All reviewers agree that this contribution is timely, and recognize its potential impact. We thank **R1** in particular for providing an excellent, concise summary of our contributions. While no major technical/scientific issues were raised in any of the reviews, our manuscript is nevertheless stronger by virtue of incorporating clarifications requested by **R2**, **R3**, and the addition of relevant references and stronger baselines (see Fig. below) suggested by **R3**. To reiterate our central contributions: we provided a rigorous analysis and proposed a solution for a critical technical issue that now enables learning of interpretable representations with coupled autoencoders. We applied this development to an unprecedented patch-seq dataset consisting of thousands of samples. Our optimization framework provides a novel, principled way of assessing the cell type hypothesis (e.g. Zeng and Sanes, 2017), and the results suggest that neuronal identities can be consistent to a surprisingly high degree across transcriptomic(T) and electrophysiological(E) modalities.

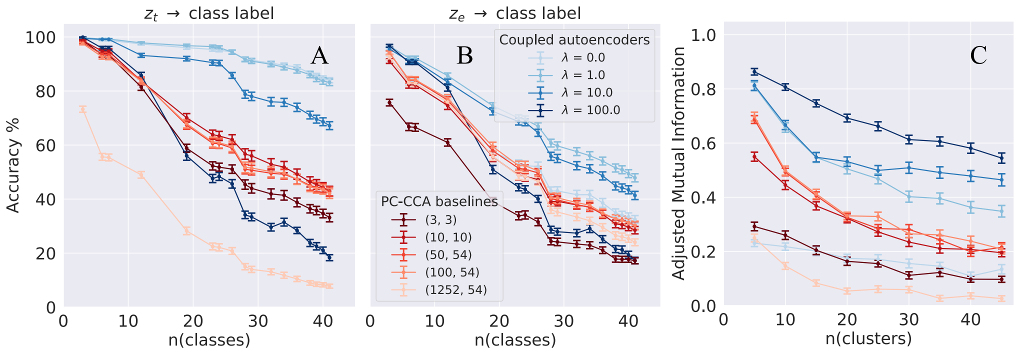
Recognizing that the reviewers did not point to any technical or scientific flaws in the *Improvements* section, we respectfully hope that the clarifications and analysis provided here will warrant substantially higher scores.

Misc. (R3) The dataset consists of 1252 differentially expressed genes, selected after excluding sex/mitochondrial genes. E recordings on the 2945-1518 neurons did not satisfy predefined quality control criteria. Indeed, an important strength of our approach is the ability to work with partially matched datasets. **(R2,R3)** We have added these and other clarifications, annotated well-known classes and hierarchical structure (Exc vs. Inh, Sst, Vip, PValb classes etc. in Fig. 2C) to make the connection to biology evident, and included 2D and 3D $\lambda = 0$ representations for completeness.

Sec. 2.1-2.4: **(R3)** lines 37-42: There are no explicit transformation matrices required to go from one representation space to another for the coupled autoencoder. We rephrase this now to avoid confusion. **(R3)** line 248: α is first defined in Sec. 2.1, and used consistently in Sec. 2.2 and 2.4. As studied in Sec. 2.4, it represents the relative noise level in the different modalities. Since this ratio is not measured explicitly, we heuristically set $\alpha = 0.1$ for all patch-seq experiments to capture the understanding that the T data is of higher resolution and quality. **(R2)** line 80: $\mathcal{E}$ and $\mathcal{D}$ can indeed be nonlinear; the statement only implies that they are *at least* capable enough to represent any linear transformation. **(R2)** The objective function contains two or more reconstruction error terms, and so all α_i's cannot be absorbed into λ. **(R2)** The suggestion to use a distance metric following normalization of individual representations would not prevent representations from collapsing. Proposition 2 proves why such strategies are guaranteed to fail, and formalizes exactly this non-trivial understanding of the problem.

Feng *et al.* 2014: **(R2,R3)** Feng *et al.* do not specify any normalization (Batch Norm. (BN) paper appeared in 2015). tSNE transforms used in that paper hide the shrinking problem, and their representations display poor alignment for all parameter values (Fig. 11 in Feng *et al.* - squares vs. pluses). Without normalization, the representations asymptotically collapse to a point (Prop 1). With uncoordinated normalization (e.g., BN), they asymptotically collapse to a line (Prop 2, k-CBNAE). Our proposed solution (C_{MSV}) avoids both problems, and is efficient and robust (Fig 1C,D and Sec. 2.3).

Representation quality: **(R2,R3)** Our optimization framework trades off the consistency of representations across modalities against the fidelity of representations to raw data. The ultimate test of whether coupled representations are biased by either modality is the cross modal data prediction ability (as quantified in Fig. 4C). Representations z_t and z_e in Fig. 2C show consistency across modalities (dot positions), and capture biologically relevant transcriptomic hierarchy of cell classes (colors, Fig 2B-C). **(R3)** line 228, Fig. 3B: As coupling (λ) increases, z_t and z_e become more consistent (Fig. 3C), at the expense of z_t capturing less of the transcriptomic hierarchy.(Fig. 3A). It is precisely because of this 'handshake' that $\lambda = 10$ is marginally lower than $\lambda = 1$ in Fig. 3B. While we do not tune α and λ, as multimodal datasets mature, it would be appropriate to optimize these parameters based on cross-modal data prediction ability (e.g. $x_t \rightarrow z_t \rightarrow \tilde{x}_e$: start from raw T data x_t, obtain the representation z_t, and pass it through the E decoder to predict raw E data x_e). We explore this systematically in Fig. 4C, where we show within and across-modality data prediction accuracies relative to reconstruction accuracy of individual, uncoupled ($\lambda = 0$) networks. From among the coupling strengths evaluated, $\lambda = 10$ strikes a desirable balance between measures of consistency in the latent space (Fig. 3B,C, example in Fig. 2C) while capturing known cell type hierarchies (Fig. 3A), and prediction accuracy (Fig. 4).



Revised Fig. 3. Coupled AE representations outperform additional CCA baselines: Tuples (t,e) in the legend indicate the number principle components for T and E data used as input for CCA alignment. Clusters of coupled AE representations ($\lambda \in \{1, 10\}$) agree with transcriptomic class labels. (A,B) and are consistent across modalities (C)