Removing Inter-Experimental Variability from Functional Data in Systems Neuroscience

Anonymous Author(s)
Affiliation
Address
email

Abstract

Integrating data from multiple experiments is common practice in systems neuroscience but it requires *inter-experimental variability* to be negligible compared to the biological signal of interest. This requirement is rarely fulfilled; systematic changes between experiments can drastically affect the outcome of complex analysis pipelines. Modern machine learning approaches designed to adapt models across multiple data domains offer flexible ways of removing inter-experimental variability where classical statistical methods often fail. While applications of these methods have been mostly limited to single-cell genomics, in this work, we develop a theoretical framework for domain adaptation in systems neuroscience. We implement this in an adversarial optimization scheme that removes inter-experimental variability while preserving the biological signal. We compare our method to previous approaches on a large-scale dataset of two-photon imaging recordings of retinal bipolar cell responses to visual stimuli. This dataset provides a unique benchmark as it contains biological signal from well-defined cell types that is obscured by large inter-experimental variability. In a supervised setting, we compare the generalization performance of cell type classifiers across experiments, which we validate with anatomical cell type distributions from electron microscopy data. In an unsupervised setting, we remove inter-experimental variability from the data which can then be fed into arbitrary downstream analyses. In both settings, we find that our method achieves the best trade-off between removing inter-experimental variability and preserving biological signal. Thus, we offer a flexible approach to remove inter-experimental variability and integrate datasets across experiments in systems neuroscience.

1 Introduction

Systems neuroscientists are often concerned with identifying and characterizing how properties of neurons vary along certain dimensions of interest. Differences in these properties between neurons form the basis for sorting them into discrete categories. Both the advance of large-scale data acquisition techniques in experimental neuroscience as well as the development of more efficient and powerful data analysis methods allow collecting and analyzing datasets of increasing size; and hence the discovery of more subtle variations in neural function between cell types [e.g. 1–4]. However, as data acquisition is often an incremental process, it has become common practice to pool data from multiple experiments. This practice ignores variability in the data stemming from external factors, which include non-biological ones (e.g. sample handling resulting in small differences in tissue quality, or temperature fluctuations affecting the rates of biochemical processes) but potentially also unforeseen biological ones (e.g. subtle genetic variations) [3, 5, 6]. Such variability due to external factors, here referred to as *inter-experimental variability*, can confound and obscure the biological

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

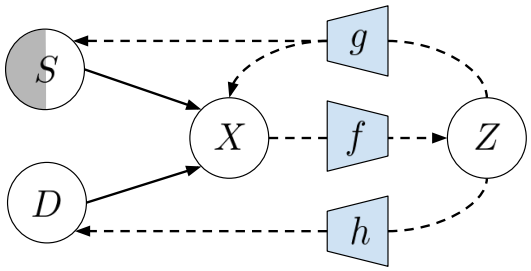

Figure 1: **Problem Setting**. Arrows represent given (solid lines) and modeled (dashed lines) relations. Capital letters denote random variables, small letters transformations (see Section 3). The setting with known S (white circle) is supervised, while unknown S (gray circle) is an unsupervised setting.

signals of interest. In some cases, the source of inter-experimental variability is known and can be modeled [5], but if this is not the case, a method for removing it from the data is required.

The issue of inter-experimental variability in systems neuroscience is analogous to the problem of *domain shift* in machine learning, where the data distribution changes between training ('source') and test ('target') data, causing an algorithm to fail when deployed on data from an unseen target domain [7–10]. Methods that address this issue have to perform some form of *domain adaptation*, i.e. adapting the algorithm to work both on the training as well as some (usually unseen) test domain [11]. One popular machine learning approach for domain adaptation deploys an adversarial two-player game to learn a domain-free representation of the input data, on which a classifier can then be built that is invariant for domain shifts [12]. In single-cell genomics, a number of different studies have leveraged this general setup for removing inter-experimental variability (see Section 2), but related works in systems neuroscience are lacking, despite the recognized need for such approaches [3, 5].

Here, we contribute to closing this gap as follows:

- We cast the removal of inter-experimental variability from functional data in systems neuroscience in the theoretical framework of domain adaptation (Figure 1 and Section 3).
- We adapt and evaluate different approaches and demonstrate improved performance of cell type assignment, while preserving the biological signals of interest (Table 1 and Figure 4).
- We demonstrate that our method produces cell type predictions on a new dataset that are best aligned with anatomical data (Table 2 and Figure 5).
- Finally, we showcase in a downstream analysis that the corrected data (Figure 3) clearly exhibits biological effects that were obscured by inter-experimental variability (Figure 6).

2 Related Work

As mentioned before, few studies have addressed the issue of inter-experimental variability in systems neuroscience. Zhao et al. [5] proposed a specialized solution to deal with the specific effects of temperature fluctuations on the response kinetics of retinal neurons by modeling them explicitly. In another study, Shah et al. [3] built encoding models to predict the responses of retinal ganglion cells across different experiments [see also 13]. They used this to study the systematic changes across experiments and compare it to covariates such as the gender of an animal. Crucially, both approaches offer specialized solutions that do not represent general purpose correction methods.

In single-cell genomics, a number of approaches for removing inter-experimental variability from data have been developed [14–21]. Two such methods are *Harmony* [22] and *scGen* [23]. *Harmony* performs iterative clustering using a variant of soft k-means until convergence to align cells from different datasets in a joint embedding. *scGen*, on the other hand, combines a variational autoencoder adapted for scRNA-seq data with latent space arithmetics to predict gene expression, while removing inter-experimental variability between datasets. In this paper, we compare our approach against these methods as they have been found to perform particularly well in two benchmarking studies [24, 25].

3 Theoretical Framework

The generative process of data denoted by a random variable X with image $\mathcal{X}$ is depicted in Figure 1. The biological signal shared across experiments (e.g. variation due to cell types) is represented by a random variable S ('signal') with image $\mathcal{S}$. We define D ('domain') as a random variable with image $\mathcal{D}$ that represents inter-experimental variability. Now, our objective is to learn a function f that transforms the data into a new random variable $Z := f(X)$ with image $\mathcal{Z}$. Importantly, we distinguish two settings: (*i*) *unsupervised* — where S is unknown and we simply try to retain in Z as much information about the data as possible while removing inter-experimental variability; (*ii*) *supervised* — where S is known and we additionally try to retain in Z as much information about S as possible.

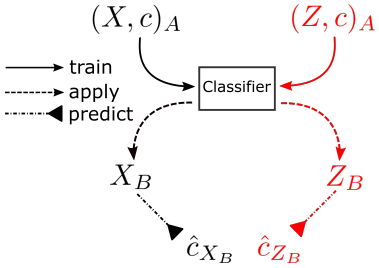

Figure 2: **Workflow.** Evaluating the preservation of signal information: A classifier gets trained on the labeled dataset *A* (either $(X, c)_A$ or $(Z, c)_A$) and applied to dataset *B* to predict labels (either $\hat{c}_{X_B}$ or $\hat{c}_{Z_B}$). The predicted labels are then used for further evaluation.

These objectives can be formulated in terms of mutual information, giving the unsupervised loss function

$$\mathcal{L} = I(Z; D) - I(Z; X) \tag{1}$$

and, provided knowledge about S, we obtain the supervised loss function

$$\mathcal{L}_+ = \mathcal{L} - I(Z; S). \tag{2}$$

Now, $I(Z; D)$ attains its minimum for $I(f(X); D) = 0$ because of the non-negativity of mutual information. And $I(Z; S)$ attains its maximum for $I(f(X); S) = I(X; S)$ because of the data processing inequality. If f were a bijection, it would follow that $I(f(X); S) = I(X; S)$, but also $I(f(X); D) = I(X; D)$. But by assumption, $I(X; D) > 0$ (otherwise there is no inter-experimental variability and we are done) and so we would have $I(f(X); D) > 0$. Thus, at the minimum of $I(Z; D)$, f cannot be a bijection. Generally, if there is an interaction between recording, signal and domain i.e. $I(X; S; D) \neq 0$, then there will be a *trade-off* between maximizing $I(f(X); S)$ and minimizing $I(f(X); D)$. This trade-off becomes even more apparent in the unsupervised setting where $I(f(X); X)$ and $I(f(X); D)$ are clearly competing.

Mutual information quantifies the dependence between two variables but it is difficult to estimate [26–29]. Instead, we measure dependency through nonlinear regression with an appropriate distance metric d.[1] Usually D is a discrete random variable indicating the experiment of a recording, and so, to estimate $I(Z; D)$, we can perform classification with a classifier function $h : \mathcal{Z} \to \mathcal{D}$, minimizing the standard cross-entropy $d_{\text{CE}}(h(Z), D)$ (Figure 1). Lemma 10 in [31] shows that this gives a variational lower bound to $I(Z; D)$ [see also 32]. In some cases S may also be discrete (e.g. cell types) and we can do the same, in other cases it might be a (high-dimensional) continuous random variable and so, to approximate $I(Z; S)$, we can perform regression, minimizing the mean squared error $d_{\text{MSE}}(g(Z), S)$. Similarly, in the unsupervised setting, to approximate $I(Z; X)$, we minimize $d_{\text{MSE}}(g(Z), X)$. To keep notation simple, in the unsupervised setting we define the mapping $g : \mathcal{X} \to \mathcal{X}$, and in the supervised setting $g : \mathcal{X} \to \mathcal{X} \times \mathcal{S}$. Putting this together, in the unsupervised setting our objective is

$$\min \mathcal{L} \longrightarrow \min_{g,f} \max_{h} d(g(f(X)), S) - \lambda\, d(h(f(X)), D) \tag{3}$$

where we have introduced a hyperparameter λ that mitigates the trade-off discussed above. In the supervised setting our objective becomes

$$\min \mathcal{L}_+ \longrightarrow \min_{g,f} \max_{h} d(g(f(X)), (X, S)) - \lambda\, d(h(f(X)), D). \tag{4}$$

In both equations we find a $\min - \max$ optimization, where h is trying to predict D from Z, tightening the lower bound on $I(Z; D)$ [see 31], while f is trying to prevent that by removing information about D from Z, effectively lowering $I(Z; D)$. Practically, this optimization scheme has become a standard adversarial setting in machine learning, for instance, in the training of generative adversarial networks [33] or for discriminative adversarial domain adaptation [12].

[1] If the joint probability density of two random variables is a bivariate normal distribution, then the mutual information is proportional to their linear correlation [30].

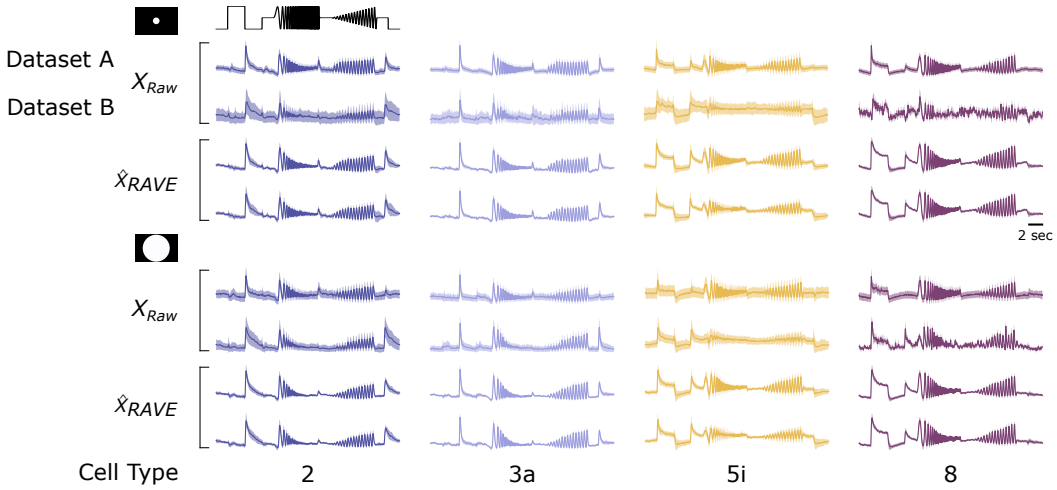

Figure 3: **Exemplary Cell Type Responses to the Chirp Stimuli.** Four cell type mean responses with SD to the local (top panel) and full-field (bottom panel) chirp of raw data X and reconstructed data $\hat{X}_{\texttt{RAVE}}$ for both datasets. Each row shows the mean responses of one cell type.

4 Methods

4.1 Datasets

To test our model approach, we use two datasets of two-photon imaging recordings [34–36] from the 14 mouse retinal bipolar cell (BC) [37] types' responses to two visual stimuli, a local and full-field chirp stimulus (Figure 3). The axon terminals of BCs stratify at distinct, cell type-specific depths within the second synaptic layer of the retina, the inner plexiform layer (IPL). The functional BC data were obtained by imaging the glutamate output at their axon terminals using the genetically encoded glutamate-sensing fluorescent reporter iGluSnFR [38]. In our study, we refer to these two datasets as *A* [2] and *B* [39] (for further preprocessing see Appendix).

In [2], an anatomy-guided clustering approach to cluster the BCs into the 14 functional types was applied to dataset *A*, thus providing functional reference cell type labels, which do not exist for dataset *B*. However, even if both datasets recorded the same cell types, they suffer from inter-experimental variability making it difficult to match and, for example, to use dataset *A* to predict the cell type labels for dataset *B*. Here, the inter-experimental variability may originate from, at least, the three following differences between the datasets: (*i*) dataset *A* recorded BCs mostly at certain IPL depths ('ChAT-bands', which are landmarks within the IPL [40]) using tangential scans parallel to the retinal layers, whereas dataset *B* used axial scans employing an electrically tunable lens to record from BCs across the entire IPL simultaneously [5] resulting in different sampling distributions; (*ii*) the chirp stimulus used in dataset *B* differs slightly as the sinusoidal intensity modulation of the increasing frequency is marginally slower; (*iii*) dataset *A* did not employ a gamma correction of the display device to linearize its intensity curve, resulting in slightly different stimulus contrasts [41].

4.2 Models

All methods described below learn a function ϕ that either transforms a data sample $x \in \mathcal{X}$ into an embedding $z := \phi(x),\ z \in \mathcal{Z}$ or a reconstruction $\hat{x} := \phi(x),\ \hat{x} \in \mathcal{X}$ from which inter-experimental variability has been removed to a varying degree. We randomly split the data into training, validation and test set and train all models with empirical risk minimization. We optimize hyperparameters through random search [42] on the validation set and report performances on the test set which is only used for final evaluation. The training setup, model architectures, data processing, and random search details for all tested methods are fully described in the Appendix. Usually we have $\dim(\mathcal{Z}) << \dim(\mathcal{X})$ and so, for different downstream evaluations, we map between these representations: (*i*) with least squares reconstructions ($\mathcal{Z} \rightarrow \mathcal{X}$), or (*ii*) principle component projections ($\mathcal{X} \rightarrow \mathcal{Z}$) (see Appendix).

4.2.1 Unsupervised Model

We parameterize the functions f, g and h (Figure 1) with neural networks. In the unsupervised model, the function $g : \mathcal{Z} \to \mathcal{X}$ provides a reconstruction $\hat{x} := g(z),\ \hat{x} \in \mathcal{X}$ of the data. With the concurrent task (3) of minimizing the predictability of the domain D, this reconstruction should only contain parts of the original data that are indiscernible across experiments. Since the purpose of our method is to `Remove`, `Adversarially`, `Variability` from datasets collected in different `Experiments`, we term our model `RAVE`.

4.2.2 Supervised Model

In the supervised setting, we have partial knowledge about the biological signal S. The function $g : \mathcal{Z} \to \mathcal{X} \times \mathcal{S}$ now returns a reconstruction as well as a prediction of that signal $(\hat{x}, \hat{s}) := g(z),\ \hat{s} \in \mathcal{S}$. When optimizing (4), this additional task is equivalent to discriminative adversarial domain adaptation [12]. In the particular data that we work with, we have two datasets $\mathcal{D} := \{A, B\}$, but the biological signal S consists of cell type labels which are only available in the first dataset A. Thus, more accurately, this presents a *semi-supervised* scenario where one wishes to classify a newly recorded dataset according to some existing classification scheme. We term this extended version of our model `RAVE+`.

4.2.3 Comparison Models

We test three different methods for comparison with our model. Our simplest comparison model is a linear model that projects out the contrast between the dataset indicator variables (see Appendix). This has an analytic solution and no hyperparameters, serving as a baseline to get an estimate of the correction quality achieved by a standard method in classical statistics. The other two methods (*scGen* and *Harmony*) are run in an unsupervised learning mode without cell type information. Even though *scGen* could be utilized to run in a supervised mode with cell type information, this is not specified in a semi-supervised setting with only partial cell type labels available.

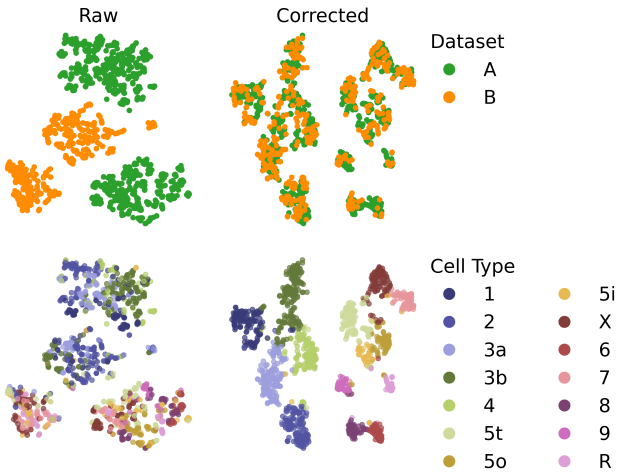

Figure 4: **Dataset Embeddings**. t-SNE embeddings of raw (left column) and corrected data by `RAVE+` (right column). Embedded cells are colored by dataset (top row) and cell type (bottom row).

4.3 Performance Evaluation

For evaluating the correction performed by the various methods, we analyze their output with respect to dataset-mixing (achieved by removing inter-experimental variability) and preservation of signal information.

4.3.1 Dataset-Mixing

The Rand index [43] measures similarity between two clusterings; the adjusted Rand index (ARI) is the Rand index adjusted for chance level (see Appendix) which was recently used by Tran et al. [25] to assess the quality of dataset-mixing in genomics. It takes as input the true and the predicted labels for a set of samples. We define $ARI_{dom}(z) := ARI(d, \hat{d}_z)$ with d the original domain labels (of the test set) and $\hat{d}_z$ the domain labels predicted by a classifier trained on z. On the raw data ($z = x$), we expect ARI_{dom} to be high due to inter-experimental variability. After successful correction (with z the output of a model), we expect ARI_{dom} to be low indicating good dataset-mixing.

In addition, we compute the accuracy (Acc_{dom}) of a domain classifier with the objective to predict the domain labels based on the input data. For low dataset-mixing, we expect a high Acc_{dom} as it should

Model	Corr.$_A$ ↑	Corr.$_B$ ↑	Acc$_{dom}$ ↓	Acc$_{type}$ ↑	ARI$_{dom}$ ↓	ARI$_{type}$ ↑
Raw	100	100	99.0 (0.1)	83.0 (0.5)	99.0 (0.3)	10.0 (3.3)
Linear	**98.9** (0.6)	**95.8** (2.8)	99.3 (0.2)	**83.0** (0.5)	97.4 (0.8)	5.7 (1.9)
Harmony	71.4 (11.2)	71.6 (13.8)	97.2 (0.3)	81.4 (0.3)	89.4 (1.1)	26.5 (2.5)
scGen	77.8 (10.1)	76.7 (11.3)	99.4 (0.1)	82.8 (0.4)	97.6 (0.5)	10.0 (2.5)
`RAVE`	60.4 (14.2)	61.3 (16.7)	71.9 (0.5)	75.1 (1.1)	15.3 (0.9)	75.0 (2.3)
`RAVE+`	55.8 (13.4)	52.7 (17.9)	**66.3** (0.8)	78.8 (0.4)	**3.8** (1.4)	**87.0** (1.8)

Table 1: **Model Comparison.** All entries in percentage. Mean and standard deviation metric scores across 10 random seeds. Bold font in each row indicates best score. $Corr._A$ ($Corr._B$) is the correlation of corrected data from dataset *A* (B) with its raw data. Acc_{dom} (Acc_{type}) is the accuracy of the domain (cell type) classifier. For ARI_{dom} and ARI_{type} see Section 4.3.

be trivial for the classifier to differentiate the datasets. However, after removing inter-experimental variability, Acc_{dom} is supposed to be close to chance level ($\sim 64\%$, cf. Table 1), which would indicate successful dataset-mixing. For the domain classifier, we use a random forest classifier with cross-validated hyperparameters for each model (see Appendix). This is crucial, because a powerful encoder f might hide (through multiple nonlinear transformations) domain information from a simple classifier, but still recover that information in an equally powerful decoder g. Conversely, we observe that overly expressive random forest classifiers, tend to overfit on the training set, thus underestimating the preserved domain or type information on the test set.

4.3.2 Preservation of Signal Information

In the unsupervised setting, to assess the amount of information preserved about the original data x, during the process of removing inter-experimental variability, we evaluate the rank correlation $Corr.(x, \hat{x})$ between input x and reconstruction $\hat{x}$. In the (semi-) supervised setting, we have reference cell type labels c_A for dataset A. To estimate how much of this information has been preserved, we fit cell type classifiers ψ (random forests, like above, see Appendix) that predict those labels $\hat{c}_A := \psi(z_A)$ (Figure 2). If a method succeed at preserving signal information in z after removing inter-experimental variability, then we expect the classifier to have a high accuracy (Acc_{type}). Deteriorating classification performance between raw data $\psi(x_A)$ and model output $\psi(z_A)$ would indicate signal loss.

Additionally, we would like to evaluate how well cell types can be distinguished and how biologically plausible they are for the unlabeled dataset *B*. To this end, we apply the classifiers $\hat{c}_B = \psi(z_B)$ obtaining cell type predictions $\hat{c}_B$. One direct comparison is between the distributions over cell types in $\hat{c}_A$ and $\hat{c}_B$ (Figure 5A). However, we can also evaluate the accuracy of these predictions by making use of BC axonal stratification profiles obtained from electron-microscopy (EM) data [44, 45]. From those data, we know where in the IPL a BC type stratifies its axon terminals. Thus, we can compare the distribution over IPL depth for the predicted cell types ($\hat{c}_B$) with the distributions expected from EM data. We quantify the difference between the expected and predicted distributions by calculating the Jensen-Shannon divergence. We define the depth score (DS) as the mean Jensen-Shannon divergence between those two distributions (Table 2). Additionally, we evaluate the robustness of cell type labels $\hat{c}_B$ by fitting the classifier ten times with different seeds and calculating the average ARI between different runs, giving the ARI_{type} score (Table 1).

5 Results

5.1 Unsupervised Removal of Inter-Experimental Variability

All methods tested here (*Harmony*, *scGen*, `RAVE`, `RAVE+` and *Linear*) succeed at retaining a significant amount of information about X in $\hat{X}$, reflected by high mean correlations between X and $\hat{X}$ (Table 1). Corr$(X, \hat{X})$ reaches similar levels for data from both datasets, suggesting that both datasets are modified to find a midway representation. This impression is confirmed when visualizing X_A, X_B and $\hat{X}_A$ and $\hat{X}_B$ next to each other (Figure 3). We show mean traces for exemplary cell types from

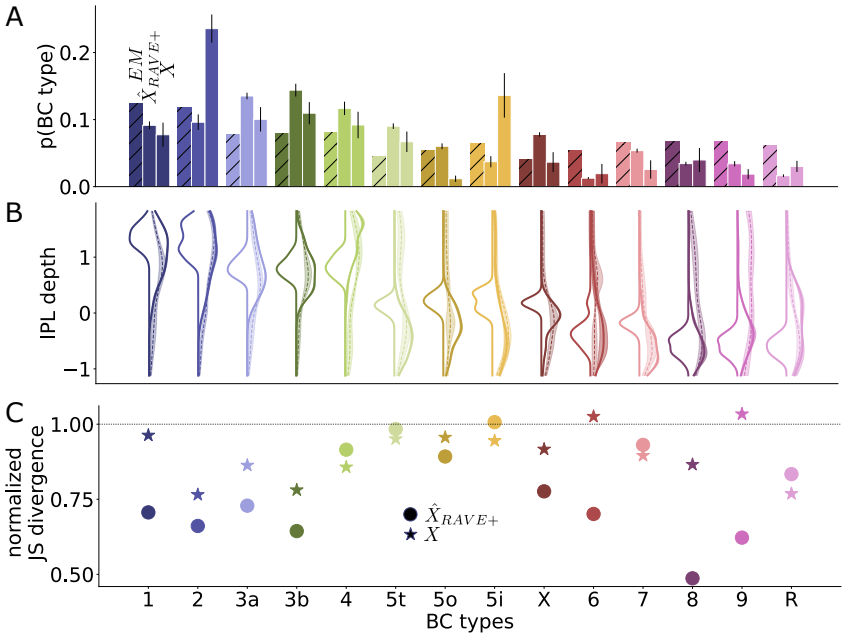

Figure 5: **Distribution Across BC Types and IPL Depths.** We compare the expected and predicted distribution of BCs from dataset *B* across the 14 types and across IPL depth. **(A)** Probability that a BC belongs to a certain type as estimated from EM data; as estimated from BC type labels predicted on $\hat{X}_{\texttt{RAVE+}}$; and as estimated from BC type labels predicted on X. Error bars indicate SD across 10 seeds of the classifier. (**B**) Distributions per cell type over IPL depth for EM data (distribution shown to the left), RAVE+ output (solid line to the right) and raw data (dashed line to the right). Shaded area around the distributions shown to the right indicate SD across 10 seeds of the classifier. (**C**) Normalized JS divergences corresponding to the distributions in B).

dataset *A*, and mean traces of cells from dataset *B* whose cell type labels we predict twice, first based on X (left pathway in Figure 2) and then again based on $\hat{X}$ (right pathway in Figure 2). As expected, inter-experimental variability obscuring the common signal S behind X_A and X_B causes the cell type assignment to fail; the similarity between responses of cells assigned to the same cell type, but coming from the different datasets is low (Figure 3, BC type 5i). Repeating the classification pipeline based on X with the same classifier architecture and different seeds yields highly variable cell type predictions for dataset *B* (Table 1, ARI_{type}) despite high prediction accuracy on dataset *A* (Table 1, Acc_{type}). This demonstrates a failure in transferring to dataset *B*, and not the classification itself. This results on the raw data X affected by inter-experimental variability were expected; however, the same pattern - low ARI_{type} (dataset *B*) and high Acc_{type} (dataset *A*) - is observed for *Harmony*, *scGen* and the linear model. This suggests that these methods fail at removing inter-experimental variability. The high domain accuracy achieved by a classifier trained on the outputs of these models confirms this conclusion. RAVE, on the other hand, succeeds at significantly lowering domain accuracy Acc_{dom}, while at the same time maintaining high scores for ARI_{type} and Acc_{type}.

5.2 Supervised Removal of Inter-Experimental Variability

RAVE+ extends RAVE to the (semi-) supervised setting where (partial) signal source information is present. RAVE+ excels at removing inter-experimental variability (Table 1, Acc_{dom} and ARI_{dom}) and at the same time retaining signal information (Table 1, Acc_{type} and ARI_{type}). A low dimensional t-SNE [46] embedding (Figure 4) shows that cells from datasets *A* and *B* are mapped onto the same cell type "islands". The distribution of types across IPL depth predicted by a classifier trained on $Z_{\texttt{RAVE+}}$ matches the expected anatomical distributions better than for all other methods (Figure 5 and Table 2). This provides a valuable validation of the estimate $\hat{S}_B$ learned by RAVE+ in the absence of ground truth knowledge of S_B.

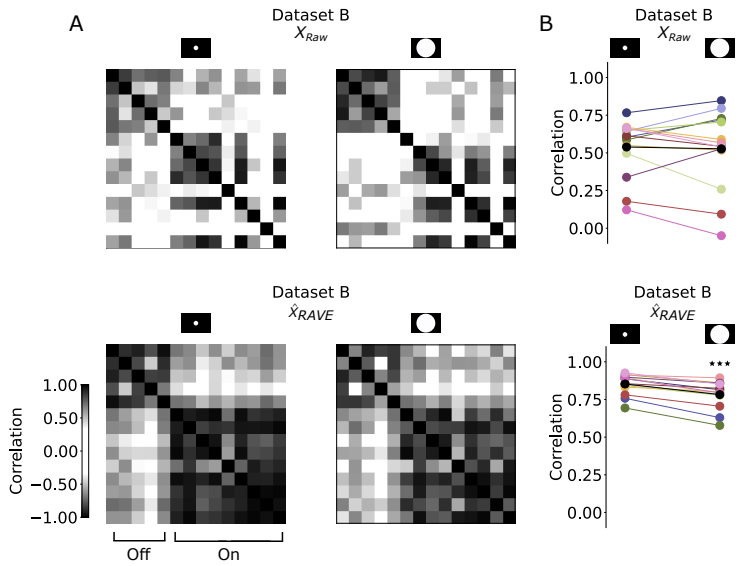

Figure 6: **Removing Inter-Experimental Variability Reveals Biological Effects.** (**A**) Correlation between mean responses per cell type cluster to local (left) and full-field (right) chirp of raw data X (top) and reconstructed data $\hat{X}_{\texttt{RAVE}}$ (bottom) responses from dataset *B*. (**B**) Mean correlation for each cell type mean response with all others types of the same response polarity between local and full-field chirp shown for raw (top) and corrected (bottom) responses. (X: mean correlation p_{local} = 0.54 and $p_{full\text{-}field}$ = 0.53, P = n.s.; $\hat{X}_{\texttt{RAVE}}$: mean correlation p_{local} = 0.85 and $p_{full\text{-}field}$ = 0.78, P < 0.0001, n = 14, non-parametric paired Wilcoxon signed-rank test).

5.3 Downstream Analyses on Reconstructed Traces

As in our unsupervised setting, it is common that no particular signal source information is available and that one wants to remove inter-experimental variability from the data to perform further downstream analyses. We show that a previously demonstrated biological effect, obscured by inter-experimental variability in X, emerges when performing the same analyses on the reconstructed traces $\hat{X}$ obtained from `RAVE`. Full-field visual stimulation has been shown to decorrelate responses from different BC types compared to local stimulation due to inhibitory feedback from amacrine cells (see Figure 3A, B in [2]). We expect this fundamental effect to be present in dataset *B*, but cannot reproduce it if we assign cells of dataset *B* to cell types based on the raw data (Figure 2, left pathway; and Figure 6A and B, top panel). However, using the reconstructed traces $\hat{X}_{\texttt{RAVE}}$, the expected effect is unmasked (Figure 2, right pathway; and Figure 6A and B, bottom panel). Here, the mean cell type responses to the local chirp are more correlated with one another than the full-field responses. This effect can also be seen when comparing the mean correlations for each cell type with all other cell types of the same response polarity (On and Off polarity) between the local and full-field chirp responses (Figure 6B).

BC Type	1	2	3a	3b	4	5t	5o	5i	X	6	7	8	9	R	∅	EM
Raw	48	43	49	42	57	**59**	65	59	60	55	**51**	54	58	50	53	23
Linear	45	34	46	42	56	62	62	57	61	55	51	41	58	50	52	34
Harmony	39	39	45	40	**51**	63	60	55	61	44	57	58	58	38	51	25
scGen	58	43	52	47	54	61	57	56	64	54	59	58	57	40	54	28
`RAVE`	59	**20**	45	40	56	60	68	**50**	60	**38**	55	35	47	**32**	48	**20**
`RAVE+`	**41**	37	**45**	**39**	55	64	**57**	61	**52**	39	61	**27**	**35**	50	**47**	**20**

Table 2: **Depth Score Comparison.** All entries in percentage, lower is better. Bold font in each row indicates best score. Depth Score - Jensen-Shannon (JS) divergence between predicted types and EM depth distribution. EM - JS score for P_{EM}(type) vs. P_{model}(type).

6 Limitations

Our method is limited to datasets where neurons were presented with the same stimulus. For other kinds of data, such as neural recordings from free behavioral paradigms where each trial will be different, it will be difficult to 'align' neural responses in a meaningful way. One solution to this could be to learn a shared embedding space [see 47], from which domain effects are removed, but distinct encoders f_i and decoders g_i for different trials i. In another setting, where different stimuli are presented between experiments, one might resort to an approach like Shah et al. [3]. Nevertheless, we do acknowledge that the data in our applications consists of *ex vivo* retinal recordings which have little to no attentional effects or task-dependent noise correlations like they would be present in *in vivo* cortical data. We are optimistic that our framework of adversarially removing inter-experimental variability is still a promising approach in those settings, under the constraint that a much more severe trade-off may need to be made between retaining signal and removing domain shifts.

7 Discussion

We have presented a framework to remove inter-experimental variability from functional recordings in systems neuroscience. To the best of our knowledge, this is the first application of domain adaptation methods to this kind of data. Using our unsupervised (`RAVE`) and (semi-) supervised (`RAVE+`) approaches, we demonstrated that we are able to remove inter-experimental variability while retaining signal information, which allows us to robustly predict cell type labels for a new dataset. We validated those predictions using an anatomy-based comparison to existing EM data.

Furthermore, our unsupervised approach `RAVE` is able to remove inter-experimental variability without cell type information. By using the corrected dataset B, we could unmask biological effects, obscured by inter-experimental variability, that had been previously described for dataset A. Thus, by allowing the integration and alignment of functional recordings across experiments, we have shown that biological effects in the data become more pronounced when using our model approaches. Inter-experimental variability is ubiquitous and we hope that this method will become a helpful resource to many experimenters as we make the code toolbox available as Open Source upon publication.

Ethically, we believe that our method can also make a contribution to systems neuroscience research in the context of the 3Rs (Replacement, Reduction and Refinement) for animal ethics: By enabling detection of more subtle biological signals after removal of inter-experimental variability, fewer animals may be needed to test a specific hypothesis. Lastly, we acknowledge the fact that any machine learning method or approach can be utilized for malevolent purposes. Therefore, we condemn, without any exceptions, any use of `RAVE(+)`, in any manner whatsoever, for the usage for weapons or warlike applications.

References

[1] Tom Baden, Philipp Berens, Katrin Franke, Miroslav Román Rosón, Matthias Bethge, and Thomas Euler. The functional diversity of retinal ganglion cells in the mouse. *Nature*, 529(7586):345–350, 2016.

[2] Katrin Franke, Philipp Berens, Timm Schubert, Matthias Bethge, Thomas Euler, and Tom Baden. Inhibition decorrelates visual feature representations in the inner retina. *Nature*, 2017.

[3] Nishal Shah, Nora Brackbill, Ryan Samarakoon, Colleen Rhoades, and Alexandra Kling. Individual variability of neural computations in the primate retina. pages 1–22, 2021.

[4] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[5] Zhijian Zhao, David A. Klindt, André Maia Chagas, Klaudia P. Szatko, Luke Rogerson, Dario A. Protti, Christian Behrens, Deniz Dalkara, Timm Schubert, Matthias Bethge, Katrin Franke, Philipp Berens, Alexander S. Ecker, and Thomas Euler. The temporal structure of the inner retina at a single glance. *Scientific Reports*, 10(1):1–17, 2020.

[6] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.

[7] Jing Jiang. A literature survey on domain adaptation of statistical classifiers. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*, 3:1–12, 2008.

[8] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. 2008.

[9] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[10] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[11] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[12] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[13] Fabian Sinz, Alexander S Ecker, Paul Fahey, Edgar Walker, Erick Cobos, Emmanouil Froudarakis, Dimitri Yatsenko, Zachary Pitkow, Jacob Reimer, and Andreas Tolias. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *Advances in Neural Information Processing Systems*, 31:7199–7210, 2018.

[14] Gabriele Beate Schweikert, Christian Widmer, Bernhard Schölkopf, and Gunnar Rätsch. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *NIPS*, volume 8, pages 1433–1440, 2008.

[15] Jonathan Bryan Dayton. Adversarial deep neural networks effectively remove nonlinear batch effects from gene-expression data. 2019.

[16] Tongxin Wang, Travis S Johnson, Wei Shao, Zixiao Lu, Bryan R Helm, Jie Zhang, and Kun Huang. Bermuda: a novel deep transfer learning method for single-cell rna sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome biology*, 20(1):1–15, 2019.

[17] Jingshu Wang, Divyansh Agarwal, Mo Huang, Gang Hu, Zilu Zhou, Chengzhong Ye, and Nancy R Zhang. Data denoising with transfer learning in single-cell transcriptomics. *Nature methods*, 16(9):875–878, 2019.

[18] Dongfang Wang, Siyu Hou, Lei Zhang, Xiliang Wang, Baolin Liu, and Zemin Zhang. imap: integration of multiple single-cell datasets by adversarial paired transfer networks. *Genome biology*, 22(1):1–24, 2021.

[19] Xiangjie Li, Kui Wang, Yafei Lyu, Huize Pan, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Muredach P Reilly, Gang Hu, and Mingyao Li. Deep learning enables accurate

clustering with batch effect removal in single-cell rna-seq analysis. *Nature communications*, 11(1):1–14, 2020.

[20] Eugene Lin, Sudipto Mukherjee, and Sreeram Kannan. A deep adversarial variational autoencoder model for dimensionality reduction in single-cell rna sequencing analysis. *BMC bioinformatics*, 21(1):1–11, 2020.

[21] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):1–14, 2019.

[22] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

[23] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

[24] Malte D Luecken, Maren Buttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *BioRxiv*, 2020.

[25] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21(1):1–32, 2020.

[26] Rudy Moddemeijer. On estimation of entropy and mutual information of continuous distributions. *Signal processing*, 16(3):233–248, 1989.

[27] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

[28] Janett Walters-Williams and Yan Li. Estimation of mutual information: A survey. In *International Conference on Rough Sets and Knowledge Technology*, pages 389–396. Springer, 2009.

[29] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[30] Izrail Moiseevich Gel'fand and Akiva Moiseevich Yaglom. Computation of the amount of information about a stochastic function contained in another such function. *Uspekhi Matematicheskikh Nauk*, 12(1):3–52, 1957.

[31] Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

[32] David Barber and Felix V Agakov. Information maximization in noisy channels: A variational approach. *Advances in Neural Information Processing Systems*, 16:201–208, 2003.

[33] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[34] Winfried Denk, James H Strickler, and Watt W Webb. Two-photon laser scanning fluorescence microscopy. *Science*, 248(4951):73–76, 1990.

[35] Winfried Denk and Peter B Detwiler. Optical recording of light-evoked calcium signals in the functionally intact retina. *Proceedings of the National Academy of Sciences*, 96(12):7035–7040, 1999.

[36] Thomas Euler, Susanne E Hausselt, David J Margolis, Tobias Breuninger, Xavier Castell, Peter B Detwiler, and Winfried Denk. Eyecup scope—optical recordings of light stimulus-evoked fluorescence signals in the retina. *Pflügers Archiv-European Journal of Physiology*, 457(6):1393–1414, 2009.

[37] Thomas Euler, Silke Haverkamp, Timm Schubert, and Tom Baden. Retinal bipolar cells: elementary building blocks of vision. *Nature Reviews Neuroscience*, 15(8):507–519, 2014.

[38] Jonathan S. Marvin, Bart G. Borghuis, Lin Tian, Joseph Cichon, Mark T. Harnett, Jasper Akerboom, Andrew Gordus, Sabine L. Renninger, Tsai Wen Chen, Cornelia I. Bargmann, Michael B. Orger, Eric R. Schreiter, Jonathan B. Demb, Wen Biao Gan, S. Andrew Hires, and Loren L. Looger. An optimized fluorescent probe for visualizing glutamate neurotransmission. *Nature Methods*, 10(2):162–170, 2013.

[39] Klaudia P Szatko, Maria M Korympidou, Yanli Ran, Philipp Berens, Deniz Dalkara, Timm Schubert, Thomas Euler, and Katrin Franke. Neural circuits in the mouse retina support color vision in the upper visual field. *Nature communications*, 11(1):1–14, 2020.

[40] DI Vaney. 'coronate'amacrine cells in the rabbit retina have the 'starburst'dendritic morphology. *Proceedings of the Royal society of London. Series B. Biological sciences*, 220(1221):501–508, 1984.

[41] Katrin Franke, André Maia Chagas, Zhijian Zhao, Maxime J.Y. Zimmermann, Philipp Bartel, Yongrong Qiu, Klaudia P. Szatko, Tom Baden, and Thomas Euler. An arbitrary-spectrum spatial visual stimulator for vision research. *eLife*, 8:1–28, 2019.

[42] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[43] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[44] Jinseop S. Kim, Matthew J. Greene, Aleksandar Zlateski, Kisuk Lee, Mark Richardson, Srinivas C. Turaga, Michael Purcaro, Matthew Balkam, Amy Robinson, Bardia F. Behabadi, Michael Campos, Winfried Denk, and H. Sebastian Seung. Space-time wiring specificity supports direction selectivity in the retina. *Nature*, 509(7500):331–336, 2014.

[45] Matthew J. Greene, Jinseop S. Kim, and H. Sebastian Seung. Analogous Convergence of Sustained and Transient Inputs in Parallel On and Off Pathways for Retinal Motion Computation. *Cell Reports*, 14(8):1892–1900, 2016.

[46] Geoffrey Hinton and Sam T Roweis. Stochastic neighbor embedding. In *NIPS*, volume 15, pages 833–840. Citeseer, 2002.

[47] Rohan Gala, Nathan Gouwens, Zizhen Yao, Agata Budzillo, Osnat Penn, Bosiljka Tasic, Gabe Murphy, Hongkui Zeng, and Uygar Sümbül. A coupled autoencoder approach for multi-modal analysis of cell types. *arXiv preprint arXiv:1911.05663*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 5.
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix and Table 1.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See sections 4.1, 4.2.3, 4.3 and Appendix.
   (b) Did you mention the license of the assets? [No] The assets are free license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The usage was discussed with the creators.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] It does not.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]