We thank the reviewers for their helpful comments and feedback. We answer their questions below.

Reviewer #2 and #3 mention a lack of theoretical or algorithmic novelty as the only weakness. We believe that rigorous large-scale empirical studies are as important as the introduction of new methods and theories. We would like to highlight the novelty of our work along other axes:

- We show, for the first time, that direct feedback alignment (DFA) can match backpropagation (BP) performance on challenging tasks. This is in contrast with past studies failing to scale DFA past simple datasets like MNIST.
- Our study is unprecedented in the variety of architectures and tasks considered, the hyper-parameter tuning effort undertaken for an alternative training method, and the breadth of the controls employed. Traditionally, alternative methods are seldom evaluated on tasks beyond image recognition with convolutional networks.
- Our implementation in the supplementary material scales to large architectures and demanding tasks, solving a practical issue encountered in previous work (Bartunov et al., 2018).

Reviewer #1 The reviewer first questions the rationale behind our choice of tasks. Our selection was built to represent a diverse set of topics representative of modern deep learning. Neural view synthesis is a challenging subset of implicit neural representations (see the recent Sitzmann et al., 2020), an active area of research rich with potential theoretical links for future studies. Recommender systems based on click-through rate (CTR) prediction are employed at-scale in the industry, and represent a real-world deep learning scenario. Beyond considerations on the applications, we also sampled varied architectures: deep fully-connected networks (MLP) with NeRF, hybrid models combining MLP with other techniques for CTR prediction, structured networks with graph convolutions, and finally attention-based architectures. We deliberately avoided tasks involving convolutions–as DFA has already been shown to be incompatible with them. However, we did not have expectations on the performance of DFA on our selection of tasks and architectures.

The reviewer then asks why DFA performs well in some tasks and not in others. The main factor influencing the performance of DFA is the architecture. The exact mechanics explaining the successes and shortcomings of DFA is a broader topic, requiring a paper of its own. We refrain from speculation, and we hope our survey can provide inspiration for further theoretical and empirical studies of DFA, by showing that convolutions are the exception, and not the norm.

The performance of DFA lags behind backpropagation more significantly in the natural language processing (NLP) task. Indeed, the Transformer is the most complex architecture considered in our paper: it requires much more careful tuning to train well. The training of Transformers is an open topic even for BP: practices like learning rate warm-up, cosine schedule, and RAdam are fairly new (Vaswani et al., 2017, Liu et al., 2019, Popel and Bojar, 2018). Fully adapting these principles to DFA requires a substantial amount of work, beyond the scope of a single submission. We are willing to rephrase the relevant part in the abstract in the camera-ready version. However, we can not find any controversial sentence in the introduction and, as noted by Reviewer #2, we state unambiguously that there remains a clear performance gap between BP and DFA in the NLP task in lines 266-268.

Reviewer #2 The missing abbreviations will be added in the camera-ready version of the paper.

Reviewer #3 Weight transport violations only occur in the Transformer architecture. Line 259 will be clarified and a mention of weight transport violation will be added in the legend of Table 5 in the camera-ready version. Also, we will include citations to Landsell et al. and Crafton et al. and fix the typo in equation 1. Finally, we share the enthusiasm for learning to learn with DFA, and hope our paper can motivate research in this direction.

Reviewer #4 The reviewer asks for details on the computational advantage offered by DFA. Assuming layers of similar sizes and sufficient parallel processing power, the speed-up factor over BP in the backward pass is equal to the number of layers, thanks to to to the parallelization of the backward pass. Furthermore, DFA reduces communication overhead where the model is spread across multiple devices. Advantages of this "backward unlocking" are laid out in Jaderberg, et al. 2016. We can't reduce this to an overall factor, as many specifics of the architecture come into play.

The reviewer inquires about the adaptation of DFA to architectures beyond the vanilla MLP used in Section 2. The strategy used for attention layers is presented in Appendix D, and adaptations to the random matrix for graphs and NLP are described in the Appendix C (lines 639 and 654). These will be added to the main text in the camera-ready version. More broadly, we introduce a random feedback $\mathbf{Be}$ after every non-linearity, in the spirit of Nøkland, 2016. We do not introduce any specific structure or operation to build the feedback. The method is applicable to any computational graph. We use a unique global feedback matrix (as in Launay et al., 2019), initialized from $\mathcal{U}(-1, 1)$, and normalized with the square root of the output dimension of every layer. We will add this detail in the camera-ready version. Regarding the necessity of the derivative of the activation, Gilmer et al., 2017 point to its importance in the update rule to perform better than a linear model. Concerning different activations functions, we stayed as close to the original models as possible (using ReLU), but Nøkland, 2016 shows that the method works for different choices of activation functions too.