We thank the reviewers for their very insightful and helpful comments and address their questions and remarks below.

Relationship to Mean Teacher/Temporal Ensembling: We are grateful to reviewers for pointing out this relevant related work that we missed in our original literature review. We will add the two citations in the related work section along with a paragraph on semi-supervised learning and connect MT and our current ablation study. Indeed, Table 5b (line 7: no predictor, $\beta = 0$, 0.2%) corresponds to using an MT-like approach in unsupervised learning (i.e., removing MT's classification loss) and we show that this approach does collapse. BYOL's novelty over MT is to perform well even without labels or classification loss, thanks to the addition of a predictor.

Code and reproducibility: We will release an open-source version of our full pretraining pipeline, the pretrained checkpoints, and the linear evaluation pipeline on ImageNet within the next two weeks. To further improve reproducibility and accessibility, we will also provide a single-GPU setup for pretraining on the smaller Imagenette dataset.

Importance of a near-optimal predictor: As the predictor is only applied to the online branch, its role is to make the architecture asymmetric rather than just making the network deeper. Table 5b already shows the importance of combining a predictor and a target network: the representation does collapse when either is removed. Furthermore, new experiments show that keeping the predictor near-optimal at all times is key to preventing collapse, which may be one of the roles of BYOL's target network. We further found that we can remove the target network without collapse by making the predictor near-optimal, either by (i) using an optimal *linear* predictor (obtained by linear regression on the current batch) before back-propagating the error through the network (52.5% top-1 accuracy at 300 epochs), or (ii) increasing the learning rate of the predictor (66.5% top-1). By contrast, increasing the learning rates of both projector *and* predictor (without target network) yields poor results ($\approx 25\%$ top-1).

Explaining BYOL's non-collapse: Similarly to GANs, BYOL uses two sets of parameters that are not minimizing the same objective. Thus, there is no *a priori* reason for BYOL's dynamics to converge to a global minimum of $||\bar{q}_\theta(z_\theta) - \bar{z}'_\theta||^2$, as they are not following the gradient of this loss ($\mathcal{L}_\theta^{\texttt{BYOL}}$ uses $\bar{z}'_\xi$). While these dynamics still admit undesirable equilibria where all images are mapped to the same constant projection (e.g., all zeros), BYOL's empirical performance seems to indicate that such equilibria may be unstable. We hypothesize that maintaining a near optimal predictor at all times is key to avoid collapsed solutions. When using an optimal predictor, BYOL minimizes the (expected) conditional variance of the target projection given the online projection . With a fixed target network, adding more information to the online projection can reduce this conditional variance, but cannot increase it. For example, training dynamics will always tend not to collapse features from the online network, as for any constant C and variables X and Y, $\mathrm{Var}(Y|X) \leq \mathrm{Var}(Y|C)$. More generally BYOL is encouraged to keep features from the online projection diverse by latching onto any source of variability Z (stemming, e.g., from noise in training dynamics) distinct from existing features, as $\mathrm{Var}(Y|X,Z) \leq \mathrm{Var}(Y|X,h(X))$ for any variables X, Y, Z and any function h. We will add these additional discussions and experiments to our submission to clarify the role of the predictor.

Note on Fig 3a/footnote 4 (batch size): When dividing the batch size by N, we also average gradients for N steps. Without batch-norm (for BYOL), the two computations would be exactly equivalent. With batch-norm for BYOL, the two computations only differ in how the batch-norm statistics are computed.

Note on robustness: As described in Section 5, contrastive methods need to make the discrimination task challenging, which requires many negative examples (large batch size) and absence of uninformative features that are easy to discriminate (strong transformations). They stop learning once their prediction is sufficiently similar to the positives compared to the negatives. Instead BYOL does not rely on comparing positive and negative pairs and keeps latching on new information thanks to the predictor. It should therefore not be as sensitive either to batch-size or transformations.

Answers to Reviewer 1: We thank you for your positive comments, and we hope to have answered your questions in explaining BYOL's non-collapse. As an optimal predictor seems sufficient to favor a diverse representation and stabilize the training, it removes the need for negative examples that are required to balance contrastive objectives.

Answers to Reviewer 2: We thank you for your in-depth questions and comments. Non-collapse and L.745-756: see discussion above. Table 1b vs 2b: Linear evaluation trains the linear layer on top of the representation with 100% of the labels, while semi-supervised learning only uses 1 or 10% of the dataset to finetune the full network (including the linear layer). This explains the difference in performance. Note that the same trend is observed in SimCLR and other related literature. L.224-228: these changes allow us to use the same hyperparameters as those optimized for SimCLR in their paper; these parameters are however not optimal at 1000 epochs, though the gap is low. L.241-251: the cosine similarity in BYOL is not between the same terms, and crucially, involves the output of the predictor. See also note on robustness above. ℓ^2 vs cosine: we wanted to emphasize that BYOL also works well with just an ℓ^2 loss (no normalization, as per table 19). Table 5: table 5 provides the trend of the target update, we further sweep on this hyperparameter in our main experiments. L.559: No, it is disjoint. Equation 5: eq. (5) is a generalization of the InfoNCE loss with added temperature parameter α, in expanded form. We will add the derivation from the I_{NCE} equation (10) of Poole et al. in Appendix F.4., to clarify differences with our equation (5).

Answers to Reviewer 3: Thank you for your insights and your time. Reproducibility: see discussion above. Detection results: we are using the same setup as in MoCo's Table 4 (fine-tune the representation on `trainval2007` only, not `trainval2007+12`), for which BYOL gains $+2.6 AP_{50}$ over MoCo. L34: see note on robustness above.