We thank all reviewers for their insightful comments. Reviewer-specific comments follow.

Reviewer 1

“unclear how much...on real distributions": We will improve the writing in the revised version to explicitly note that our results are in the context of the proposed datasets. More importantly, note that the key design characteristic of our slab-structured datasets—*multiple independent features of varying predictive power and simplicity*—is motivated by recent empirical findings that differentially characterize learned features and desired features: syntactic cues vs. semantic meaning [27], non-robust vs. robust features [19] and statistical regularities vs. high-level concepts [21]. **Regarding Theorem 1**, Lemma F.3 actually provides a precise description about the classifier after the t^{th} mini-batch gradient descent step. We subsequently use this characterization to bound the train and test loss in Lemma F.2. **Section 4.3 and Appendix**: Thank you for your suggestions—we shall shorten section 4.3 and add a brief description of our results presented in the appendices in the revised version.

Reviewer 2

Role of initialization: Our results on simplicity bias are robust to the exact choice of *random* initialization. That said, we agree that understanding the tradeoff between (non-random) initialization and simplicity bias is an important next step—we plan to study this tradeoff by varying the extent to which the model (at initialization) is aligned with the complex slab features in the slab-structured and MNIST-CIFAR datasets. **Amended Cross Entropy**: Thank you for this suggestion. We are in the process of porting the open-sourced code (https://github.com/jhjacobsen/fully-invertible-revnet) to train invertible neural networks with amended cross entropy on the synthetic datasets and will add the results to Appendix E in the revised version. **Related work and Code**: Thank you for sharing relevant works. We will discuss connections to both papers—shortcut learning (arXiv:2004.07780) and concurrent work (arXiv:2006.12433)—in the revised version. We shall open-source our code and datasets as well.

Reviewer 3

Slab-structured datasets: Please see the response to Reviewer 1. Our focus on distilling empirical findings on real datasets into synthetic datasets clearly sets us apart from pathological worst-case data distributions used to prove no-free-lunch theorems. We believe that the slab-structured datasets can be used (a) as testbeds to develop algorithms that improve the robustness of neural networks and (b) to gain theoretical insights that cannot be obtained from “linearly separable" data [4], as they do not capture the failure modes of NNs observed in practice. Several papers have studied such “principled synthetic datasets" to obtain insights on initial learning rate [R1], adversarial examples [R2], long-term dependencies [R3], GAN dynamics [R4], generalization vs. robustness [R5], and bias in generative models [R6]. **Regarding overly general claims**: We would like to clarify our viewpoint: we show that simplicity bias is *an* important factor that *jointly*, but not exclusively, contributes to the adversarial vulnerability, poor OOD performance and suboptimal generalization (see lines 13, 92, 261, 296). We agree with reviewer 3—the paper should not give readers an impression that our claims hold for *all* datasets and for *every* setting. To address this misunderstanding, we will rectify our writing in the revised version to explicitly note that our empirical & theoretical results hold in the context of the proposed slab-structured and MNIST-CIFAR datasets. **Defining simplicity**: Regarding “natural notion of simplicity", our definition (line 161) is equivalent to the minimal width of one-hidden-layer ReLU NNs required to perfectly fit a given dataset. “Number of linear classifiers...not justified": The number of pieces in piecewise linear functions determines the VC dimension of this concept class (e.g., see [R7]). “mnist necessarily less complex than cifar10?" [R8] use spectrally normalized margin distributions to show that mnist is less complex than cifar10. Moreover, linear models trained on the mnist & cifar blocks individually attain 99.9% & 68.9% test accuracy respectively, even though the cifar block is almost fully predictive of its labels (line 251).

Reviewer 4

“...if drop-out or batch-norm have any influence on the obtained results": In Appendix C, we validate our results on extreme simplicity bias across architectures, activation functions, optimizers and regularization methods such as L2 regularization and dropout. We will add another section on the effect of batch-norm in the revised version.

[R1] Li, Yuanzhi, et al. "Towards explaining regularization effect of initial large learning rate in neural networks." NeurIPS (2019).

[R2] Gilmer, Justin, et al. "Adversarial spheres." arXiv:1801.02774 (2018).

[R3] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation (1997)

[R4] Li, Jerry, et al. "On the limitations of first-order approximation in gan dynamics." ICML (2018).

[R5] Raghunathan, Aditi, et al. "Understanding and mitigating the tradeoff between robustness and accuracy." ICML (2020).

[R6] Zhao, Shengjia, et al. "Bias and generalization in deep generative models: An empirical study." NeurIPS (2018).

[R7] Bartlett, P., et al. "Nearly-tight VC-dimension & pseudodimension bounds for piecewise linear neural networks." JMLR (2019).

[R8] Bartlett, P., et al.. "Spectrally-normalized margin bounds for neural networks." NeurIPS (2017).