We thank all reviewers for their comments. We will correct all typos and address all minor comments in the final paper.

**Reviewer #1:** Biological plausibility of C1-C3 and $\phi$? **C1:** Given an input-target pair, C1 means that a model makes a prediction from the input, before it learns from the target, which is quite natural for biological neural systems. **C2 and $\phi$:** C2 is realized by $\phi$, and it is very plausible that $\phi$ can be performed by biological neurons, because some types of neurons are well-known to respond predominantly to changes in their input [71]. **C3:** Though it is unnatural for biological neural systems to have a specific integration step, we show in the supplementary material that relaxing C3 results in BP with a different learning rate for different layers, which does not violate the core of BP.

Weight transport problem? In the predictive coding model, the errors are back-propagated by correct weights, because the model includes feedback connections that also learn. The weight modification rules for corresponding feedforward and feedback weights are the same, which ensures that they remain equal if initialized to equal values (see [48]).

Other criticisms for BP (spiking and recurrent neural networks) remain? As discussed in the related work section, they are left as future work, while this work addresses two of the most crucial ones: local plasticity and autonomy.

Related work of https://arxiv.org/abs/2006.04182? This work was published on 7 Jun 2020 (after the NeurIPS'20 deadline). It includes another approximation to BP but no equivalence; we will add it to the related work.

**Reviewer #2:** Weight transport problem? Please see line 8 in the response to Reviewer #1.

Locality of learning rule? (the update of $\theta^{l+1}_{i,j}$ depends on $x^l_i$)? According to Eq. 9, the update of $\theta^{l+1}_{i,j}$ actually depends on $x^{l+1}_j$ rather than $x^l_i$ (we suspect the reviewer might have misread Eq. 9), so it is local.

Confusion about inference? Alg. 2 explicitly states C1 in the second "Require". In line 158, we make a note that C1 is omitted in Fig. 2 for simplicity. This note will be included in the caption of Fig. 2.

Full autonomy during the initial convergence of prediction? During the prediction phase, the error nodes change due to feedforward input, while during learning, the error nodes change due to feedback input. We will clarify that, in order to prevent learning during prediction, $\phi$ is equal to 1 only if the change in error node is caused by feedback input.

**Reviewer #3:** Limitations of PCNs. They are well-discussed in [10,48]; we will add a summary of them. Note that some limitations have been addressed (e.g., 1-to-1 connections are addressed in [10]). We will also review key studies illustrating that PCNs are widely used and informative models of information processing in the brain.

Model in [48] already autonomous, because plasticity triggered by network convergence? The plasticity trigger in [48] requires global information (i.e., total error in all error nodes), while our trigger $\phi$ needs only local information, thus, is more plausible for biological implementation.

Experiment of Fa-Z-IL? $\phi$ with $t_d > 4$ succeeds in all detections (Table 3), i.e., Fa-Z-IL with such $\phi$ coincides with Z-IL (BP). We will include the classification results of such Fa-Z-IL, which produces the same results as Z-IL (BP).

Clarification of full autonomy? Fa-Z-IL does require the input to be presented before the teacher to satisfy C1. We consider this to be a requirement of the learning setup and will make it explicit in the paper, moderating our claim of autonomy. However, we consider such requirement to be much weaker, compared to switching computational rules (BP) and detecting convergence of global variables (IL). We leave the study of removing this requirement or putting it inside an autonomous neural system as future research.

Moderate the title? We will modify the title to the more specific "Can the Brain Do Backpropagation? — Exact Implementation of Backpropagation in Predictive Coding Networks".

From Z-IL to Fa-Z-IL? We will add the statement that Fa-Z-IL loses formal equivalence to BP, but with $t_d > 4$, empirical equivalence always remains.

C1 in [48] should be acknowledged? We will acknowledge this.

Clarification of [48]? In line 42, we have stated that some previous works are equivalent to BP when feedback is sufficiently weak (i.e., teacher weakly perturbs the network); we will add such clarification when introducing [48].

Discuss about [46]? We will add a section outlining the differences of learning rules between [46] and Z-IL, along which we point out that some steps may serve similar general purposes. A deeper study on the connections of the two substantially different learning rules is left as future research.

**Reviewer #4:** Moderate the title? Will be changed to "Can the Brain Do Backpropagation? — Exact Implementation of Backpropagation in Predictive Coding Networks".

Strong claim? We will moderate the sentence pointed out by the reviewer.

Classification accuracy? The classification accuracy of all situations in Fig. 3 will be added to the final paper: the averaged accuracies of BP, IL, Z-IL, and Fa-Z-IL are 94.16%, 93.78%, 94.16%, and 94.16%, respectively.

Details of the two criteria? The divergence of the test error is the L1 distance between the corresponding test errors, averaged over 64 training iterations (the test error is evaluated after each training iteration). The divergence of the final weights is the sum of the L2 distance between the corresponding weights, after the last training iteration.

Extra description of Fig. 2? Lines 133–141 are the description of Fig. 2; this will be included in the caption of Fig. 2.