We thank the reviewers for their positive feedback and comments. We will incorporate their suggestions into the next version of the paper. We address points made by each individual reviewer in turn.

Reviewer 1

- In the revision, we will present simplifications of Theorem 2 and Corollary 4. For instance, if we fix n, δ, and ϵ and choose η so that $N_\eta = 1$ (i.e., at least one model is η-similar to all others), then Theorem 2 bounds the number of testable models as
$$k \leq \frac{c_1}{(1-\eta)^{c_2}},$$
where $c_1, c_2 \geq 0$ are constants that depend on n, δ, and ϵ. Then, as the similarity η of the model collection grows, the number of testable models k grows as well.
- In Corollary 4, α quantifies the strength of the similarity assumption. For $\alpha = 0$, there is no similarity requirement, and the confidence interval is wide. However, as α grows, the similarity requirement becomes restrictive while the confidence interval becomes increasingly tight. We will make this dependence clear and simplify the stated bound.
- In the adaptive query case, the $(n+1)^{k-1}$ bound on the number of queries is standard in the adaptive data analysis literature. The bound is indeed conservative and reflects the worst-case behavior of the analyst. Our analysis helps illustrate why this worst-case behavior does not attain in practice. If the queries are similar, then the effective number of queries can be much smaller than $(n+1)^{k-1}$.

Reviewer 2

We begin with the main point of Reviewer 2. The reviewer is concerned that the paper's conclusions could be derived solely from simulation work rather than mathematical analysis. Simply put, **the y-axes of our plots heavily rely upon our theoretical contributions.** As a result, we are not aware of a way to derive our paper's conclusions purely from simulation. Since the review is not specific about what types of simulations would yield similar insights into overfitting, we would appreciate a clarification in the updated review.

To expand on this further: the bulk of the theoretical content is careful numerical calculations to obtain sharp generalization bounds where the improvement due to similar models is apparent on data from ImageNet and CIFAR-10. The standard union bound does not take advantage of model similarity; the refinement in equation (3) is the basis for our calculations that highlight the beneficial effect of model similarity. These sharp bounds allow us to empirically demonstrate that model similarity offers protection against overfitting on heavily used benchmark datasets. Consequently, the mathematical analysis is not merely an accompaniment to our empirical results, but an important part in their development.

Furthermore, generalization bounds of the type we present in Theorem 2 and Corollary 4 are a core focus in statistical learning theory. Our bounds demonstrate a link between similarity and protection against overfitting in both the non-adaptive and adaptive case. The settings considered are purposefully simple to highlight the main ideas, and, as suggested by Reviewer 1, we will further simplify the bounds to improve their pedagogic value. Finally, generalization bounds can often be vacuous when instantiated with concrete values from practical settings. A key contribution of our work is generalization bounds that provide meaningful numerical guarantees on both ImageNet and CIFAR-10 data.

We now address the remaining comments of Reviewer 2:

- We agree an important future direction is to expand the scope of benchmarks considered. We plan to conduct the same similarity analysis on both Kaggle competitions and tasks in natural language processing.
- Our analysis explicitly considers the case where models cluster into smaller sets with large η. The definition of an *η-similarity cover* is a clustering such that each cluster has similarity at least η. Figure 3 shows one such example: for $\eta \approx 0.8$, the models can be partitioned into roughly 200 clusters.
- Thank you for pointing out the typo for p_w. In lines 238 and 244, p_w should be replaced with $1 - p_w$. Otherwise, the definitions and subsequent discussion are unchanged.

Reviewer 3

- We agree an important direction for future work is exploring the extent to which our findings transfer to other settings. A concrete next step is evaluating model similarity on data from Kaggle competitions, which includes a diverse set of sample sizes, model classes, and data types. Nevertheless, image classification on ImageNet and CIFAR-10 is a natural starting place for this line of inquiry. Recent research has shown that both benchmarks do not suffer from adaptive overfitting despite almost a decade of intense activity, and model similarity offers an important piece of the explanation for this surprising phenomenon.
- We thank the reviewer for pointing out the lack of citations in the introduction. We will update the introduction with appropriate citations. We would appreciate any pointers to additional related work in the updated review.