Author Response - Submission 7862 (NeurIPS 2020): We thank the reviewers for their valuable comments and for recommending acceptance. Reviewer #1 says *"it's a novel approach"* and *"a step in the right direction"*. Reviewer #2 says it *"connects ideas [...] in a new and interesting way"* and *"is solidly grounded in the theory behind scattering transforms"*. Reviewer #3 says it has *"good motivation and efficient and elegant solution"* and *"the experimental methodology seems quite nice"*. Reviewer #4 says it *"identifies an innovative solution to an important limitation of GCNs"* and mentions *"the experiment on reducing the amount of labelled nodes is very informative"*. All typos and minor comments will be fixed in the final version, and the addition of a 9th page will be used to address comments about the paper being a bit packed (albeit well written). Further questions and concerns are addressed in the following:

Reviewer #1: *"the additional information sketched by the two lemmas in section 7 are not verified empirically"* **Reply:** We will add concrete examples of (colored) cyclic and bipartite graphs to demonstrate and verify the properties shown in these lemmas. ⧫ *"Scattering Transforms typically average over high frequencies (to achieve stability) and are known to lose high-frequency information."* **Reply:** Respectfully, this statement is inaccurate. One of the main strengths of traditional scattering is its ability to capture and aggregate high frequency information thanks to the demodulation provided by the complex modulus (while losing the phase). This has been demonstrated in multiple applications involving image and audio textures, audio source separation, and inverse problems (all critically depending on high frequency information). The analogy in the graph case is not perfect (e.g., due to ill-defined demodulation), but we will add further discussion clarifying this point in the final version. ⧫ *"Have you considered comparing with [1]"* **Reply:** The arxiv version of [1] was first posted on June 10 (then published in ICML in July), and thus not available before the NeurIPS submission deadline. Nevertheless, we will add discussion referring to its alternative GFT and (relaxed) wavelet construction. However, direct comparison is not applicable here, since it does not provide node-level features, but rather graph-level ones, while considering classification of multiple signals over a single graph, rather than nodes on it. ⧫ *"Can you include performance of graph scattering methods in Tables 1-2."* **Reply:** To our knowledge, Zou & Lerman (ACHA 2019) is the only previous graph scattering work reporting node classification (others focus solely on graph-level tasks). Their paper only reports 81.9% on Cora, below both GAT and our method, but their GitHub code achieves 69.4% on DBLP (better than GAT, but still significantly below our method) and underperforms even GCN on Citeseer (67.5%) and PubMed (69.8%). We will include these results in the tables, as requested. Moreover, we will extend the ablation study in the supplement to further emphasize the impact of bandpass information in scattering channels compared to lowpass GCN ones. It should be noted that since the addition of bandpass information is the main distinction between our method and others in Sec. 8, the impact of such information is evidently nonnegligible.

Reviewer #2: *"[line 224] is ambiguous and needs to be clarified"* **Reply:** We will revise this sentence to clarify the "completely lazy" random walk is the stationary one (only having self-loops) represented by the identity, while the nonresting one R contains no self-loops. ⧫ *"How is the residual layer helping in this architecture and why? [...] Why do we need the extra threshold on the wavelets in Fig 1e?"* **Reply:** The motivation for this layer is discussed in Sec. 6 and its impact is verified empirically in the ablation study in the supplement. Briefly, the wavelet cascade in scattering extracts high frequencies and may capture undesirable features (e.g., distinguishing labeled and unlabeled nodes). The residual convolution alleviates such artifacts while maintaining the localization of node features. We note that traditional scattering transforms also typically use a final smoothing step (e.g., via lowpass or moments), as pointed out also by R1.

Reviewer #3: We thank the reviewer for the detailed feedback on the notations. We will make sure to clarify every mentioned instance. ⧫ *"l121: '$h_j^l = \ldots$': j doesn't appear in the [R]HS?"* **Reply:** The reviewer is correct - there was a missing index j on θ^ℓ_{ij}, which are the individual weights in the matrix Θ^ℓ (see Eq. 2, lines 125-126). We will fix this missing index in all instances including the one noticed by the reviewer and two others in lines 122-123. ⧫ *"[...] notations [w]ere often misleading and inconsistent with prior work on scattering"* **Reply:** We note that we largely tried to follow prior geometric scattering notations (e.g., Gao et al. 2019), while bridging certain discrepancies with notations used in GCN work. That being said, we acknowledge the reviewer's comment that our use of Φ and J may have been somewhat misleading. We will therefore rename Φ to U, which is sometimes used for the wavelet cascade (albeit here without the outermost nonlinearity that is added later in lines 179-180 – we will clarify this slight abuse of notation in a footnote), and rename J to p, following the notation from Mallat 2012 (Definition 2.2, page 10 there) for the scattering pathways. We will also make an effort to verify no other notation conflicts remain in the paper, which have not been explicitly pointed out by the reviewers, and we will revise the notations to resolve them as necessary.

Reviewer #4: *"I'd be interested to hear whether the proposed approach could also benefit from an attention mechanism similar to GAT."* **Reply:** While out of scope for this work, this is an excellent point, and indeed we are currently in the process of developing a scattering attention network that combines these ideas showing promising preliminary results. ⧫ *"[details regarding the] setup for the experiment where the training size is reduced"* **Reply:** In our experiments, we decrease the number of labeled nodes (i.e., "hiding" node labels) for training on the fixed graph while the nodes for validation and test are the same. In a low training size regime, reducing the training size means that the nearest labeled node can be further away. ⧫ *"A schematic illustration of the geometric scattering may help with the accessibility"* **Reply:** We agree and will add one, similar to Fig. 2 from Gao et al. 2019 or Fig. 1 from Gama et al. NeurIPS 2019.