbf{W}'^\top$, $\mathbf{W} = \pm\mathbf{W}'^\top\mathbf{W}'$ named *sum*, *prod* and *neg-prod* respectively where *prod* (*neg-prod*) variants have only non-negative (non-positive) eigenvalues.

Complexity and number of parameters. If we treat the number of layers as a constant, the discrete gradient flow scales as $\mathcal{O}(|\mathsf{V}|pd + |\mathsf{E}|d^2)$, where p and d are input feature and hidden dimension respectively, with $p \geq d$ usually. Note that GCN has complexity $\mathcal{O}(|\mathsf{E}|pd)$ and in fact *our model is faster than GCN* as confirmed in Figure 5 in Appendix D. Since $\psi_{\text{EN}}, \psi_{\text{DE}}$ are single linear layers (MLPs), we can bound the number of parameters by $pd + d^2 + 3d + dk$, with k the number of label classes, in the (DD)-variant while in the (D)-variant we have $pd + 3d + dk$. Further ablation studies appear in Figure 4 in the Appendix showing that (DD) outperforms *sum* and GCN – especially in the lower hidden dimension regime – on real-world benchmarks with varying homophily.

Synthetic experiments and ablation studies. To investigate our claims in a controlled environment we use the synthetic Cora dataset of [51, Appendix G]. Graphs are generated for target levels of homophily via preferential attachment – see Appendix D.3 for details. Figure 2 confirms the spectral analysis and offers a better understanding in terms of performance and smoothness of the predictions. Each curve – except GCN – represents one version of $\mathbf{W}$ as in 'methodology' and we implement eq. (11) with $\beta = 0$, $\mathbf{\Omega} = \mathbf{0}$. Figure 2 (top) reports the test accuracy vs true label homophily. *Neg-prod* is better than *prod* on low-homophily and viceversa on high-homophily. This confirms Proposition 3.1 where we have shown that the gradient flow can lead to a HFD dynamics – that are generally desirable with low-homophily – through the negative eigenvalues of $\mathbf{W}$. Conversely, the *prod* configuration (where we have an attraction-only dynamics) struggles in low-homophily scenarios *even though a residual connection is present*. Both *prod* and *neg-prod* are 'extreme' choices and serve the purpose of highlighting that by turning off one side of the spectrum this could be the more damaging depending on the underlying homophily. In general though 'neutral' variants like *sum* and (DD) are indeed more flexible and better performing. In fact, (DD) outperforms GCN especially in low-homophily scenarios, confirming Theorem 4.3 where we have shown that without a residual connection convolutional models are LFD – and hence more sensitive to underlying homophily – irrespectively of the spectrum of $\mathbf{W}$. This is further confirmed in Figure 3.

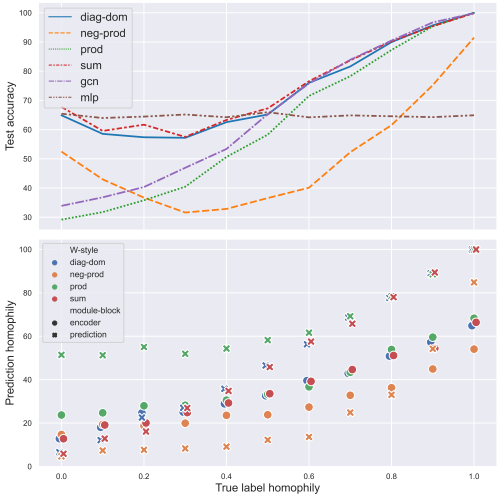

Figure 2: Experiments on synthetic datasets with controlled homophily.

In Figure 2 (bottom) we compute the homophily of the prediction (cross) for a given method and we compare with the homophily (circle) of the prediction read from the encoding (i.e. *graph-agnostic*). The homophily here is a proxy to assess whether the evolution is *smoothing*, the goal being explaining the smoothness of the prediction via the spectrum of $\mathbf{W}$ as per our theoretical analysis. For *neg-prod* the homophily after the evolution is lower than that of the encoding, supporting the analysis that negative eigenvalues of $\mathbf{W}$ enhance high-frequencies. The opposite behaviour occurs in the case of *prod* and explains that in the low-homophily regime *prod* is under-performant due to the prediction

	Texas	**Wisconsin**	**Cornell**	**Film**	**Squirrel**	**Chameleon**	**Citeseer**	**Pubmed**	**Cora**
Hom level	**0.11**	**0.21**	**0.30**	**0.22**	**0.22**	**0.23**	**0.74**	**0.80**	**0.81**
#Nodes	183	251	183	7,600	5,201	2,277	3,327	18,717	2,708
#Edges	295	466	280	26,752	198,493	31,421	4,676	44,327	5,278
#Classes	5	5	5	5	5	5	7	3	6
GGCN	84.86 ± 4.55	86.86 ± 3.29	85.68 ± 6.63	37.54 ± 1.56	55.17 ± 1.58	71.14 ± 1.84	77.14 ± 1.45	89.15 ± 0.37	87.95 ± 1.05
GPRGNN	78.38 ± 4.36	82.94 ± 4.21	80.27 ± 8.11	34.63 ± 1.22	31.61 ± 1.24	46.58 ± 1.71	77.13 ± 1.67	87.54 ± 0.38	87.95 ± 1.18
H2GCN	84.86 ± 7.23	87.65 ± 4.98	82.70 ± 5.28	35.70 ± 1.00	36.48 ± 1.86	60.11 ± 2.15	77.11 ± 1.57	89.49 ± 0.38	87.87 ± 1.20
GCNII	77.57 ± 3.83	80.39 ± 3.40	77.86 ± 3.79	37.44 ± 1.30	38.47 ± 1.58	63.86 ± 3.04	77.33 ± 1.48	90.15 ± 0.43	88.37 ± 1.25
Geom-GCN	66.76 ± 2.72	64.51 ± 3.66	60.54 ± 3.67	31.59 ± 1.15	38.15 ± 0.92	60.00 ± 2.81	78.02 ± 1.15	89.95 ± 0.47	85.35 ± 1.57
PairNorm	60.27 ± 4.34	48.43 ± 6.14	58.92 ± 3.15	27.40 ± 1.24	50.44 ± 2.04	62.74 ± 2.82	73.59 ± 1.47	87.53 ± 0.44	85.79 ± 1.01
GraphSAGE	82.43 ± 6.14	81.18 ± 5.56	75.95 ± 5.01	34.23 ± 0.99	41.61 ± 0.74	58.73 ± 1.68	76.04 ± 1.30	88.45 ± 0.50	86.90 ± 1.04
GCN	55.14 ± 5.16	51.76 ± 3.06	60.54 ± 5.30	27.32 ± 1.10	53.43 ± 2.01	64.82 ± 2.24	76.50 ± 1.36	88.42 ± 0.50	86.98 ± 1.27
GAT	52.16 ± 6.63	49.41 ± 4.09	61.89 ± 5.05	27.44 ± 0.89	40.72 ± 1.55	60.26 ± 2.50	76.55 ± 1.23	87.30 ± 1.10	86.33 ± 0.48
MLP	80.81 ± 4.75	85.29 ± 3.31	81.89 ± 6.40	36.53 ± 0.70	28.77 ± 1.56	46.21 ± 2.99	74.02 ± 1.90	75.69 ± 2.00	87.16 ± 0.37
CGNN	71.35 ± 4.05	74.31 ± 7.26	66.22 ± 7.69	35.95 ± 0.86	29.24 ± 1.09	46.89 ± 1.66	76.91 ± 1.81	87.70 ± 0.49	87.10 ± 1.35
GRAND	75.68 ± 7.25	79.41 ± 3.64	82.16 ± 7.09	35.62 ± 1.01	40.05 ± 1.50	54.67 ± 2.54	76.46 ± 1.77	89.02 ± 0.51	87.36 ± 0.96
Sheaf (max)	85.95 ± 5.51	89.41 ± 4.74	84.86 ± 4.71	37.81 ± 1.15	56.34 ± 1.32	68.04 ± 1.58	76.70 ± 1.57	89.49 ± 0.40	86.90 ± 1.13
GRAFF (DD)	88.38 ± 4.53	87.45 ± 2.94	83.24 ± 6.49	36.09 ± 0.81	54.52 ± 1.37	71.08 ± 1.75	76.92 ± 1.70	88.95 ± 0.52	87.61 ± 0.97
GRAFF (D)	88.11 ± 5.57	88.83 ± 3.29	84.05 ± 6.10	37.11 ± 1.08	47.36 ± 1.89	66.78 ± 1.28	77.30 ± 1.85	90.04 ± 0.41	88.01 ± 1.03
GRAFF-timedep (DD)	87.03 ± 4.49	87.06 ± 4.04	82.16 ± 7.07	35.93 ± 1.23	53.97 ± 1.45	69.56 ± 1.20	76.59 ± 1.53	88.26 ± 0.41	87.38 ± 1.05

Table 1: Results on heterophilic and homophilic datasets

being smoother than the true homophily. (DD) and *sum* variants adapt better to the true homophily. We note how the encoding compensates when the dynamics can only either attract or repulse (i.e. the spectrum of $\mathbf{W}$ has a sign) by decreasing or increasing the initial homophily respectively.

Real world experiments. We test GRAFF against a range of datasets with varying homophily [37, 33, 31] (see Appendix D.4 for additional details). We use results provided in [45, Table 1], which includes standard baselines as GCN [27], GraphSAGE [23], GAT [42], PairNorm [48] and recent models tailored towards the heterophilic setting (GGCN [45], Geom-GCN [31], H2GCN [51] and GPRGNN [13]). For Sheaf [5], a recent top-performer on heterophilic datasets, we took the best performing variant (out of six provided) for each dataset. We also include continuous baselines CGNN [44] and GRAND [10] to provide empirical evidence for Proposition 4.1. Splits taken from [31] are used in all the comparisons. The GRAFF model discussed in 'methodology' is a very simple architecture with shared parameters across layers and run-time smaller than GCN and more recent models like GGCN designed for heterophilic graphs (see Figure 5 in the Appendix). Nevertheless, it achieves competitive results on all datasets, performing on par or better than more complex recent models. Moreover, comparison with the 'time-dependent' (DD) variant confirms that by sharing weights across layers we do not lose performance. We note that on heterophilic graphs short integration time is usually needed due to the topology being harmful and the negative eigenvalues of $\mathbf{W}$ leading to exponential behaviour (see Appendix D).

6 Conclusions

In this work, we developed a framework for GNNs where the evolution can be interpreted as minimizing a multi-particle learnable energy. This translates into studying the interaction between the spectrum of the graph and the spectrum of the 'channel-mixing' leading to a better understanding of when and why the induced dynamics is low (high) frequency dominated. From a theoretical perspective, we refined existing asymptotic analysis of GNNs to account for the role of the spectrum of the channel-mixing as well. From a practical perspective, our framework allows for 'educated' choices resulting in a simple convolutional model that achieves competitive performance on homophilic and heterophilic benchmarks while being faster than GCN. Our results refute the folklore of graph convolutional models being too simple for heterophilic benchmarks.

Limitations and future works. We limited our attention to a *constant* bilinear form $\mathbf{W}$, which might be excessively rigid. It is possible to derive non-constant alternatives that are *aware* of the features or the position in the graph. The main challenge amounts to matching the requirement for local 'heterogeneity' with efficiency: we reserve this question for future work. Our analysis is also a first step into studying the interaction of the graph and 'channel-mixing' spectra; we did not explore other dynamics that are neither LFD nor HFD as per our definitions. The energy formulation points to new models more 'physics' inspired; this will be explored in future work.

Societal impact. Our work sheds light on the actual dynamics of GNNs and could hence improve their understanding, which is crucial for assessing their impact on large-scale applications. We also show that instances of our framework achieve competitive performance on heterophilic data despite being faster than GCN, providing evidence for efficient methods with reduced footprint.

References

[1] U. Alon and E. Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021.

[2] M. Balcilar, G. Renton, P. Héroux, B. Gaüzère, S. Adam, and P. Honeine. Analyzing the expressive power of graph neural networks in a spectral perspective. In *International Conference on Learning Representations*, 2020.

[3] M. Biloš, J. Sommer, S. S. Rangapuram, T. Januschowski, and S. Günnemann. Neural flows: Efficient alternative to neural odes. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[4] D. Bo, X. Wang, C. Shi, and H. Shen. Beyond low-frequency information in graph convolutional networks. *In AAAI. AAAI Press*, 2021.

[5] C. Bodnar, F. Di Giovanni, B. P. Chamberlain, P. Liò, and M. M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. *arXiv preprint arXiv:2202.04579*, 2022.

[6] S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[7] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[8] C. Cai and Y. Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020.

[9] B. Chamberlain, J. Rowbottom, D. Eynard, F. Di Giovanni, X. Dong, and M. Bronstein. Beltrami flow and neural diffusion on graphs. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] B. Chamberlain, J. Rowbottom, M. I. Gorinova, M. Bronstein, S. Webb, and E. Rossi. Grand: Graph neural diffusion. In *International Conference on Machine Learning*, pages 1407–1418. PMLR, 2021.

[11] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, pages 1725–1735. PMLR, 2020.

[12] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[13] E. Chien, J. Peng, P. Li, and O. Milenkovic. Adaptive universal generalized pagerank graph neural network. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[14] F. R. Chung and F. C. Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

[15] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[16] J. Eells and J. H. Sampson. Harmonic mappings of riemannian manifolds. *American journal of mathematics*, 86(1):109–160, 1964.

[17] M. Eliasof, E. Haber, and E. Treister. Pde-gcn: Novel architectures for graph neural networks motivated by partial differential equations. *Advances in Neural Information Processing Systems*, 34, 2021.

[18] M. Geiger, L. Petrini, and M. Wyart. Landscape and training regimes in deep learning. *Physics Reports*, 924:1–18, 2021.

[19] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017.

[20] C. Goller and A. Kuchler. Learning task-dependent distributed representations by backpropagation through structure. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, pages 347–352. IEEE, 1996.

[21] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

[22] E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34, 2018.

[23] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[24] D. K. Hammond, P. Vandergheynst, and R. Gribonval. The spectral graph wavelet transform: Fundamental theory and fast computation. In *Vertex-Frequency Analysis of Graph Signals*, pages 141–175. Springer, 2019.

[25] M. He, Z. Wei, H. Xu, et al. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation. *Advances in Neural Information Processing Systems*, 34, 2021.

[26] R. Kimmel, N. Sochen, and R. Malladi. From high energy physics to low level vision. In *International Conference on Scale-Space Theories in Computer Vision*, pages 236–247. Springer, 1997.

[27] T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, 2017.

[28] J. Klicpera, S. Weißenberger, and S. Günnemann. Diffusion improves graph learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[29] H. Nt and T. Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*, 2019.

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

[31] H. Pei, B. Wei, K. C. Chang, Y. Lei, and B. Yang. Geom-gcn: Geometric graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[32] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *PAMI*, 12(7):629–639, 1990.

[33] B. Rozemberczki, C. Allen, and R. Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021.

[34] T. K. Rusch, B. P. Chamberlain, J. Rowbottom, S. Mishra, and M. M. Bronstein. Graph-coupled oscillator networks. In *International Conference on Machine Learning*, 2022.

[35] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.

[36] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[37] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

[38] A. Sperduti. Encoding labeled graphs by labeling raam. *Advances in Neural Information Processing Systems*, 6, 1993.

[39] M. Thorpe, T. M. Nguyen, H. Xia, T. Strohmer, A. Bertozzi, S. Osher, and B. Wang. Grand++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2021.

[40] J. Topping, F. Di Giovanni, B. P. Chamberlain, X. Dong, and M. M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. *International Conference on Learning Representations*, 2022.

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

[42] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[43] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR, 2019.

[44] L.-P. Xhonneux, M. Qu, and J. Tang. Continuous graph neural networks. In *International Conference on Machine Learning*, pages 10432–10441. PMLR, 2020.

[45] Y. Yan, M. Hashemi, K. Swersky, Y. Yang, and D. Koutra. Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks. *arXiv preprint arXiv:2102.06462*, 2021.

[46] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[47] H. Yuan, H. Yu, S. Gui, and S. Ji. Explainability in graph neural networks: A taxonomic survey. *arXiv preprint arXiv:2012.15445*, 2020.

[48] L. Zhao and L. Akoglu. Pairnorm: Tackling oversmoothing in gnns. *arXiv preprint arXiv:1909.12223*, 2019.

[49] D. Zhou and B. Schölkopf. Regularization on discrete spaces. In *Joint Pattern Recognition Symposium*, pages 361–368. Springer, 2005.

[50] K. Zhou, X. Huang, D. Zha, R. Chen, L. Li, S.-H. Choi, and X. Hu. Dirichlet energy constrained learning for deep graph neural networks. *Advances in Neural Information Processing Systems*, 34:21834–21846, 2021.

[51] J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, and D. Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in Neural Information Processing Systems*, 33:7793–7804, 2020.

[52] O. Shchur, M. Mumme, A. Bojchevski, and S. Günnemann. Pitfalls of graph neural network evaluation. In NIPS workshop, 2018.

[53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019.

[54] M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[55] L. Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] , in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] in the **Societal impact** paragraph in Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] in Appendix A, Appendix B and Appendix C.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code and README in SM, dataloaders in code
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Splits and hyperparameters provided in code zip
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Standard deviations are stated in results table
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] in appendix D
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] datasets and standard libraries cited in appendix D
   (b) Did you mention the license of the assets? [Yes] industry standard libraries and benchmark datasets were used in accordance with licences
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] code provided in SM zip
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] no personal data is contained within benchmarking datasets
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs and additional details of Section 2

A.1 Discussion on continuous Dirichlet energy and harmonic maps

In this subsection we briefly expand on the formulation of continuous Dirichlet energy in Section 2 to provide more context. Consider a smooth map $f : (M, g) \to (N, h)$, where N is usually a larger manifold we embed M into, and g, h are Riemannian metrics on domain and codomain respectively. The *Dirichlet energy* of f is defined by

$$\mathcal{E}(f, g, h) := \frac{1}{2}\int_M |df|_g^2 d\mu(g),$$

with $|df|_g$ the norm of the Jacobian of f measured with respect to g and h. If (M, g) is standard Euclidean space $\mathbb{R}^n$, $N = \mathbb{R}^d$ and h is a constant positive semi-definite matrix, then we can rewrite the Dirichlet energy in a more familiar form as

$$\mathcal{E}(f, h) = \frac{1}{2}\int_{\mathbb{R}^n} \text{trace}\left(Df^\top h Df\right) d\mu = \frac{1}{2}\sum_{q,r=1}^{d}\sum_{j=1}^{n}\int_{\mathbb{R}^n} h_{qr}\partial_j f^q \partial_j f^r(x)dx.$$

The Dirichlet energy measures the smoothness of the map f, and indeed if h is the identity in $\mathbb{R}^d$, then we recover the classical definition

$$\mathcal{E}(f) = \frac{1}{2}\sum_{r=1}^{d}\int_{\mathbb{R}^n} ||\nabla f^r||^2(x)dx.$$

Gradient flow of Dirichlet energy. Minimizers of $\mathcal{E}$ - referred to as *harmonic maps* - are important objects in geometry: to mention a few, geodesics, minimal isometric immersions and maps $f : M \to \mathbb{R}^d$ solving $\Delta_g f = 0$ are all instances of harmonic maps. To identify such critical points, one computes the first variation of the energy $\mathcal{E}$ along an arbitrary direction $\partial_t f$, which can be written as

$$d\mathcal{E}_f(\partial_t f) = -\int_M \langle \tau_g(f), \partial_t f\rangle_h d\mu(g).$$

for some tensor field τ with explicit form

$$(\tau_{g_M}(f))^\alpha := \Delta_{g_M} f^\alpha + {}^{h_N}\Gamma^\alpha_{\beta\gamma}\partial_i f^\beta \partial_j f^\gamma g_M^{ij},$$

for $1 \leq \alpha \leq \dim(N)$, with $\{y^\alpha\}$ local coordinates on N and $\Gamma^\alpha_{\beta\gamma}$ Christoffel symbols. It follows that harmonic maps are identified by the condition $\tau_g(f)) = 0$. In [16], the pivotal idea of harmonic map flow – which has shaped much of modern research in geometric analysis – was introduced for the first time: in order to identify minimizers of $\mathcal{E}$, an input map f_0 is evolved along the direction of (minus) the gradient of the energy $\mathcal{E}$ leading to the dynamics

$$\partial_t f = \tau_g(f). \tag{14}$$

As a special case, when the target space is the classical Euclidean space one recovers the *heat equation* induced by the input Riemannian structure. We also note that when (M, g) is a surface representing an image and $f : (u_1, u_2) \mapsto (u_1, u_2, \phi(u_1, u_2))$ with ϕ a color map, then eq. (14) becomes

$$\partial_t \phi = \text{div}(C_g \nabla \phi), \tag{15}$$

with C_g a constant depending on the metric on M. If we now let g to depend on ϕ, one can recover the celebrated Perona-Malik flow [26].

A.2 Review of Kronecker product and properties of Laplacian kernel

Additional notations and conventions used throughout the appendix. Any graph G is taken to be *connected*. We order the eigenvalues of the graph Laplacian as $0 = \lambda_0^{\mathbf{\Delta}} \leq \lambda_1^{\mathbf{\Delta}} \leq \ldots \leq \lambda_{n-1}^{\mathbf{\Delta}} = \rho_{\mathbf{\Delta}} \leq 2$ with associated orthonormal basis of eigenvectors $\{\boldsymbol{\phi}_i^{\mathbf{\Delta}}\}_{i=0}^{n-1}$ so that in particular we have $\mathbf{\Delta}\boldsymbol{\phi}_0^{\mathbf{\Delta}} = \mathbf{0}$. Moreover, given a symmetric matrix $\mathbf{B}$, we generally denote the spectrum of $\mathbf{B}$ by $\text{spec}(\mathbf{B})$. Finally, if we write $\mathbf{F}(t)/||\mathbf{F}(t)||$ we always take the norm to be the Frobenius one and tacitly assume that the dynamics is s.t. the solution is not zero.

Kronecker product. In this subsection we summarize a few relevant notions pertaining the Kronecker product of matrices that are going to be applied throughout our spectral analysis of gradient flow equations for GNNs in both the continuous and discrete time setting.

Given a matricial equation of the form

$$\mathbf{Y} = \mathbf{A}\mathbf{X}\mathbf{B},$$

we can vectorize $\mathbf{X}$ and $\mathbf{Y}$ by stacking columns into $\mathrm{vec}(\mathbf{X})$ and $\mathrm{vec}(\mathbf{Y})$ respectively, and rewrite the previous system as

$$\mathrm{vec}(\mathbf{Y}) = \left(\mathbf{B}^\top \otimes \mathbf{A}\right) \mathrm{vec}(\mathbf{X}). \tag{16}$$

If $\mathbf{A}$ and $\mathbf{B}$ are symmetric with spectra $\mathrm{spec}(\mathbf{A})$ and $\mathrm{spec}(\mathbf{B})$ respectively, then the spectrum of $\mathbf{B} \otimes \mathbf{A}$ is given by $\mathrm{spec}(\mathbf{A}) \cdot \mathrm{spec}(\mathbf{B})$. Namely, if $\mathbf{A}\mathbf{x} = \lambda^{\mathbf{A}}\mathbf{x}$ and $\mathbf{B}\mathbf{y} = \lambda^{\mathbf{B}}\mathbf{y}$, for $\mathbf{x}$ and $\mathbf{y}$ non-zero vectors, then $\lambda^{\mathbf{B}}\lambda^{\mathbf{A}}$ is an eigenvalue of $\mathbf{B} \otimes \mathbf{A}$ with eigenvector $\mathbf{y} \otimes \mathbf{x}$:

$$(\mathbf{B} \otimes \mathbf{A})\, \mathbf{y} \otimes \mathbf{x} = (\lambda^{\mathbf{B}}\lambda^{\mathbf{A}})\mathbf{y} \otimes \mathbf{x}. \tag{17}$$

One can also define the *Kronecker sum* of matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{d\times d}$ as

$$\mathbf{A} \oplus \mathbf{B} := \mathbf{A} \otimes \mathbf{I}_d + \mathbf{I}_n \otimes \mathbf{B}, \tag{18}$$

with spectrum $\mathrm{spec}(\mathbf{A} \oplus \mathbf{B}) = \{\lambda^{\mathbf{A}} + \lambda^{\mathbf{B}} :\ \lambda^{\mathbf{A}} \in \mathrm{spec}(\mathbf{A}),\ \lambda^{\mathbf{B}} \in \mathrm{spec}(\mathbf{B})\}$.

Additional details on $\mathcal{E}^{\mathrm{Dir}}$ and the choice of Laplacian. We recall that the classical graph Dirichlet energy $\mathcal{E}^{\mathrm{Dir}}$ is defined by

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) = \frac{1}{2}\mathrm{trace}\left(\mathbf{F}^\top \boldsymbol{\Delta}\mathbf{F}\right),$$

where the (unusual) extra factor of $\frac{1}{2}$ is to avoid rescaling the gradient flow by 2 – which is the more common convention. We can use the Kronecker product to rewrite the Dirichlet energy as

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) = \frac{1}{2}\,\mathrm{vec}(\mathbf{F})^\top (\mathbf{I}_d \otimes \boldsymbol{\Delta})\mathrm{vec}(\mathbf{F}), \tag{19}$$

from which we immediately derive that $\nabla_{\mathrm{vec}(\mathbf{F})}\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}) = (\mathbf{I}_d \otimes \boldsymbol{\Delta})\mathrm{vec}(\mathbf{F})$ – since $\boldsymbol{\Delta}$ is *symmetric* – and hence recover the gradient flow in eq. (3) leading to the graph heat equation across each channel.

Before we further comment on the characterizations of LFD and HFD dynamics, we review the main choices of graph Laplacian and the associated harmonic signals (i.e. how we can characterize the kernel spaces of the given Laplacian operator). Recall that throughout the appendix we always assume that the underlying graph G is *connected*. The symmetrically normalized Laplacian $\boldsymbol{\Delta} = \mathbf{I} - \bar{\mathbf{A}}$ is symmetric, positive semi-definite with harmonic space of the form [14]

$$\ker(\boldsymbol{\Delta}) := \mathrm{span}(\mathbf{D}^{\frac{1}{2}}\mathbf{1}_n :\ \mathbf{1}_n = (1,\dots,1)^\top). \tag{20}$$

This confirms that if a given GNN evolution $\dot{\mathbf{F}}(t) = \mathrm{GNN}_\theta(\mathbf{F}(t), t)$ with initial condition $\mathbf{F}(0)$ over-smooths as per Definition 2.1 – i.e. $\boldsymbol{\Delta}\mathbf{f}^r(t) \to \mathbf{0}$ for $t \to \infty$ for each column $1 \le r \le d$ – then the only information persisting in the asymptotic regime is the degree and any dependence on the input features is lost, as studied in [30, 8]. A slightly different behaviour occurs if instead of $\boldsymbol{\Delta}$, we consider the unnormalized Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ with kernel $\mathrm{span}(\mathbf{1}_n)$, meaning that if $\mathbf{L}\mathbf{f}^r(t) \to \mathbf{0}$ as $t \to \infty$ for each $1 \le r \le d$, then any node would be embedded to a single point, hence making any separation task impossible. The same consequence applies to the random walk Laplacian $\boldsymbol{\Delta}_{\mathrm{RW}} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$. In particular, we note that generally a row-stochastic matrix is not symmetric – if it was, then this would in fact be doubly-stochastic – and the same applies to the random-walk Laplacian (a special exception is given by the class of *regular* graphs). In fact, in general any dynamical system governed by $\boldsymbol{\Delta}_{\mathrm{RW}}$ (or simply $\mathbf{D}^{-1}\mathbf{A}$) is not the gradient flow of an energy due to the lack of symmetry, as further confirmed below in eq. (22).

A.3 Additional details on LFD and HFD characterizations

In this subsection we provide further details and justifications for Definition 2.1 and Definition 2.2. We first prove the following simple properties.

Lemma A.1. *Assume we have a (continuous) process $t \mapsto \mathbf{F}(t) \in \mathbb{R}^{n\times d}$, for $t \ge 0$. The following equivalent characterizations hold:*

(i) $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \to 0$ *for* $t \to \infty$ *if and only if* $\boldsymbol{\Delta}\mathbf{f}^r(t) \to \mathbf{0}$*, for* $1 \le r \le d$*.*

(ii) $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \to \rho_{\boldsymbol{\Delta}}/2$ *for* $t \to \infty$ *if and only if for any sequence* $t_j \to \infty$ *there exist a subsequence* $t_{j_k} \to \infty$ *and a unit limit* $\mathbf{F}_\infty$ *– depending on the subsequence – such that* $\boldsymbol{\Delta}\mathbf{f}^r_\infty = \rho_{\boldsymbol{\Delta}}\mathbf{f}^r_\infty$*, for* $1 \le r \le d$*.*

Proof. (i) Given $\mathbf{F}(t) \in \mathbb{R}^{n\times d}$, we can vectorize it and decompose it in the orthonormal basis $\{\mathbf{e}_r \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}} : \ 1 \le r \le d, \ 0 \le i \le n-1\}$, with $\{\mathbf{e}_r\}_{r=1}^d$ canonical basis in $\mathbb{R}^d$, and write

$$\mathrm{vec}(\mathbf{F}(t)) = \sum_{r,i} c_{r,i}(t)\mathbf{e}_r \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}}, \quad c_{r,i}(t) := \langle \mathrm{vec}(\mathbf{F}(t)), \mathbf{e}_r \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}} \rangle.$$

We can then use eq. (19) to compute the Dirichlet energy as

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) = \frac{1}{2}\sum_{r=1}^{d}\sum_{i=0}^{n-1} c_{r,i}^2(t)\lambda_i^{\boldsymbol{\Delta}} \equiv \frac{1}{2}\sum_{r=1}^{d}\sum_{i=1}^{n-1} c_{r,i}^2(t)\lambda_i^{\boldsymbol{\Delta}} \ge \frac{1}{2}\mathrm{gap}(\boldsymbol{\Delta})\sum_{r=1}^{d}\sum_{i=1}^{n-1} c_{r,i}^2(t),$$

where we have used the convention above that the eigenvector $\boldsymbol{\phi}_0^{\boldsymbol{\Delta}}$ is in the kernel of $\boldsymbol{\Delta}$. Therefore

$$\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) \to 0 \quad \Longleftrightarrow \quad \sum_{r=1}^{d}\sum_{i=1}^{n-1} c_{r,i}^2(t) \to 0, \quad t \to \infty,$$

which occurs if and only if

$$(\mathbf{I}_d \otimes \boldsymbol{\Delta})\mathrm{vec}(\mathbf{F}(t)) = \sum_{r=1}^{d}\sum_{i=1}^{n-1} c_{r,i}(t)\lambda_i^{\boldsymbol{\Delta}}\mathbf{e}_r \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}} \to 0.$$

(ii) The argument here is similar. Indeed we can write $\mathbf{Q}(t) = \mathbf{F}(t)/||\mathbf{F}(t)||$ with $\mathbf{Q}(t)$ a unit-norm signal. Namely, we can vectorize and write

$$\mathrm{vec}(\mathbf{Q}(t)) = \sum_{r,i} q_{r,i}(t)\mathbf{e}_r \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}}, \quad \sum_{r,i} q_{r,i}^2(t) = 1.$$

Then $\mathcal{E}^{\mathrm{Dir}}(\mathbf{Q}(t)) \to \rho_{\boldsymbol{\Delta}}/2$ if and only if

$$\sum_{r,i} q_{r,i}^2(t)\lambda_i^{\boldsymbol{\Delta}} \to \rho_{\boldsymbol{\Delta}}, \quad t \to \infty,$$

which holds if and only if

$$\begin{aligned} &\sum_r q_{r,\rho_{\boldsymbol{\Delta}}}^2(t) \to 1 \\ &\qquad q_{r,i}^2(t) \to 0, \quad i : \lambda_i^{\boldsymbol{\Delta}} < \rho_{\boldsymbol{\Delta}}, \end{aligned} \tag{21}$$

given the unit norm constraint. This is equivalent to the Rayleigh quotient of $\mathbf{I}_d \otimes \boldsymbol{\Delta}$ converging to its maximal value $\rho_{\boldsymbol{\Delta}}$. When this occurs, for any sequence $t_j \to \infty$ we have that $q_{r,i}^2(t_j) \le 1$, meaning that we can extract a converging subsequence that due to eq. (21) will converge to a unit eigenvector $\mathbf{Q}_\infty$ of $\mathbf{I}_d \otimes \boldsymbol{\Delta}$ satisfying $(\mathbf{I}_d \otimes \boldsymbol{\Delta})\mathbf{Q}_\infty = \rho_{\boldsymbol{\Delta}}\mathbf{Q}_\infty$. Conversely assume for a contradiction that there exists a sequence $t_j \to \infty$ such that $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t_j)/||\mathbf{F}(t_j)||) < \rho_{\boldsymbol{\Delta}}/2 - \epsilon$, for some $\epsilon > 0$. Then eq. (21) fails to be satisfied along the sequence, meaning that no subsequence converges to a unit norm eigenvector $\mathbf{F}_\infty$ of $\mathbf{I}_d \otimes \boldsymbol{\Delta}$ with associated eigenvalue $\rho_{\boldsymbol{\Delta}}$ which is a contradiction to our assumption.

□

Before we address the formulation of low(high)-frequency-dominated dynamics, we solve explicitly the system $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)$ in $\mathbb{R}^{n\times d}$, with some initial condition $\mathbf{F}(0)$. We can vectorize the equation and solve $\dot{\mathrm{vec}}(\mathbf{F}(t)) = (\mathbf{I}_d \otimes \bar{\mathbf{A}})\mathrm{vec}(\mathbf{F}(t))$, meaning that

$$\mathrm{vec}(\mathbf{F}(t)) = \sum_{r=1}^{d}\sum_{i=0}^{n-1} e^{(1-\lambda_i^{\boldsymbol{\Delta}})t} c_{r,i}(0)\mathbf{e}_r \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}}, \qquad c_{r,i}(0) := \langle \mathrm{vec}(\mathbf{F}(0)), \mathbf{e}_r \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}} \rangle.$$

Consider any initial condition $\mathbf{F}(0)$ such that

$$\sum_{r=1}^{d}|c_{r,0}| = \sum_{r=1}^{d}\left|\langle \mathrm{vec}(\mathbf{F}(0)), \mathbf{e}_r \otimes \boldsymbol{\phi}_0^{\boldsymbol{\Delta}}\rangle\right| > 0,$$

which is satisfied for each $\mathrm{vec}(\mathbf{F}(0)) \in \mathbb{R}^{nd} \setminus \mathcal{U}^{\perp}$, where $\mathcal{U}^{\perp}$ is the orthogonal complement of $\mathbb{R}^d \otimes \mathrm{span}(\boldsymbol{\phi}_0^{\boldsymbol{\Delta}})$. Since $\mathcal{U}^{\perp}$ is a lower-dimensional subspace, its complement is dense. Accordingly for a.e. $\mathbf{F}(0)$, we find that the solution satisfies

$$||\mathrm{vec}(\mathbf{F}(t))||^2 = e^{2t}\left(\sum_{r=1}^{d} c_{r,0}^2 + \mathcal{O}(e^{-2\mathrm{gap}(\boldsymbol{\Delta})t})\right) = e^{2t}\left(||P^{\perp}_{\ker(\boldsymbol{\Delta})}\mathrm{vec}(\mathbf{F}(0))||^2 + \mathcal{O}(e^{-2\mathrm{gap}(\boldsymbol{\Delta})t})\right),$$

with $P^{\perp}_{\ker(\boldsymbol{\Delta})}$ the projection onto $\mathbb{R}^d \otimes \ker(\boldsymbol{\Delta})$. We see that the norm of the solution increases exponentially, however *the dominant term is given by the projection onto the lowest frequency signal* and in fact

$$\frac{\mathrm{vec}(\mathbf{F}(t))}{||\mathrm{vec}(\mathbf{F}(t))||} = \frac{P^{\perp}_{\ker(\boldsymbol{\Delta})}\mathrm{vec}(\mathbf{F}(0)) + \mathcal{O}(e^{-\mathrm{gap}(\boldsymbol{\Delta})t})(\mathbf{I} - P^{\perp}_{\ker(\boldsymbol{\Delta})})\mathrm{vec}(\mathbf{F}(0))}{\left(||P^{\perp}_{\ker(\boldsymbol{\Delta})}\mathrm{vec}(\mathbf{F}(0))||^2 + \mathcal{O}(e^{-2\mathrm{gap}(\boldsymbol{\Delta})t})\right)^{\frac{1}{2}}} \to \mathrm{vec}(\mathbf{F}_{\infty}),$$

such that $(\mathbf{I}_d \otimes \boldsymbol{\Delta})\mathrm{vec}(\mathbf{F}_{\infty}) = \mathbf{0}$ which means $\boldsymbol{\Delta}\mathbf{f}^r_{\infty} = \mathbf{0}$, for each column $1 \leq r \leq d$. Equivalently, one can compute $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||)$ and conclude that the latter quantity converges to zero as $t \to \infty$ by the very same argument.

In fact, this motivates further the nomenclature LFD and HFD. Without loss of generality we focus now on the high-frequency case. Assume that we have a HFD dynamics $t \mapsto \mathbf{F}(t)$, i.e. $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \to \rho_{\boldsymbol{\Delta}}/2$, then we can vectorize the solution and write $\mathrm{vec}(\mathbf{F}(t)) = ||\mathbf{F}(t)||\mathrm{vec}(\mathbf{Q}(t))$, for some time-dependent unit vector $\mathrm{vec}(\mathbf{Q}(t)) \in \mathbb{R}^{nd}$:

$$\mathrm{vec}(\mathbf{Q}(t)) = \sum_{r.i} q_{r,i}(t)\mathbf{e}_r \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}}, \quad \sum_{r,i} q^2_{r,i}(t) = 1.$$

By Lemma A.1 and more explicitly eq. (21), we derive that the coefficients $\{q_{r,\rho_{\boldsymbol{\Delta}}}\}$ associated with the eigevenctors $\mathbf{e}_r \otimes \boldsymbol{\phi}^{\boldsymbol{\Delta}}_{\rho_{\boldsymbol{\Delta}}}$ are dominant in the evolution hence justifying the name *high-frequency dominated* dynamics.

We note that the next result covers Lemma 2.3.

Lemma A.2. *Consider a dynamical system $\dot{\mathbf{F}}(t) = \mathrm{GNN}_{\theta}(\mathbf{F}(t), t)$, with initial condition $\mathbf{F}(0)$.*

- *(i) GNN_{θ} is* LFD *if and only if $(\mathbf{I}_d \otimes \boldsymbol{\Delta})\frac{\mathrm{vec}(\mathbf{F}(t))}{||\mathbf{F}(t)||} \to \mathbf{0}$ if and only if for each sequence $t_j \to \infty$ there exist a subsequence $t_{j_k} \to \infty$ and $\mathbf{F}_{\infty}$ (depending on the subsequence) s.t. $\frac{\mathbf{F}(t_{j_k})}{||\mathbf{F}(t_{j_k})||} \to \mathbf{F}_{\infty}$ satisfying $\boldsymbol{\Delta}\mathbf{f}^r_{\infty} = \mathbf{0}$, for each $1 \leq r \leq d$.*
- *(ii) GNN_{θ} is* HFD *if and only if for each sequence $t_j \to \infty$ there exist a subsequence $t_{j_k} \to \infty$ and $\mathbf{F}_{\infty}$ (depending on the subsequence) s.t. $\frac{\mathbf{F}(t_{j_k})}{||\mathbf{F}(t_{j_k})||} \to \mathbf{F}_{\infty}$ satisfying $\boldsymbol{\Delta}\mathbf{f}^r_{\infty} = \rho_{\boldsymbol{\Delta}}\mathbf{f}^r_{\infty}$, for each $1 \leq r \leq d$.*

Proof. (i) Since $\boldsymbol{\Delta}\mathbf{f}^r(t) \to \mathbf{0}$ for each $1 \leq r \leq d$ if and only if $(\mathbf{I}_d \otimes \boldsymbol{\Delta})\mathrm{vec}(\mathbf{F}(t)) \to \mathbf{0}$, we conclude that the dynamics is LFD if and only if $(\mathbf{I}_d \otimes \boldsymbol{\Delta})\frac{\mathrm{vec}(\mathbf{F}(t))}{||\mathbf{F}(t)||} \to \mathbf{0}$ due to (i) in Lemma A.1. Consider a sequence $t_j \to \infty$. Since $\mathrm{vec}(\mathbf{F}(t_j))/||\mathbf{F}(t_j)||$ is a bounded sequence we can extract a converging subsequence t_{j_k}: $\mathrm{vec}(\mathbf{F}(t_{j_k}))/||\mathbf{F}(t_{j_k})|| \to \mathrm{vec}(\mathbf{F}_{\infty})$. If the dynamics is LFD, then $(\mathbf{I}_d \otimes \boldsymbol{\Delta})\frac{\mathrm{vec}(\mathbf{F}(t_{j_k}))}{||\mathbf{F}(t_{j_k})||} \to \mathbf{0}$ and hence we conclude that $\mathrm{vec}(\mathbf{F}_{\infty}) \in \ker(\mathbf{I}_d \otimes \boldsymbol{\Delta})$. Conversely, assume that for any sequence $t_j \to \infty$ there exists a subsequence t_{j_k} and $\mathbf{F}_{\infty}$ such that $\frac{\mathbf{F}(t_{j_k})}{||\mathbf{F}(t_{j_k})||} \to \mathbf{F}_{\infty}$ satisfying $\boldsymbol{\Delta}\mathbf{f}^r_{\infty} = \mathbf{0}$, for each $1 \leq r \leq d$. If for a contradiction we had $\varepsilon > 0$ and $t_j \to \infty$ such that $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t_j)/||\mathbf{F}(t_j)|| \geq \varepsilon$ – for j large enough – then by (i) in Lemma A.1 there exist $1 \leq r \leq d$, $i > 0$ and a subsequence t_{j_k} satisfying

$$|\langle\left(\frac{\mathrm{vec}(\mathbf{F}(t_{j_k}))}{||\mathbf{F}(t_{j_k})||}\right), \mathbf{e}_r \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}}\rangle| > \delta(\varepsilon) > 0,$$

meaning that there is no subsequence of $\{t_{j_k}\}$ s.t. $(\mathbf{I}_d \otimes \mathbf{\Delta})\mathrm{vec}(\mathbf{F}(t_{j_k}))/||\mathbf{F}(t_{j_k})|| \to \mathbf{0}$, providing a contradiction.

(ii) This is equivalent to (ii) in Lemma A.1.

□

Remark. *We note that in Lemma 2.3 an* LFD *dynamics does not necessarily mean that the normalized solution converges to the kernel of* $\mathbf{I}_d \otimes \mathbf{\Delta}$ *– i.e. one in general has always to pass to subsequences. Indeed, we can consider the simple example* $t \mapsto \mathrm{vec}(\mathbf{F}(t)) := \cos(t)\mathbf{e}_{\bar{r}} \otimes \boldsymbol{\phi}_0^{\mathbf{\Delta}}$*, for some* $1 \leq \bar{r} \leq d$*, which satisfies* $\mathbf{\Delta}\mathbf{f}^r(t) = \mathbf{0}$ *for each* r*, but it is not a convergent function due to its oscillatory nature. Same argument applies to* HFD*.*

A.4 Details and proofs on $\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}$ and its gradient flow

By direct computation one verifies that the definition in eq. (4) can be equivalently written as

$$\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathbf{F}) = \frac{1}{2}\langle \mathrm{vec}(\mathbf{F}), (\mathbf{W}^\top\mathbf{W} \otimes \mathbf{\Delta})\mathrm{vec}(\mathbf{F})\rangle,$$

from which we immediately derive $\nabla_{\mathrm{vec}(\mathbf{F})}\mathcal{E}^{\mathrm{Dir}}_{\mathbf{W}}(\mathrm{vec}(\mathbf{F})) = (\mathbf{W}^\top\mathbf{W} \otimes \mathbf{\Delta})\mathrm{vec}(\mathbf{F})$ which proves eq. (5). We can now address the proof of Proposition 2.4.

Proof of Proposition 2.4. We can vectorize the gradient flow system in eq. (5) and use the spectral characterization of $\mathbf{W}^\top\mathbf{W} \otimes \mathbf{\Delta}$ in eq. (17) to write the solution explicitly as

$$\mathrm{vec}(\mathbf{F}(t)) = \sum_{r,i} e^{-(\lambda_r^{\mathbf{W}}\lambda_i^{\mathbf{\Delta}})t} c_{r,i}(0)\boldsymbol{\phi}_r^{\mathbf{W}} \otimes \boldsymbol{\phi}_i^{\mathbf{\Delta}},$$

where $\{\lambda_r^{\mathbf{W}}\}_r = \mathrm{spec}(\mathbf{W}^\top\mathbf{W}) \subset \mathbb{R}_{\geq 0}$ with associated basis of orthonormal eigenvectors given by $\{\boldsymbol{\phi}_r^{\mathbf{W}}\}_r$. Then

$$\begin{aligned}
\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) &= \frac{1}{2}\langle \mathrm{vec}(\mathbf{F}(t)), (\mathbf{I}_d \otimes \mathbf{\Delta})\mathrm{vec}(\mathbf{F}(t))\rangle = \frac{1}{2}\sum_{r,i} e^{-2t(\lambda_r^{\mathbf{W}}\lambda_i^{\mathbf{\Delta}})} c_{r,i}^2(0)\lambda_i^{\mathbf{\Delta}} \\
&= \frac{1}{2}\sum_{r:\lambda_r^{\mathbf{W}}=0,i} c_{r,i}^2(0)\lambda_i^{\mathbf{\Delta}} + \frac{1}{2}\sum_{r:\lambda_r^{\mathbf{W}}>0,i>0} c_{r,i}^2(0)e^{-2t(\lambda_r^{\mathbf{W}}\lambda_i^{\mathbf{\Delta}})}\lambda_i^{\mathbf{\Delta}} \\
&= \mathcal{E}^{\mathrm{Dir}}((P^{\mathrm{ker}}_{\mathbf{W}} \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0))) + \frac{1}{2}\sum_{r:\lambda_r^{\mathbf{W}}>0,i>0} c_{r,i}^2(0)e^{-2t(\lambda_r^{\mathbf{W}}\lambda_i^{\mathbf{\Delta}})}\lambda_i^{\mathbf{\Delta}} \\
&\leq \mathcal{E}^{\mathrm{Dir}}((P^{\mathrm{ker}}_{\mathbf{W}} \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0))) + \frac{\rho_{\mathbf{\Delta}}}{2}e^{-2t\mathrm{gap}(\mathbf{W}^\top\mathbf{W})\mathrm{gap}(\mathbf{\Delta})}||\mathbf{F}(0)||^2,
\end{aligned}$$

where we recall that $P^{\mathrm{ker}}_{\mathbf{W}}$ is the projection onto $\ker(\mathbf{W}^\top\mathbf{W})$ and that by convention the index $i = 0$ is associated with the lowest graph frequency $\lambda_0^{\mathbf{\Delta}} = 0$ – by assumption G is connected. This proves that the dynamics is in fact smoothing as per Definition 2.1. By the very same argument we find that

$$\mathrm{vec}(\mathbf{F}(t)) \to (\mathbf{I}_d \otimes P^{\mathrm{ker}}_{\mathbf{\Delta}})\mathrm{vec}(\mathbf{F}(0)) + (P^{\mathrm{ker}}_{\mathbf{W}} \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0)), \quad t \to \infty,$$

with $P^{\mathrm{ker}}_{\mathbf{\Delta}}$ the orthogonal projection onto $\ker\mathbf{\Delta}$ – the other terms decay exponentially to zero. We first focus on the first quantity, which we can write as

$$(\mathbf{I}_d \otimes P^{\mathrm{ker}}_{\mathbf{\Delta}})\mathrm{vec}(\mathbf{F}(0)) = \sum_r c_{r,0}(0)\boldsymbol{\phi}_r^{\mathbf{W}} \otimes \boldsymbol{\phi}_0^{\mathbf{\Delta}},$$

which has matrix representation $\boldsymbol{\phi}_0^{\mathbf{\Delta}}\boldsymbol{\phi}_\infty^\top \in \mathbb{R}^{n\times d}$ with

$$\boldsymbol{\phi}_\infty := \sum_r c_{r,0}(0)\boldsymbol{\phi}_r^{\mathbf{W}}.$$

By eq. (20) we deduce that the i-th row of $\boldsymbol{\phi}_0^{\mathbf{\Delta}}\boldsymbol{\phi}_\infty^\top \in \mathbb{R}^{n\times d}$ is the d-dimensional vector $\sqrt{d_i}\boldsymbol{\phi}_\infty$. We now focus on the term

$$(P^{\mathrm{ker}}_{\mathbf{W}} \otimes \mathbf{I}_n)\mathrm{vec}(\mathbf{F}(0)) = \sum_{r:\lambda_r^{\mathbf{W}}=0,j} c_{r,j}(0)\boldsymbol{\phi}_r^{\mathbf{W}} \otimes \boldsymbol{\phi}_j^{\mathbf{\Delta}}$$

which has matrix representation $\sum_{r:\lambda_r^{\mathbf{W}}=0,j} c_{r,j}(0)\boldsymbol{\phi}_j^{\boldsymbol{\Delta}}(\boldsymbol{\phi}_r^{\mathbf{W}})^\top$. In particular, the i-th row is given by

$$\sum_{r:\lambda_r^{\mathbf{W}}=0,j} c_{r,j}(0)(\boldsymbol{\phi}_j^{\boldsymbol{\Delta}})_i\boldsymbol{\phi}_r^{\mathbf{W}} = P_{\mathbf{W}}^{\ker}\mathbf{f}_i(0).$$

This completes the proof of Proposition 2.4. □

B Proofs and additional details of Section 3

B.1 Spectral analysis of the channel-mixing: the continuous case

Consider the generalized energy $\mathcal{E}^{\text{tot}}$ in eq. (7). We can use vectorization to rewrite it as

$$\mathcal{E}^{\text{tot}}(\text{vec}(\mathbf{F})) = \frac{1}{2}\langle\text{vec}(\mathbf{F}), (\boldsymbol{\Omega}\otimes\mathbf{I}_n)\text{vec}(\mathbf{F})\rangle - \frac{1}{2}\langle\text{vec}(\mathbf{F}), (\mathbf{W}\otimes\bar{\mathbf{A}})\text{vec}(\mathbf{F})\rangle,$$

from which the gradient flow in eq. (8) follows. In particular, given a system as in eq. (8):

$$\text{vec}(\dot{\mathbf{F}}(t)) = -(\boldsymbol{\Omega}\otimes\mathbf{I}_n)\text{vec}(\mathbf{F}(t)) + (\mathbf{W}\otimes\bar{\mathbf{A}})\text{vec}(\mathbf{F}(t)),$$

if this is the gradient flow of $\mathbf{F}\mapsto\mathcal{E}^{\text{tot}}(\mathbf{F})$, then we would have

$$\nabla^2_{\text{vec}(\mathbf{F})}\mathcal{E}^{\text{tot}}(\mathbf{F}) = \boldsymbol{\Omega}\otimes\mathbf{I}_n - \mathbf{W}\otimes\bar{\mathbf{A}}, \tag{22}$$

which must be symmetric due to the Hessian of a function being symmetric. The latter means

$$(\boldsymbol{\Omega}^\top - \boldsymbol{\Omega})\otimes\mathbf{I}_n = (\mathbf{W}^\top - \mathbf{W})\otimes\bar{\mathbf{A}},$$

which is satisfied if and only if both $\boldsymbol{\Omega}$ and $\mathbf{W}$ are *symmetric*. This shows that eq. (8) *is the gradient flow of $\mathcal{E}^{\text{tot}}$ if and only if $\boldsymbol{\Omega}$ and $\mathbf{W}$ are symmetric*.

We now rely on the spectral decomposition of $\mathbf{W}$ to rewrite $\mathcal{E}^{\text{tot}}$ explicitly in terms of attractive and repulsive interactions. If we have a spectral decomposition $\mathbf{W} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$, we can separate the positive eigenvalues from the negative ones and write

$$\mathbf{W} = \mathbf{U}\boldsymbol{\Lambda}_+\mathbf{U}^\top + \mathbf{U}\boldsymbol{\Lambda}_-\mathbf{U}^\top := \mathbf{W}_+ - \mathbf{W}_-.$$

Since $\mathbf{W}_+ \succeq 0, \mathbf{W}_- \succeq 0$, we can use the Choleski decomposition to write $\mathbf{W}_+ = \boldsymbol{\Theta}_+^\top\boldsymbol{\Theta}_+$ and $\mathbf{W}_- = \boldsymbol{\Theta}_-^\top\boldsymbol{\Theta}_-$ with $\boldsymbol{\Theta}_+, \boldsymbol{\Theta}_- \in \mathbb{R}^{d\times d}$. Equation (9) follows then by direct computation: namely

$$\begin{aligned}
\mathcal{E}^{\text{tot}}(\mathbf{F}) &= \frac{1}{2}\sum_i\langle\mathbf{f}_i, \boldsymbol{\Omega}\mathbf{f}_i\rangle - \frac{1}{2}\sum_{i,j}\bar{a}_{ij}\langle\mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle \\
&= \frac{1}{2}\sum_i\langle\mathbf{f}_i, (\boldsymbol{\Omega}-\mathbf{W})\mathbf{f}_i\rangle + \frac{1}{2}\sum_i\langle\mathbf{f}_i, \mathbf{W}\mathbf{f}_i\rangle - \frac{1}{2}\sum_{i,j}\bar{a}_{ij}\langle\boldsymbol{\Theta}_+\mathbf{f}_i, \boldsymbol{\Theta}_+\mathbf{f}_j\rangle + \frac{1}{2}\sum_{i,j}\bar{a}_{ij}\langle\boldsymbol{\Theta}_-\mathbf{f}_i, \boldsymbol{\Theta}_-\mathbf{f}_j\rangle \\
&= \frac{1}{2}\sum_i\langle\mathbf{f}_i, (\boldsymbol{\Omega}-\mathbf{W})\mathbf{f}_i\rangle + \frac{1}{4}\sum_{i,j}||\boldsymbol{\Theta}_+(\nabla\mathbf{F})_{ij}||^2 - \frac{1}{4}\sum_{i,j}||\boldsymbol{\Theta}_-(\nabla\mathbf{F})_{ij}||^2,
\end{aligned}$$

where we have used that $\sum_{i,j}\frac{1}{d_i}||\boldsymbol{\Theta}_+\mathbf{f}_i||^2 = \sum_i||\boldsymbol{\Theta}_+\mathbf{f}_i||^2$.

Proof of Proposition 3.1. Once we compute the spectrum of $\mathbf{W}\otimes\bar{\mathbf{A}}$ via eq. (17), we can write the solution as – recall that $\bar{\mathbf{A}} = \mathbf{I}_n - \boldsymbol{\Delta}$ so we can rephrase the eigenvalues of $\bar{\mathbf{A}}$ in terms of the eigenvalues of $\boldsymbol{\Delta}$:

$$\text{vec}(\mathbf{F}(t)) = \sum_{r,i} e^{\lambda_r^{\mathbf{W}}(1-\lambda_i^{\boldsymbol{\Delta}})t}c_{r,i}(0)\boldsymbol{\phi}_r^{\mathbf{W}}\otimes\boldsymbol{\phi}_i^{\boldsymbol{\Delta}},$$

with $\mathbf{W}\boldsymbol{\phi}_r^{\mathbf{W}} = \lambda_r^{\mathbf{W}}\boldsymbol{\phi}_r^{\mathbf{W}}$, for $1\le r\le d$, where $\{\boldsymbol{\phi}_r^{\mathbf{W}}\}_r$ is an orthonormal basis of eigenvectors in $\mathbb{R}^d$. We can then calculate the Dirichlet energy along the solution as

$$\mathcal{E}^{\text{Dir}}(\mathbf{F}(t)) = \frac{1}{2}\langle\text{vec}(\mathbf{F}(t)), (\mathbf{I}_d\otimes\boldsymbol{\Delta})\text{vec}(\mathbf{F}(t))\rangle = \frac{1}{2}\sum_{r,i}e^{2\lambda_r^{\mathbf{W}}(1-\lambda_i^{\boldsymbol{\Delta}})t}c_{r,i}^2(0)\lambda_i^{\boldsymbol{\Delta}}.$$

We now consider two cases:

- If $\lambda_r^{\mathbf{W}} > 0$, then $\lambda_r^{\mathbf{W}}(1-\lambda_i^{\mathbf{\Delta}}) \leq \lambda_+^{\mathbf{W}}$.
- If $\lambda_r^{\mathbf{W}} < 0$, then $\lambda_r^{\mathbf{W}}(1-\lambda_i^{\mathbf{\Delta}}) \leq |\lambda_-^{\mathbf{W}}|(\rho_{\mathbf{\Delta}}-1) := \rho_-$, with eigenvectors $\boldsymbol{\phi}_r^{\mathbf{W}} \otimes \boldsymbol{\phi}_{\rho_{\mathbf{\Delta}}}^{\mathbf{\Delta}}$ for each r s.t. $\mathbf{W}\boldsymbol{\phi}_r^{\mathbf{W}} = \lambda_-^{\mathbf{W}}\boldsymbol{\phi}_r^{\mathbf{W}}$ – without loss of generality we can assume that $\rho_{\mathbf{\Delta}}$ is a simple eigenvalue for $\mathbf{\Delta}$. In particular, if $\lambda_r^{\mathbf{W}} < 0$ and $\lambda_r^{\mathbf{W}}(1-\lambda_i^{\mathbf{\Delta}}) < \rho_-$, then

$$\lambda_r^{\mathbf{W}}(1-\lambda_i^{\mathbf{\Delta}}) < \max\{|\lambda_-^{\mathbf{W}}|(\lambda_{n-2}^{\mathbf{\Delta}}-1), |\lambda_{-,2}^{\mathbf{W}}|(\rho_{\mathbf{\Delta}}-1)\},$$

where $\lambda_{-,2}^{\mathbf{W}}$ is the second most negative eigenvalue of $\mathbf{W}$ and $\lambda_{n-2}^{\mathbf{\Delta}}$ is the second largest eigenvalue of $\mathbf{\Delta}$. In particular, we can write

$$\lambda_{n-2}^{\mathbf{\Delta}} = \rho_{\mathbf{\Delta}} - \mathrm{gap}(\rho_{\mathbf{\Delta}}\mathbf{I}_n - \mathbf{\Delta}), \quad |\lambda_{-,2}^{\mathbf{W}}| = |\lambda_-^{\mathbf{W}}| - \mathrm{gap}(|\lambda_-^{\mathbf{W}}|\mathbf{I}_d + \mathbf{W}). \tag{23}$$

From (i) and (ii) we derive that if $\lambda_r^{\mathbf{W}}(1-\lambda_i^{\mathbf{\Delta}}) \neq \rho_-$, then

$$\begin{aligned}\lambda_r^{\mathbf{W}}(1-\lambda_i^{\mathbf{\Delta}}) - \rho_- &< -\min\{\rho_- - \lambda_+^{\mathbf{W}}, \rho_- - |\lambda_-^{\mathbf{W}}|(\lambda_{n-2}^{\mathbf{\Delta}}-1), \rho_- - |\lambda_{-,2}^{\mathbf{W}}|(\rho_{\mathbf{\Delta}}-1)\} \\ &= -\min\{\rho_- - \lambda_+^{\mathbf{W}}, |\lambda_-^{\mathbf{W}}|\mathrm{gap}(\rho_{\mathbf{\Delta}}\mathbf{I} - \mathbf{\Delta}), \mathrm{gap}(|\lambda_-^{\mathbf{W}}|\mathbf{I} + \mathbf{W})(\rho_{\mathbf{\Delta}}-1)\} = -\epsilon_{\mathrm{HFD}},\end{aligned} \tag{24}$$

where we have used eq. (23). Accordingly, if $\rho_- > \lambda_+^{\mathbf{W}}$, then

$$\begin{aligned}\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) &= e^{2t\rho_-}\left(\frac{\rho_{\mathbf{\Delta}}}{2}\sum_{r:\lambda_r^{\mathbf{W}}=\lambda_-^{\mathbf{W}}} c_{r,\rho_{\mathbf{\Delta}}}^2(0) + \frac{1}{2}\sum_{r,i:\lambda_r^{\mathbf{W}}(1-\lambda_i^{\mathbf{\Delta}})\neq\rho_-} e^{2(\lambda_r^{\mathbf{W}}(1-\lambda_i^{\mathbf{\Delta}})-\rho_-)t}c_{r,i}^2(0)\right) \\ &= e^{2t\rho_-}\left(\frac{\rho_{\mathbf{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}(e^{-2t\epsilon_{\mathrm{HFD}}})\right).\end{aligned}$$

By the same argument we can factor out the dominant term and derive the following limit for $t \to \infty$ and for a.e. $\mathbf{F}(0)$ since $P_{\mathbf{W}}^{\rho_-}\mathrm{vec}(\mathbf{F}(0)) = \mathbf{0}$ only if $\mathrm{vec}(\mathbf{F}(0))$ belongs to a lower dimensional subspace of $\mathbb{R}^{nd}$:

$$\frac{\mathrm{vec}(\mathbf{F}(t))}{\mathrm{vec}(\mathbf{F}(t))} = \frac{P_{\mathbf{W}}^{\rho_-}\mathrm{vec}(\mathbf{F}(0)) + \mathcal{O}(e^{-\epsilon_{\mathrm{HFD}}t})((\mathbf{I}-P_{\mathbf{W}}^{\rho_-})\mathrm{vec}(\mathbf{F}(0)))}{\left(||P_{\mathbf{W}}^{\rho_-}\mathrm{vec}(\mathbf{F}(0))||^2 + \mathcal{O}(e^{-2\epsilon_{\mathrm{HFD}}t})\right)^{\frac{1}{2}}} \to \frac{P_{\mathbf{W}}^{\rho_-}\mathrm{vec}(\mathbf{F}(0))}{||P_{\mathbf{W}}^{\rho_-}\mathrm{vec}(\mathbf{F}(0))||},$$

where the latter is a unit vector $\mathrm{vec}(\mathbf{F}_\infty)$ satisfying $(\mathbf{I}_d \otimes \mathbf{\Delta})\mathrm{vec}(\mathbf{F}_\infty) = \rho_{\mathbf{\Delta}}\mathrm{vec}(\mathbf{F}_\infty)$, which completes the proof. □

B.2 Propagating with $-\mathbf{\Delta}$: a perspective in terms of channel-mixing spectrum

In this subsection we briefly review the special case of eq. (8) where $\mathbf{\Omega} = \mathbf{W}$, and comment on why we generally expect a framework where the propagation is governed by the graph vector field $\bar{\mathbf{A}}$ to be more flexible than one with $-\mathbf{\Delta}$. If $\mathbf{\Omega} = \mathbf{W}$, the gradient flow in eq. (8) becomes

$$\dot{\mathbf{F}}(t) = -\mathbf{\Delta}\mathbf{F}(t)\mathbf{W}. \tag{25}$$

We note that once vectorized, the solution to the dynamical system can be written as

$$\mathrm{vec}(\mathbf{F}(t)) = \sum_{r=1}^{d}\sum_{i=0}^{n-1} e^{-\lambda_r^{\mathbf{W}}\lambda_i^{\mathbf{\Delta}}t}c_{r,i}(0)\boldsymbol{\phi}_r^{\mathbf{W}} \otimes \boldsymbol{\phi}_i^{\mathbf{\Delta}}.$$

In particular, we immediately deduce the following counterpart to Proposition 3.1

Corollary B.1. *If* $\mathrm{spec}(\mathbf{W}) \cap \mathbb{R}_- \neq \emptyset$*, then eq.* (25) *is* HFD *for a.e.* $\mathbf{F}(0)$.

Differently from eq. (8) the lowest frequency component is *always preserved independent of the spectrum of* $\mathbf{W}$. This means that the system cannot learn eigenvalues of $\mathbf{W}$ to either magnify or suppress the low-frequency projection. In contrast, this can be done if $\mathbf{\Omega} = \mathbf{0}$, or equivalently one replaces $-\mathbf{\Delta}$ with $\mathbf{A}$ providing a further justification in terms of the interaction between graph spectrum and channel-mixing spectrum for why graph convolutional models use the normalized adjacency rather than the Laplacian for propagating messages [27].

B.3 A more general family of energies: gradient flow with non-linear activations

Consider a more general pairwise energy including a non-linear differentiable activation map σ of the form

$$\mathcal{E}^{\text{pair}}_{\sigma,\mathbf{W}}(\mathbf{F}) = \frac{1}{2}\sum_{i,j}\bar{a}_{i,j}\sigma\left(\mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle\right).$$

We temporarily assume that $\mathbf{\Omega} = \mathbf{0}$. The gradient flow follows from direct computation:

$$\dot{\mathbf{F}}(t) = \boldsymbol{\mathcal{A}}_\sigma(\mathbf{F}(t))\mathbf{F}(t)\mathbf{W}, \quad (\boldsymbol{\mathcal{A}}_\sigma(\mathbf{F}(t)))_{ij} := \bar{a}_{ij}\sigma'\left(\mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle\right). \tag{26}$$

In particular, we see that the non-linear activations in general may induce a type of attention mechanism where the diffusion along edges is controlled by the derivative of σ evaluated on the inner product of features induced by $\mathbf{W}$. A similar structure is investigated in [17]. We also observe that analogous conclusions can be deduced if $\mathbf{\Omega} \neq \mathbf{0}$ and the external energy term $\mathcal{E}^{\text{ext}}_{\mathbf{\Omega}}$ includes a non-linear activation map σ as in the pairwise contribution.

C Proofs and additional details of Section 4

We first explicitly report here the expansion of the discrete gradient flow in eq. (11) after m layers to further highlight how this is not equivalent to a single linear layer with a message passing matrix $\bar{\mathbf{A}}^m$ as for SGCN [43]. For simplicity we suppress the source term.

$$\begin{aligned}
\mathbf{F}(t+\tau) &= \mathbf{F}(t) + \tau\left(-\mathbf{F}(t)\mathbf{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}\right)\\
\text{vec}(\mathbf{F}(t+\tau)) &= \left(\mathbf{I}_{nd} + \tau\left(-\mathbf{\Omega}\otimes\mathbf{I}_n + \mathbf{W}\otimes\bar{\mathbf{A}}\right)\right)\text{vec}(\mathbf{F}(t))\\
\text{vec}(\mathbf{F}(m\tau)) &= \sum_{k=0}^{m}\binom{m}{k}\tau^k\left(-\mathbf{\Omega}\otimes\mathbf{I}_n + \mathbf{W}\otimes\bar{\mathbf{A}}\right)^k\text{vec}(\mathbf{F}(0))
\end{aligned} \tag{27}$$

and we see how the message passing matrix $\bar{\mathbf{A}}$ actually enters the expansion after m layers with each power $0 \leq k \leq m$. This is not surprising, given that *we are discretizing a linear dynamical system, meaning that we are approximating an exponential matrix.*

C.1 Comparison with continuous GNNs: details and proofs

We prove the following result which covers Proposition 4.1.

Proof of Proposition 4.1. We structure the proof by following the numeration in the statement.

(i) From direct computation we find

$$\begin{aligned}
\frac{d\mathcal{E}^{\text{Dir}}(\mathbf{F}(t))}{dt} &= \frac{1}{2}\frac{d}{dt}\left(\langle\text{vec}(\mathbf{F}(t)), (\mathbf{I}_d\otimes\mathbf{\Delta})\text{vec}(\mathbf{F}(t))\rangle\right)\\
&= -\langle\text{vec}(\mathbf{F}(t)), (\mathbf{K}^\top(t)\mathbf{K}(t)\otimes\mathbf{\Delta}^2)\text{vec}(\mathbf{F}(t))\rangle \leq 0,
\end{aligned}$$

since $\mathbf{K}^\top(t)\mathbf{K}(t)\otimes\mathbf{\Delta}^2 \succeq 0$. Note that we have used that $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \mathbf{AC}\otimes\mathbf{BD}$.

(ii) We consider the dynamical system

$$\dot{\mathbf{F}}_{\text{CGNN}}(t) = -\mathbf{\Delta}\mathbf{F}(t) + \mathbf{F}(t)\tilde{\mathbf{\Omega}} + \mathbf{F}(0).$$

We can write $\text{vec}(\mathbf{F}(t)) = \sum_{r,i} c_{r,i}(t)\boldsymbol{\phi}_r^{\tilde{\mathbf{\Omega}}}\otimes\boldsymbol{\phi}_i^{\mathbf{\Delta}}$, leading to the system

$$\dot{c}_{r,i}(t) = (\lambda_r^{\tilde{\mathbf{\Omega}}} - \lambda_i^{\mathbf{\Delta}})c_{r,i}(t) + c_{r,i}(0), \quad 0 \leq i \leq n-1,\ 1 \leq r \leq d.$$

We can solve explicitly the system as

$$\begin{aligned}
c_{r,i}(t) &= c_{r,i}(0)\left(e^{(\lambda_r^{\tilde{\mathbf{\Omega}}} - \lambda_i^{\mathbf{\Delta}})t}\left(1 + \frac{1}{\lambda_r^{\tilde{\mathbf{\Omega}}} - \lambda_i^{\mathbf{\Delta}}}\right) - \frac{1}{\lambda_r^{\tilde{\mathbf{\Omega}}} - \lambda_i^{\mathbf{\Delta}}}\right), \quad \text{if } \lambda_r^{\tilde{\mathbf{\Omega}}} \neq \lambda_i^{\mathbf{\Delta}}\\
c_{r,i}(t) &= c_{r,i}(0)(1+t), \quad \text{otherwise.}
\end{aligned}$$

We see now that for a.e. $\mathbf{F}(0)$ the projection $(\mathbf{I}_d \otimes \boldsymbol{\phi}^{\boldsymbol{\Delta}}_{\rho_{\boldsymbol{\Delta}}}(\boldsymbol{\phi}^{\boldsymbol{\Delta}}_{\rho_{\boldsymbol{\Delta}}})^\top)\mathrm{vec}(\mathbf{F}(t))$ is never the dominant term. In fact, if there exists r s.t. $\lambda^{\tilde{\boldsymbol{\Omega}}}_r \geq \rho_{\boldsymbol{\Delta}}$, then $\lambda^{\tilde{\boldsymbol{\Omega}}}_r - \lambda^{\boldsymbol{\Delta}}_i > \lambda^{\tilde{\boldsymbol{\Omega}}}_r - \rho_{\boldsymbol{\Delta}}$, for any other non-maximal graph Laplacian eigenvalue. It follows that there is *no* $\tilde{\boldsymbol{\Omega}}$ s.t. the normalized solution maximizes the Rayleigh quotient of $\mathbf{I}_d \otimes \boldsymbol{\Delta}$, proving that CGNN is never HFD.

If we have no source, then the CGNN equation becomes

$$\dot{\mathbf{F}}(t) = -\boldsymbol{\Delta}\mathbf{F}(t) + \mathbf{F}(t)\tilde{\boldsymbol{\Omega}} \quad \Longleftrightarrow \mathrm{vec}(\dot{\mathbf{F}}(t)) = (\tilde{\boldsymbol{\Omega}} \oplus (-\boldsymbol{\Delta}))\mathrm{vec}(\mathbf{F}(t)),$$

using the Kronecker sum notation in eq. (18). It follows that we can write the vectorized solution in the basis $\{\boldsymbol{\phi}^{\tilde{\boldsymbol{\Omega}}}_r \otimes \boldsymbol{\phi}^{\boldsymbol{\Delta}}_i\}_{r,i}$ as

$$\mathrm{vec}(\mathbf{F}(t)) = e^{\lambda^{\tilde{\boldsymbol{\Omega}}}_+ t}\left(\sum_{r:\lambda^{\tilde{\boldsymbol{\Omega}}}_r = \lambda^{\tilde{\boldsymbol{\Omega}}}_+} c_{r,0}(0)\boldsymbol{\phi}^{\tilde{\boldsymbol{\Omega}}}_r \otimes \boldsymbol{\phi}^{\boldsymbol{\Delta}}_0 + \mathcal{O}(e^{-\mathrm{gap}(\lambda^{\tilde{\boldsymbol{\Omega}}}_+ \mathbf{I}_d - \tilde{\boldsymbol{\Omega}})t}) \sum_{r:\lambda^{\tilde{\boldsymbol{\Omega}}}_r < \lambda^{\tilde{\boldsymbol{\Omega}}}_+} c_{r,0}(0)\boldsymbol{\phi}^{\tilde{\boldsymbol{\Omega}}}_r \otimes \boldsymbol{\phi}^{\boldsymbol{\Delta}}_0\right)$$
$$+ e^{\lambda^{\tilde{\boldsymbol{\Omega}}}_+ t}\left(\mathcal{O}(e^{-\mathrm{gap}(\boldsymbol{\Delta})t})\left(\sum_{r,i>0} c_{r,i}(0)\boldsymbol{\phi}^{\tilde{\boldsymbol{\Omega}}}_r \otimes \boldsymbol{\phi}^{\boldsymbol{\Delta}}_i\right)\right),$$

meaning that the dominant term is given by the lowest frequency component and in fact, if we normalize we find $\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)/||\mathbf{F}(t)||) \leq e^{-\mathrm{gap}(\boldsymbol{\Delta})t}$.

(iii) Finally we consider the dynamical system induced by linear GRAND

$$\dot{\mathbf{F}}_{\mathrm{GRAND}}(t) = -\boldsymbol{\Delta}_{\mathrm{RW}}\mathbf{F}(t) = -(\mathbf{I} - \boldsymbol{\mathcal{A}}(\mathbf{F}(0)))\mathbf{F}(t).$$

Since we have no channel-mixing, without loss of generality we can assume that $d = 1$ – one can then extend the argument to any entry. We can use the Jordan form of $\boldsymbol{\mathcal{A}}$ to write the solution of the GRAND dynamical system as

$$\mathbf{f}(t) = P\mathrm{diag}(e^{J_1 t}, \dots, e^{J_n t})P^{-1}\mathbf{f}(0),$$

for some invertible matrix P of eigenvectors, with

$$e^{J_k t} = e^{-(1-\lambda^{\boldsymbol{\mathcal{A}}}_k)t}\begin{pmatrix} 1 & t & \cdots & \frac{t^{m_k-1}}{(m_k-1)!} \\ & & & \vdots \\ & & & 1 \end{pmatrix},$$

where m_k are the eigenvalue multiplicities. Since by assumption G is connected and augmented with self-loops, the row-stochastic attention matrix $\boldsymbol{\mathcal{A}}$ computed in [10] with softmax activation is *regular*, meaning that there exists $m \in \mathbb{N}$ such that $(\boldsymbol{\mathcal{A}}^m)_{ij} > 0$ for each entry (i,j). Accordingly, we can apply Perron Theorem to derive that any eigenvalue of $\boldsymbol{\mathcal{A}}$ has real part smaller than one except the eigenvalue $\lambda^{\boldsymbol{\mathcal{A}}}_0$ with multiplicity one, associated with the Perron eigenvector $\mathbf{1}_n$. Accordingly, we find that each block $e^{J_k t}$ decays to zero as $t \to \infty$ with the exception of the one $e^{J_0 t}$ associated with the Perron eigenvector. In particular, the projection of $\mathbf{f}_0$ over the Perron eigenvector is just $\mu\mathbf{1}_n$, with μ the average of the feature initial condition. This completes the proof. □

C.2 Common GNN architectures as gradient flow

We consider linear GNNs of the form

$$\mathbf{F}(t+1) = \mathbf{F}(t)\boldsymbol{\Omega} + \boldsymbol{\mathcal{A}}\mathbf{F}(t)\mathbf{W} + \beta\mathbf{F}(0)\tilde{\mathbf{W}}, \quad 0 \leq t \leq T.$$

If $\boldsymbol{\Omega} = \mathbf{0}, \beta = 0$ and $\boldsymbol{\mathcal{A}} = \bar{\mathbf{A}}$, we recover linear GCN with weights shared across layers [27, 43]. Similarly, if $\boldsymbol{\mathcal{A}} = \bar{\mathbf{A}}$ and $\beta = 0$, this is linear GraphSAGE [23] with propagation given by *symmetric* adjacency and weights shared across layers. A symmetric version of GAT [42] can be recovered if $\boldsymbol{\Omega} = \mathbf{0}, \beta = 0$ and $\boldsymbol{\mathcal{A}} = \bar{\mathbf{A}}$ is a *symmetric* attention matrix depending only on the initial encoded features – note that in general a row-stochastic matrix may not be symmetric so *a symmetrization of a row-stochastic attention matrix would generally fail to remain row-stochastic*. We believe that this point deserves further investigation. Finally GCNII [11] can be recovered by taking $\boldsymbol{\Omega} = \mathbf{0}$ and $\boldsymbol{\mathcal{A}} = \bar{\mathbf{A}}$.

Proof of Lemma 4.2. This follows from the same argument in eq. (22) once we regard the linear system in eq. (12) as a unit step size Euler discretization

$$\dot{\mathbf{F}}(t) \sim \mathbf{F}(t+1) - \mathbf{F}(t) = \mathbf{F}(t)(\boldsymbol{\Omega} - \mathbf{I}_d) + \boldsymbol{\mathcal{A}}\mathbf{F}(t)\mathbf{W} + \beta\mathbf{F}(0)\tilde{\mathbf{W}}$$

□

C.3 Spectral analysis of the channel-mixing: the discrete case

We first address the proof of the main result.

Proof of Theorem 4.3. We consider a linear dynamical system

$$\mathbf{F}(t+\tau) = \mathbf{F}(t) + \tau\bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W},$$

with $\mathbf{W}$ symmetric. We vectorize the system and rewrite it as

$$\mathrm{vec}(\mathbf{F}(t+\tau)) = (\mathbf{I}_{nd} + \tau\mathbf{W} \otimes \bar{\mathbf{A}})\mathrm{vec}(\mathbf{F}(t))$$

which in particular leads to

$$\mathrm{vec}(\mathbf{F}(m\tau)) = (\mathbf{I}_{nd} + \tau\mathbf{W} \otimes \bar{\mathbf{A}})^m\mathrm{vec}(\mathbf{F}(0)).$$

We can then write explicitly the solution as

$$\mathrm{vec}(\mathbf{F}(m\tau)) = \sum_{r,i} \left(1 + \tau\lambda_r^{\mathbf{W}}(1 - \lambda_i^{\boldsymbol{\Delta}})\right)^m c_{r,i}(0)\boldsymbol{\phi}_r^{\mathbf{W}} \otimes \boldsymbol{\phi}_i^{\boldsymbol{\Delta}}.$$

We now verify that by assumption in eq. (13) the dominant term of the solution is the projection into the eigenspace associated with the eigenvalue $\rho_- = |\lambda_-^{\mathbf{W}}|(\rho_{\boldsymbol{\Delta}} - 1)$. The following argument follows the same structure in the proof of Proposition 3.1 with the extra condition given by the step-size. First, we note that for any r such that $\lambda_r^{\mathbf{W}} > 0$, we have

$$|1 + \tau\rho_-| > |1 + \tau\lambda_+^{\mathbf{W}}| \geq |1 + \tau\lambda_r^{\mathbf{W}}(1 - \lambda_i^{\boldsymbol{\Delta}})|$$

since we required $\rho_- > \lambda_+^{\mathbf{W}}$ in eq. (13). Conversely, if $\lambda_r^{\mathbf{W}} < 0$, then

$$|1 + \tau\lambda_r^{\mathbf{W}}(1 - \lambda_i^{\boldsymbol{\Delta}})| \leq \max\{|1 + \tau\rho_-|, |1 + \tau\lambda_-^{\mathbf{W}}|\}$$

Assume that $\tau|\lambda_-^{\mathbf{W}}| > 1$, otherwise there is nothing to prove. Then $|1 + \tau\rho_-| > \tau|\lambda_-^{\mathbf{W}}| - 1$ if and only if

$$\tau|\lambda_-^{\mathbf{W}}|(2 - \rho_{\boldsymbol{\Delta}}) < 2,$$

which is precisely the right inequality in eq. (13). We can then argue exactly as in the proof of Proposition 3.1 to derive that for each index r such that $\lambda_r^{\mathbf{W}} < 0$ and $\lambda_r^{\mathbf{W}} \neq \lambda_-^{\mathbf{W}}$, then

$$|1 + \tau\lambda_r^{\mathbf{W}}(1 - \lambda_i^{\boldsymbol{\Delta}})| \leq \max\{|1 + \tau|\lambda_{-,2}^{\mathbf{W}}|(\rho_{\boldsymbol{\Delta}} - 1)|, |1 + \tau|\lambda_-^{\mathbf{W}}|(\lambda_{n-2}^{\boldsymbol{\Delta}} - 1)|\}$$

with $\lambda_{-,2}^{\mathbf{W}}$ and $\lambda_{n-2}^{\boldsymbol{\Delta}}$ defined in eq. (23). We can then introduce

$$\delta_{\mathrm{HFD}} := \max\{\lambda_+^{\mathbf{W}}, \rho_- - |\lambda_-^{\mathbf{W}}|\mathrm{gap}(\rho_{\boldsymbol{\Delta}}\mathbf{I} - \boldsymbol{\Delta}), \rho_- - (\rho_{\boldsymbol{\Delta}} - 1)\mathrm{gap}(|\lambda_-^{\mathbf{W}}|\mathbf{I} + \mathbf{W}), |\lambda_-^{\mathbf{W}}| - \frac{2}{\tau}\} \quad (28)$$

and conclude that

$$\begin{aligned}
\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t)) &= \frac{1}{2}\sum_{r,i} \left(1 + \tau\lambda_r^{\mathbf{W}}(1 - \lambda_i^{\boldsymbol{\Delta}})\right)^{2m} c_{r,i}^2(0)\lambda_i^{\boldsymbol{\Delta}} \\
&= (1 + \tau\rho_-)^{2m}\left(\frac{\rho_{\boldsymbol{\Delta}}}{2}\sum_{r:\lambda_r^{\mathbf{W}} = \lambda_-^{\mathbf{W}}} c_{r,\rho_{\boldsymbol{\Delta}}}^2(0) + \mathcal{O}\left(\left(\frac{1 + \tau\delta_{\mathrm{HFD}}}{1 + \tau\rho_-}\right)\right)^{2m}\sum_{i,r:\lambda_r^{\mathbf{W}}(1-\lambda_i^{\boldsymbol{\Delta}}) \neq \rho_-} c_{r,i}^2(0)\lambda_i^{\boldsymbol{\Delta}}\right) \\
&= (1 + \tau\rho_-)^{2m}\left(\frac{\rho_{\boldsymbol{\Delta}}}{2}||P_{\mathbf{W}}^{\rho_-}\mathbf{F}(0)||^2 + \mathcal{O}\left(\left(\frac{1 + \tau\delta_{\mathrm{HFD}}}{1 + \tau\rho_-}\right)^{2m}\right)\right).
\end{aligned}$$

In particular, we can normalize the solution and due to $(\mathbf{I}_d \otimes \mathbf{\Delta}) P^{\rho_-}_{\mathbf{W}} \mathrm{vec}(\mathbf{F}(0)) = \rho_{\mathbf{\Delta}} P^{\rho_-}_{\mathbf{W}} \mathrm{vec}(\mathbf{F}(0))$, we complete the proof for the case with residual connection.

If instead we drop the residual connection and simply consider $\dot{\mathbf{F}}(t) = \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W}$, then

$$\mathrm{vec}(\mathbf{F}(m\tau)) = (\tau \mathbf{W} \otimes \bar{\mathbf{A}})^m \mathrm{vec}(\mathbf{F}(0)).$$

Since G is not bipartite, the Laplacian spectral radius satisfies $\rho_{\mathbf{\Delta}} < 2$. Therefore, for each pair of indices (r, i) we have the following bound:

$$|\lambda^{\mathbf{W}}_r (1 - \lambda^{\mathbf{\Delta}}_i)| \leq \max\{\lambda^{\mathbf{W}}_+, |\lambda^{\mathbf{W}}_-|\},$$

and the inequality becomes strict if $i > 0$, i.e. $\lambda^{\mathbf{\Delta}}_i > 0$. The eigenvalues $\lambda^{\mathbf{W}}_+$ and $\lambda^{\mathbf{W}}_-$ are attained along the eigenvectors $\boldsymbol{\phi}^{\mathbf{W}}_+ \otimes \boldsymbol{\phi}^{\mathbf{\Delta}}_0$ and $\boldsymbol{\phi}^{\mathbf{W}}_- \otimes \boldsymbol{\phi}^{\mathbf{\Delta}}_0$ respectively. Accordingly, the dominant terms of the evolution lie in the kernel of $\mathbf{I}_d \otimes \mathbf{\Delta}$, meaning that for any $\mathbf{F}_0$ with non-zero projection in $\ker(\mathbf{I}_d \otimes \mathbf{\Delta})$ – which is satisfied by all initial conditions except those belonging to a lower dimensional subspace – the dynamics is LFD. In fact, without loss of generality assume that $|\lambda^{\mathbf{W}}_-| > \lambda^{\mathbf{W}}_+$, then

$$\begin{aligned}\mathrm{vec}(\mathbf{F}(m\tau)) = |\lambda^{\mathbf{W}}_-|^m \sum_{r:\lambda^{\mathbf{W}}_r = \lambda^{\mathbf{W}}_-} (-1)^m c_{r,0}(0) \boldsymbol{\phi}^{\mathbf{W}}_- \otimes \boldsymbol{\phi}^{\mathbf{\Delta}}_0 \\ + |\lambda^{\mathbf{W}}_-|^m \left(\mathcal{O}(\varphi(m)) \left(\mathbf{I}_{nd} - \sum_{r:\lambda^{\mathbf{W}}_r = \lambda^{\mathbf{W}}_-} (\boldsymbol{\phi}^{\mathbf{W}}_- \otimes \boldsymbol{\phi}^{\mathbf{\Delta}}_0)(\boldsymbol{\phi}^{\mathbf{W}}_- \otimes \boldsymbol{\phi}^{\mathbf{\Delta}}_0)^\top \right) \mathrm{vec}(\mathbf{F}(0)) \right),\end{aligned}$$

with $\varphi(m) \to 0$ as $m \to \infty$, which completes the proof. □

Gradient flow as spectral GNNs. We finally discuss eq. (11) from the perspective of spectral GNNs as in [2]. Let us assume that $\beta = 0$, $\mathbf{\Omega} = \mathbf{0}$. If we let $\mathbf{\Delta} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be the eigendecomposition of the graph Laplacian and $\{\lambda^{\mathbf{W}}_r\}$ be the spectrum of $\mathbf{W}$ with associated orthonormal basis of eigenvectors given by $\{\phi^{\mathbf{W}}_r\}$, and we introduce $\mathbf{z}^r(t) : \mathsf{V} \to \mathbb{R}$ defined by $z^r_i(t) = \langle \mathbf{f}_i(t), \phi^{\mathbf{W}}_r \rangle$, then we can rewrite the discretized gradient flow as

$$\mathbf{z}^r(t + \tau) = \mathbf{U}(\mathbf{I} + \tau \lambda^{\mathbf{W}}_r (\mathbf{I} - \mathbf{\Lambda}))\mathbf{U}^\top \mathbf{z}^r(t) = \mathbf{z}^r(t) + \tau \lambda^{\mathbf{W}}_r \bar{\mathbf{A}} \mathbf{z}^r(t), \quad 1 \leq r \leq d. \tag{29}$$

Accordingly, for each projection into the r-th eigenvector of $\mathbf{W}$, we have a spectral function in the graph frequency domain given by $\lambda^{\mathbf{\Delta}} \mapsto 1 + \tau \lambda^{\mathbf{W}}_r (1 - \lambda^{\mathbf{\Delta}})$. If $\lambda^{\mathbf{W}}_r > 0$ we have a *low-pass* filter while if $\lambda^{\mathbf{W}}_r < 0$ we have a *high-pass* filter. Moreover, we see that along the eigenvectors of $\mathbf{W}$, if $\lambda^{\mathbf{W}}_r < 0$ then the dynamics is equivalent to flipping the sign of the edge weights, which offers a direct comparison with methods proposed in [4, 45] where some 'attentive' mechanism is proposed to learn negative edge weights based on feature information.

The previous equation simply follows from

$$\begin{aligned} z^r_i(t + \tau) &= \langle \mathbf{f}_i(t + \tau), \boldsymbol{\phi}^{\mathbf{W}}_r \rangle = \langle \mathbf{f}_i(t) + \mathbf{W}(\bar{\mathbf{A}}\mathbf{f}(t))_i, \boldsymbol{\phi}^{\mathbf{W}}_r \rangle \\ &= z^r_i(t) + \lambda^{\mathbf{W}}_r \sum_j \bar{a}_{ij} z^r_j(t), \end{aligned}$$

which concludes the derivation of eq. (29).

D Additional details on experiments

D.1 Additional details on GRAFF

Given a gradient flow dynamical system of the form $\mathbf{F}(t + \tau) = \mathbf{F}(t) + \tau \bar{\mathbf{A}} \mathbf{F}(t) \mathbf{W}$, the vectorized solution is

$$\mathrm{vec}(\mathbf{F}(m\tau)) = \sum_{r,i} \left(1 + \tau \lambda^{\mathbf{W}}_r (1 - \lambda^{\mathbf{\Delta}}_i)\right)^m c_{r,i}(0) \boldsymbol{\phi}^{\mathbf{W}}_r \otimes \boldsymbol{\phi}^{\mathbf{\Delta}}_i.$$

We then see that the number of layers m – which coincides with the quotient of the integration time T by the step size τ – represents the degree of the polynomial computing the solution. More

precisely, on a heterophilic graph for which a HFD dynamics is more suited than an LFD dynamics, the negative eigenvalues of $\mathbf{W}$ are needed to magnify the graph high-frequencies. This in turn yields terms $\left(1+\tau\lambda_r^{\mathbf{W}}(1-\lambda_i^{\mathbf{\Delta}})\right)^m$ that would become unbounded with m growing if there is sufficient mass on the negative side of $\mathrm{spec}(\mathbf{W})$. On the other hand, terms associated with positive eigenvalues of $\mathbf{W}$ would quickly lead to over-smoothing. One then expects that on heterophilic graphs the degree m of the polynomial – i.e. the *number of layers* – should be generally smaller than that on homophilic graphs. This is confirmed in our real-world experiments where on the larger heterophilic graphs like Squirrel and Chameleon the optimal number m is an integer in $\{2, 3, 4\}$.

D.2 General Experimental details

GRAFF is implemented in PyTorch [53], using PyTorch geometric [54] and torchdiffeq [12]. Code and instructions to reproduce the experiments are available on GitHub. Hyperparameters were tuned using wandb[55] and random grid search. Experiments were run on AWS p2.8xlarge machines, each with 8 Tesla V100-SXM2 GPUs.

D.3 Additional details on synthetic ablation studies:

The synthetic Cora dataset is provided by [51, Appendix G]. They use a modified preferential attachment process to generate graphs for target levels of homophily. Nodes, edges and features are sampled from Cora proportional to a mix of class compatibility and node degree resulting in a graph with the required homophily and appropriate feature/label distribution. To validate the provided data before use we provide table 2 summarising the properties of the synthetic Cora dataset. All rows/levels of homophily have the same number of nodes (1,490), edges (5,936), features (1,433) and classes (5).

homophily	max_degree	min_degree	av_degree	density	edge_homoph	node_homoph
0.00	84.33	1.67	3.98	0.0027	0.00	0.00
0.10	71.33	2.00	3.98	0.0027	0.10	0.10
0.20	73.33	1.67	3.98	0.0027	0.20	0.20
0.30	70.00	2.00	3.98	0.0027	0.29	0.30
0.40	77.67	2.00	3.98	0.0027	0.39	0.39
0.50	76.33	2.00	3.98	0.0027	0.49	0.49
0.60	76.00	1.67	3.98	0.0027	0.59	0.60
0.70	67.67	2.00	3.98	0.0027	0.70	0.70
0.80	58.00	1.67	3.98	0.0027	0.78	0.79
0.90	58.00	1.67	3.98	0.0027	0.89	0.89
1.00	51.00	2.00	3.98	0.0027	1.00	1.00

Table 2: Summary of properties of synthetic Cora dataset

As well as the ablation shown in fig. 2 we used this dataset to perform an ablation using GCN as the baseline. We asses the impact of each of the steps necessary to augment a standard GCN model to GRAFF. This involves 5 steps; 1) add an encoder/decoder. 2) add a residual connection. 3) share the weights of $\mathbf{W}$ and $\mathbf{\Omega}$ across time/layers. 4) symmetrize $\mathbf{W}$ and $\mathbf{\Omega}$. 5) remove the non-linearity between layers. The results are shown in fig. 3 and corroborate Theorem 4.3 that adding a residual term is beneficial especially in low-homophily scenarios. We also note augmentations 3,4 and 5 are not "costly" in terms of performance.

D.4 Additional details on real-world ablation studies

For the real-world experiments in table 1 we performed 10 repetitions over the splits taken from [31]. For all datasets we used the largest connected component (LCC) apart from Citeseer where the 5th and 6th split are LCC and others require the full dataset. For Chameleon and Squirrel we added self loops and made the edges undirected as a preprocessing step. All other datasets are provided as undirected but without self loops. Each split uses 48/32/20 of nodes for training, validation and test set respectively. Table 3 summarises each of the datasets.

We used the real-world datasets to perform 2 ablation studies. First we choose 2 heterophilic datasets (Chameleon, Squirrel) and 2 homophilic (Cora, Citeseer) and observed how the size of the hidden

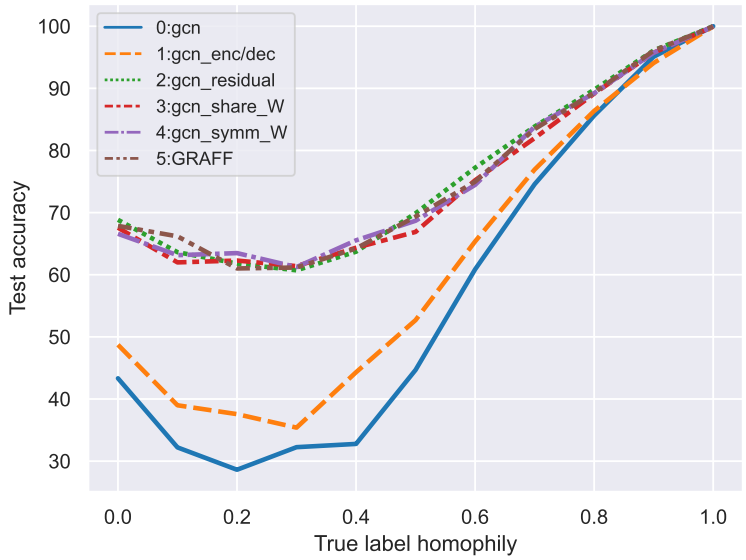

Figure 3: Experiments on synthetic Cora - GCN ablation

dataset	nodes	edges	features	classes	max_degree	min_degree	av_degree	density	edge_homoph	node_homoph
Texas	183	558	1,703	5	104	1	3.05	0.0167	0.06	0.06
Wisconsin	251	900	1,703	5	122	1	3.59	0.0143	0.18	0.16
Cornell	183	554	1,703	5	94	1	3.03	0.0165	0.3	0.3
Film	7,600	53,318	932	5	1,303	1	7.02	0.0009	0.22	0.22
Squirrel	5,201	401,907	2,089	5	1,904	2	77.27	0.0149	0.23	0.29
Chameleon	2,277	65,019	2,325	5	733	2	28.55	0.0125	0.26	0.33
Citeseer *	3,327	9,104	3,703	6	99	0	2.74	0.0008	0.74	0.71
Citeseer	2,120	7,358	3,703	6	99	1	3.47	0.0016	0.73	0.71
Pubmed	19,717	88,648	500	3	171	1	4.5	0.0002	0.8	0.79
Cora	2,485	10,138	1,433	7	168	1	4.08	0.0016	0.8	0.81

Table 3: Summary of properties of real-word datasets. All LCC except *

dimension effected performance for the structures of $\mathbf{W}$ described in section 5. For heterophilic datasets we used the splits from [31]. For homophilic datasets we used the methodology in [52], each split randomly selects 1,500 nodes for the development set, from the development set 20 nodes for each class are taken as the training set, the remainder are allocated as the validation set. The remaining nodes outside of the development set are used as the test set. This gives a lower percentage (3-6%) of training nodes. This approach was taken because less training information is needed in the homophilic setting and performance can become less sensitive to other factors, meaning less signal from the controlled variable. From fig. 4 we see that (DD) is more parameter efficient than *sum* in the heterophilic setting and (D) (a parameter light configuration) outperforms in the homophilic setting.

The second ablation study further corroborates the behaviour seen in fig. 2. We tested the structures of $\mathbf{W}$ against the real-world datasets with known homophily, again *neg-prod* outperforms *prod* in the heterophilic setting and vice-versa due the sign of their spectra.

dataset	neg_prod	prod	sum
Chameleon	67.32	58.86	68.36
Squirrel	51.39	42.11	51.29
Cora	31.80	79.65	81.17
Citeseer	32.47	67.31	67.53

Table 4: Ablation with controlled spectrum of $\mathbf{W}$ on real-world datasets

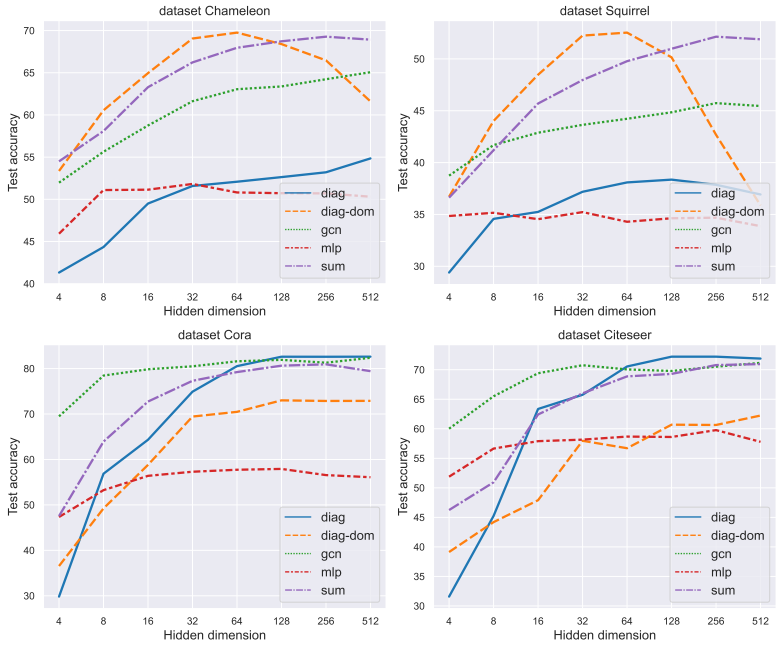

Figure 4: Ablation against hidden dimension

To validate the complexity analysis in Section 5 we performed a runtime ablation for the models between standard GCN and GRAFF described in the GCN ablation Figure 3. The average inference runtime over 100 runs for 1 split of Cora was recorded. We also include runtimes for the provided dense and sparse implementations of GGCN [45]. Adding the encoder/decoder (step 1) speeds up the model due to dimensionality reduction. Subsequent steps also reduce complexity and offer speedup with GRAFF performing the fastest.

D.5 Details on hyperparameters

Using wandb [55] we performed a random grid search with uniform sampling of the continuous variables. We provide the hyperparameters that achieved the best results from the random grid search in table 5. An implementation that uses these hyperparameters is available in the provided code with hyperparameters provided in `graff_params.py`. Input dropout and dropout are the rates applied to the encoder/decoder respectively *with no dropout applied in the ODE block*.

	w_style	lr	decay	dropout	input_dropout	hidden_dim	time	step_size
chameleon	diag_dom	0.0014	0.0004	0.37	0.43	64	3.2	1
squirrel	diag_dom	0.0058	0.0002	0.50	0.51	64	2.3	1
texas	diag_dom	0.0041	0.0354	0.33	0.39	64	0.6	0.5
wisconsin	diag	0.0029	0.0318	0.37	0.37	64	2.1	0.5
cornell	diag	0.0021	0.0184	0.30	0.44	64	2.0	1
film	diag	0.0026	0.0130	0.48	0.42	64	1.5	1
Cora	diag	0.0026	0.0413	0.34	0.53	64	3.0	0.25
Citeseer	diag	0.0001	0.0274	0.22	0.51	64	2.0	0.5
Pubmed	diag	0.0039	0.0003	0.42	0.41	64	2.6	0.5

Table 5: Selected hyperparameters for real-world datasets

E Elementwise non-linear activations and energy dissipation

In this section we investigate how to extend the energy framework to include more conventional non-linear activation maps to potentially have better expressive power – note that recent works like

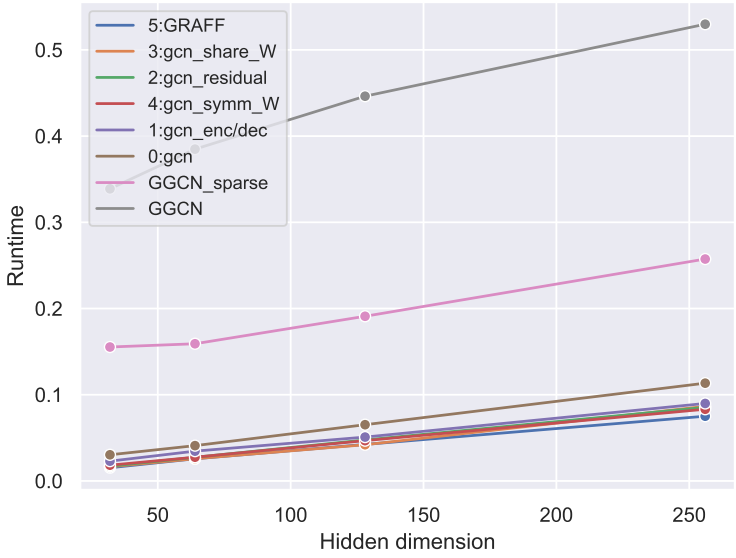

Figure 5: Runtime ablation for inference on Cora dataset

[6] in some sense argue for more non-linear layers. Namely, we consider the general energy in equation 7 with the inclusion of the source term:

$$\mathcal{E}^{\text{tot}}(\mathbf{F}) = \frac{1}{2}\sum_i \langle \mathbf{f}_i, \mathbf{\Omega}\mathbf{f}_i\rangle - \frac{1}{2}\sum_{i,j}\bar{a}_{ij}\langle \mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle + \beta\langle \mathbf{F}, \mathbf{F}(0)\rangle.$$

Since the energy is *quadratic*, its gradient flow is a linear dynamical system:

$$\dot{\mathbf{F}}(t) = -\mathbf{F}(t)\mathbf{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} - \beta\mathbf{F}(0) \tag{30}$$

The key question we explore here is: **what happens if we activate the equations with a pointwise non-linear map** σ? In general, we will not be a gradient flow of the energy $\mathcal{E}^{\text{tot}}$, however can we still say something about the behaviour of $t \mapsto \mathcal{E}^{\text{tot}}(\mathbf{F}(t))$ along the solution? The answer is affirmative and offers a novel contribution where even non-linear, residual, graph convolutional models maintain the intepretation of dissipating an energy where $\mathbf{W}$ plays the role of an edge-wise bilinear potential generating attraction and repulsion:

Proposition E.1. *Consider a non-linear map* $\sigma : \mathbb{R} \to \mathbb{R}$ *such that the function* $x \mapsto x\sigma(x) \geq 0$. *If* $t \mapsto \mathbf{F}(t)$ *solves the equation*

$$\dot{\mathbf{F}}(t) = \sigma\left(-\mathbf{F}(t)\mathbf{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} - \beta\mathbf{F}(0)\right),$$

where σ *acts elementwise, then*

$$\frac{d\mathcal{E}^{\text{tot}}(\mathbf{F}(t))}{dt} \leq 0.$$

Proof. The argument is simple and derives from direct computation. Namely, let us use the Kronecker product formalism to rewrite the gradient $\nabla_{\text{vec}(\mathbf{F})}\mathcal{E}^{\text{tot}}(\mathbf{F})$ as a vector in $\mathbb{R}^{nd}$: explicitly, we get

$$\nabla_{\text{vec}(\mathbf{F})}\mathcal{E}^{\text{tot}}(\mathbf{F}) = (\mathbf{\Omega}\otimes\mathbf{I}_n - \mathbf{W}\otimes\bar{\mathbf{A}})\text{vec}(\mathbf{F}) + \beta\text{vec}(\mathbf{F}(0)).$$

It follows then that

$$\begin{aligned}\frac{d\mathcal{E}^{\text{tot}}(\mathbf{F}(t))}{dt} &= \left(\nabla_{\text{vec}(\mathbf{F})}\mathcal{E}^{\text{tot}}(\mathbf{F}(t))\right)^\top \text{vec}(\dot{\mathbf{F}}(t)) = \\ &= \left((\mathbf{\Omega}\otimes\mathbf{I}_n - \mathbf{W}\otimes\bar{\mathbf{A}})\text{vec}(\mathbf{F}(t)) + \beta\text{vec}(\mathbf{F}(0))\right)^\top \sigma\left((-\mathbf{\Omega}\otimes\mathbf{I}_n + \mathbf{W}\otimes\bar{\mathbf{A}})\text{vec}(\mathbf{F}(t)) - \beta\text{vec}(\mathbf{F}(0))\right)\end{aligned}$$

If we introduce the notation $\mathbf{Z}(t) = (-\boldsymbol{\Omega} \otimes \mathbf{I}_n + \mathbf{W} \otimes \bar{\mathbf{A}})\mathrm{vec}(\mathbf{F}(t)) - \beta \mathrm{vec}(\mathbf{F}(0))$, then we can rewrite the derivative as

$$\frac{d\mathcal{E}^{\mathrm{tot}}(\mathbf{F}(t))}{dt} = -\mathbf{Z}(t)^\top \sigma(\mathbf{Z}(t)) = -\sum_{\alpha} \mathbf{Z}(t)^\alpha \sigma(\mathbf{Z}(t)^\alpha) \leq 0$$

by assumption on σ, which completes the proof. □

Important consequence: The previous results shows that even if the non-linear dynamical system

$$\dot{\mathbf{F}}(t) = \sigma\left(-\mathbf{F}(t)\boldsymbol{\Omega} + \bar{\mathbf{A}}\mathbf{F}(t)\mathbf{W} + \beta\mathbf{F}(0)\right),$$

is not a gradient flow for $\mathcal{E}^{\mathrm{tot}}$, the latter quantity is still decreasing along the solution meaning that the interpretation of positive (negative) eigenvalues of $\mathbf{W}$ inducing attraction (repulsion) persists given that *the energy has not changed*. **This allows us to derive that general (non-linear) graph convolutional models retain the learnable multi-particle energy-dissipation property provided that the channel-mixing matrices are symmetric and that the pointwise activation satisfies** $x\sigma(x) \geq 0$, which for example holds for $\mathrm{ReLU}, \tanh, \arctan$ and so on. *In particular, models that are energy-dissipating can fit non-linear activations*.

To further support the principle that the effects induced by $\mathbf{W}$ are similar even in this non-linear setting, we consider a simplified scenario.

Lemma E.2. *If we choose* $\boldsymbol{\Omega} = \mathbf{W} = \mathrm{diag}(\boldsymbol{\omega})$ *with* $\omega^r \leq 0$ *for* $1 \leq r \leq d$ *and* $\beta = 0$ *i.e.* $t \mapsto \mathbf{F}(t)$ *solves the dynamical system*

$$\dot{\mathbf{F}}(t) = \sigma\left(-\boldsymbol{\Delta}\mathbf{F}(t)\mathrm{diag}(\boldsymbol{\omega})\right),$$

with $x\sigma(x) \geq 0$*, then the standard graph Dirichlet energy satisfies*

$$\frac{d\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t))}{dt} \geq 0.$$

Proof. This again simply follows from directly computing the derivative:

$$\begin{aligned}\frac{d\mathcal{E}^{\mathrm{Dir}}(\mathbf{F}(t))}{dt} &= \frac{1}{4}\frac{d}{dt}\left(\sum_{r=1}^{d}\sum_{(i,j)\in\mathsf{E}}\left(\frac{f_i^r(t)}{\sqrt{d_i}} - \frac{f_j^r(t)}{\sqrt{d_j}}\right)^2\right) \\ &= \sum_{r=1}^{d}\sum_{i\in\mathsf{V}}(\boldsymbol{\Delta}\mathbf{f}^r)_i\sigma(-\omega^r(\boldsymbol{\Delta}\mathbf{f}^r)_i) = \sum_{r=1}^{d}\sum_{i\in\mathsf{V}}(\boldsymbol{\Delta}\mathbf{f}^r)_i\sigma(|\omega^r|(\boldsymbol{\Delta}\mathbf{f}^r)_i) \geq 0.\end{aligned}$$

Important consequence: The previous Lemma implies that even with non-linear activations, negative eigenvalues of the channel-mixing induce repulsion and indeed the solution becomes less smooth as measured by the classical Dirichlet Energy increasing along the solution. Generalising this result to more arbitrary choices is not immediate and we reserve this for future work.

□