We thank all reviewers for careful reading & positive comments, including **R1**: "that different algorithms can be categorized based on relatively simple metrics is surprising & interesting"; **R2**: "the results...are highly significant and novel and relevant to the NeurIPS community"; **R3**: "creative and thought-provoking approach which may inspire future other 'virtual experiments' of the kind"; **R4**: "this work has great potential for high impact in systems and computational neuroscience". We now address major reviewer concerns below. ★**How biological are the architectures, task, & learning rules evaluated ...? Why these particular choices? (R1,R2,R3,R4)**: We chose NN architecture types & training datasets that have been shown in comp. neurosci. literature to make good models of neural response patterns in primate electrophysiology & human fMRI data. We test learning rules that have competitive ML performance that cannot be ruled out by performance characteristics alone (e.g. simple hebbian rules). We use supervised & self-supervised learning objectives (without need for Imagenet category labels), & a range of different specific NN architectures, to model the fact that the loss function & architecture best suited to understanding a given brain area are generally partially, but far from exactly, known. Our work's goal is to identify statistics that will allow identification of the learning rule, *invariant* across the variability due to these types of unknowns. **R3**, good point about varying datasets / architecture classes. We've obtained results for shallow architectures with CIFAR-10 dataset – biologically, perhaps interpretable as expanding project scope to simpler *non-primate* (e.g. mouse) visual systems. We also have results for networks trained on *auditory* stimuli, using the AudioSet dataset – showing our approach holds across multiple sensory modalities with the *same* classifier. Will include these results in revision. We hope in the future to also broaden scope to e.g. RL models, as suggested by **R1**. ★**"... suspicious that [discrimination power] is driven by differences in Imagenet performance..." (R4)**: Important question. As shown in Fig S1, all learning rules except feedback alignment (FA) have high overlap in performance across hyperparameters; performance differences due to architecture swamp those from learning rule, e.g. FA aside, Alexnet with best learning rule performs $\ll$ Resnet-34 with worst learning rule. Thus, performance is a highly confounded indicator of learning rule, a key point we should have emphasized, so will move to main text as **R2** suggests. Also, we want to address experimental situations where performance is not directly measurable (animal behavior is often harder than e-phys!); & to allow for the possibility of unsupervised learning objectives not optimized for specific performance goals. Thus, it is important & nontrivial to identify features that are robust across architecture & objective fns, & have direct physical analogues in experimental measurement. ★**"The authors [show] that certain statistics are more informative than others ... not clear why this should be the case?" (R4)**: The primary intuition that certain aggregate statistics could be useful for separating learning rules comes from studying the learning dynamics of single layer perceptrons, where activation mean is a typical choice [eg. Werfel et al. 2004]. But in deep NNs, no theory yet allows us to derive optimal statistics mathematically, motivating our empirical approach. We thus included a variety of potential observables that might more robustly characterize non-linear network effects, & thus enable the classifier to *discount* differences when needed. Ideally in the future we can combine better theory with our method to sharpen feature design. We will improve discussion of this in revision. ★**"...neurons have recurrent dynamics, experiments here [only use feedforward] models... architecture is intrinsically tied to the learning algorithm!" (R4)**: We have tested our approach on recurrent convolutional models [Nayebi et al. 2018, Schrimpf et al. 2019] – just not at such large scale as the included results, since such networks are very resource-intensive. However, outcomes don't change conclusions at all, will include what we have in revision. Importantly: a main takeaway of our paper is that architecture is in some sense *not* necessarily intrinsically tied to the learning rule; otherwise, we would not have been able to reliably separate learning rules across the range of architectures considered. ★**"Relatedly, you use Adam & SGD+Momentum ...Discriminating learning rate seems like a different question of discriminating learning algorithms." (R4)**: First-order learning rules are basically characterized by two choices, namely how parameter updates are made as a function of (1) (high-dimensional) direction of gradient tensor, & (2) the magnitude of gradient tensor. Item (2) is directly tied to learning rate policy, & as adaptive methods can yield significant (if hard to predict) differences in trainability across various architectures & datasets, learning rate policy is an integral part of the learning rule. Our choice of candidate rules tested the ability of our approach to handle variation of *both* aspects. ★**"Does discriminability change when initializing from relatively good weights, rather than random?" (R4)**: While we're not exactly sure how to initialize from good weights in a task agnostic way (we used standard best practices for init), we *did* examine training the classifier solely on different portions of training trajectory, including only using late-time checkpoints after network performance stabilized – this somewhat approximates idea of using "good" weights. We found largely consistent results (Fig. S4). Interesting question for follow up work! ★**"Can a model trained with one set of hyperparameters generalize...?" (R4)**: In all reported results, we widely varied not only architecture & loss function, but also learning hyperparameters such as learning rates/regularizations (see supplement for details). We then considered two types of classifier accuracy evaluations. First, we performed standard cross-validation, e.g. random non-overlapping train/test splits. High accuracy here shows classifiers work across new mixed combinations of architecture, objective function, & learning hyperparameters. We also performed tests that held out entire classes of input types, to explore strong generalization. For example, we did architecture hold-outs, training on some architectures then testing on others, & our method still performed well in this crucial case (Fig 2b). Also see Figs. S2-3 for other such generalization tests. ★**Other comments (R1-R4)**: We cannot address all remaining comments due to space limitations, but will address them in revision, especially **R2**'s stylistic suggestions.