We thank the reviewers for their thoughtful and constructive feedback. We are pleased that all reviewers [**R1**, **R2**, **R3**, **R4**] find the paper clear; most reviewers [**R2**, **R3**, **R4**] find the problem of overcoming the implicit homophily assumption in most GNN models well-motivated and vital; and [**R2**, **R3**, **R4**] value our theoretical analysis and the grounding of our methodology. Next, we first clarify the technical contributions of our work, and then address specific comments. While we only address major discussion points here, we will incorporate all feedback in the final version.

Recap: Technical contributions & Novelty. We empirically revealed the limitation of some widely-adopted GNNs to learn over *networks with heterophily*, and *identified* a set of key designs that actually are helpful. We showed the effectiveness of these designs under heterophily through theoretical analysis (§3.1.1 - 3.1.3) and ablation studies (§5.1, lines 298–321). While we acknowledge that these designs are used in existing methods, we are the first to revisit their effectiveness *in heterophily settings with in-depth theoretical justifications and extensive empirical evaluation* (this has been largely unknown before this work). Existing models have used subsets of these designs (and tested them under strong homophily), but not all at the same time (Table 2). Thus, our purpose in designing $\mathrm{H_2GCN}$ is to exemplify how an effective combination of these designs can help a GNN better adapt to the whole spectrum of low-to-high homophily, while avoiding interference with other designs. We'll revise our paper to clarify the scope of our contributions.

[R1, R3] Concerns that the proposed designs aren't novel as they're existing techniques. We are the first to discuss the importance of these designs *under heterophily* with novel theoretical justifications and extensive empirical evaluations. While we agree that the designs are not *new*, our analysis for the heterophily setting is *novel*. We believe that showing *what* works and *why* in a challenging, rarely-studied setting advances the field. We'll make this more clear.

[R1, R3] Sufficiency of baselines. Thanks to the identified designs, we were able to spot very competitive baselines under heterophily (e.g. GCN-Cheby, GraphSAGE), which were not compared against in recent state-of-the-art works (e.g. GeomGCN [20] in ICLR'20, against which we also compare). We have put considerable effort in ensuring an extensive, rigorous comparison. That said, we appreciate **R3**'s excellent suggestion to enhance the baselines with the jumping-knowledge (JK) connections, corresponding to design D3. We use JK-Concat [34] and report results for GraphSAGE, GCN-Cheby and GCN in Table R1. JK connections improve the baselines (for fixed number of layers) in some cases though without changing our observations. We'll discuss these results in detail in the final version.

Table R1: [**R3**] Additional baselines on real benchmarks (baselines + JK). Our observations remain largely the same.

	Texas	**Wisconsin**	**Actor**	**Squirrel**	**Chameleon**	**Cornell**	**Cora Full**	**Citeseer**	**Pubmed**	**Cora**
GraphSAGE+JK	$81.89{\pm 7.32}$	$83.14{\pm 4.45}$	$34.35{\pm 0.67}$	$40.84{\pm 1.54}$	$58.09{\pm 1.92}$	$77.03{\pm 4.08}$	$65.31{\pm 0.99}$	$75.91{\pm 1.09}$	$88.34{\pm 0.47}$	$86.24{\pm 1.21}$
GCN-Cheby+JK	$77.03{\pm 7.88}$	$81.18{\pm 4.55}$	$34.70{\pm 1.05}$	$40.90{\pm 2.78}$	$59.91{\pm 2.28}$	$71.62{\pm 9.47}$	$66.09{\pm 0.12}$	$74.19{\pm 1.69}$	$88.69{\pm 0.49}$	$84.91{\pm 1.98}$
GCN+JK	$66.49{\pm 6.64}$	$74.31{\pm 6.43}$	$34.26{\pm 0.90}$	$39.43{\pm 1.00}$	$62.70{\pm 1.98}$	$64.59{\pm 8.68}$	$64.73{\pm 0.30}$	$74.53{\pm 1.60}$	$88.45{\pm 0.49}$	$85.81{\pm 1.04}$

[R1, R3] Significance / stability of results on real data. These benchmarks show the *complexity* of learning from graphs with heterophily. Our main focus is *not* to optimize for high-homophily datasets like Cora, Citeseer, Pubmed [**R1**]; we include them to show the trends across the full spectrum of low-to-high homophily. While we agree that there is not a consistent winner for *all* the datasets, we have demonstrated that $\mathrm{H_2GCN}$ variants have the best *overall* performance across the spectrum in terms of the smallest average ranking. Another clear trend is that most models utilizing some of our identified (heterophily-friendly) designs outperform other models under heterophily; deviations are related to implementation details and other designs that may interfere with our identified designs.

[R1] "Graph neural nets of multiple layers are able to model the network heterophily." This is not the case: 2-layer GCN performs poorly under heterophily (cf. Table 4) and in general can suffer from oversmoothing[1]. **"Not clear how designs D2+D3 help in heterophily."** Removal of designs D2 and D3 leads to dramatic decrease in accuracy under heterophily, as shown in the ablation studies in Fig. 3(b)-3(c) (§ 5.1; theoretical justifications in §3.1.2-3.1.3).

[R2] Differences between $\mathrm{H_2GCN}$ and baselines. We discuss the differences from GCN in lines 639–647, and from GraphSAGE in lines 653–659 (Supp. §D.2). GraphSAGE generally has more learnable parameters than $\mathrm{H_2GCN}$—e.g., $\mathrm{H_2GCN}$-2 outperforms GraphSAGE in `syn-products` with less than $\frac{1}{5}$ of the parameters (10,880 vs. 59,648).

[R3] "Thm 1 only points out a limitation of one specific (though popular) GCN variant." This in fact illustrates the point of our work: there exist GNNs that happen to make the design choices we study, but also popular GNNs that do not. Without work to shed light on *why* GNNs should use particular designs, any success on heterophily is the result of a shot in the dark. **"The adj matrix is a low-pass filter" & "Aggregation over larger neighborhood sizes makes the filters more sensitive to high frequencies" are incorrect.** We agree that "high-order polynomials of the norm adj matrix correspond to low-pass filter" is more accurate; we'll reword this. However, we have not found the latter claim in our work. In lines 198–200 we say: "*intermediate outputs from earlier rounds contain higher-frequency components than ... later rounds*"; thus, D3 helps when higher-frequency information is beneficial (e.g., in heterophily).

[R4] "Small datasets ... technical challenges (e.g., high variance) when one attempts to scale the proposed method via neighborhood sampling" These are important future directions. Our paper calls for future work in designing large-scale benchmarks exhibiting heterophily, which will hopefully inspire methodological developments.

[1] Graph Neural Networks Exponentially Lose Expressive Power For Node Classification. ICLR 2020