COVARIATE-PROFILE SIMILARITY WEIGHTING AND BAGGING STUDIES WITH THE STUDY STRAP: MULTI-STUDY LEARNING FOR HUMAN NEUROCHEMICAL SENSING

By Gabriel Loewinger[*,†,¶], Prasad Patil[*,‖] Kenneth Kishida[‡,§,**] and Giovanni Parmigiani[*,††,¶]

Harvard University[¶], *Boston University*[‖],
Wake Forest University[**] *and Dana Farber Cancer Institute* [††]

Prediction settings with multiple studies have become increasingly common. Ensembling models trained on individual studies has been shown to improve replicability in new studies. Motivated by a groundbreaking new technology in human neuroscience, we introduce two generalizations of multi-study ensemble predictions. First, while existing methods weight ensemble elements by cross-study prediction performance, we extend weighting schemes to also incorporate covariate similarity between training data and target validation studies. Second, we introduce a hierarchical resampling scheme to generate pseudo-study replicates ("study straps") and ensemble classifiers trained on these rather than the original studies themselves. We demonstrate analytically that existing methods are special cases. Through a tuning parameter, our approach forms a continuum between merging all training data and training with existing multi-study ensembles. Leveraging this continuum helps accommodate different levels of between-study heterogeneity.

Our methods are motivated by the application of Voltammetry in humans. This technique records electrical brain measurements and converts signals into neurotransmitter concentration estimates using a prediction model. Using this model in practice presents a cross-study challenge, for which we show marked improvements after application of our methods. We verify our methods in simulations and provide the `studyStrap` R package.

DRAFT

1. Introduction.

1.1. *Motivating Application.* The elucidation of the neural correlates of neurological and psychiatric diseases is critical for the development of treatments. Recently, a groundbreaking application of Fast-Scan Cyclic Voltammetry (FSCV), a technology historically used in animal models ([Rodeberg, et al., 2017]), has been implemented in awake humans to enable the monitoring of changes in neurochemical levels in the human brain with the temporal resolution (usually 10 measurements per second) and spatial resolution (probes are on the micrometer scale) necessary to investigate how these signals evolve in real-time and contribute to human behavior ([Kishida et al., 2011]; also see commentary by [Platt and Pearson, 2016]; [Kishida et al., 2016]; [Moran, et al., 2018]). This technique offers unprecedented opportunity to monitor human brain activity that is currently impossible to measure with non-invasive brain imaging techniques (e.g., functional magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalograhy (EEG)). FSCV

[*]NSF-DMS1810829
[†]T32 AI 007358
[‡]NIH, R01 DA048096; NIH, R01 MH121099; NIH, R01 NS092701; NIH, 5KL2TR00142
[§]WFSOM, Phys/Pharm Neurosurgery
Keywords and phrases: Multi-study learning, Machine Learning, Neuroscience

is an invasive technique that involves temporarily inserting a tiny recording electrode deep into the brain of participants undergoing neurosurgery for psychiatric or neurological conditions. This allows for the measurement of neurotransmitters (chemicals that brain cells use to communicate) such as dopamine, an important molecule that is thought to underlie learning, reward, and many psychiatric and neurological diseases such as drug addiction and Parkinson's disease ([Volkow, et al., 2018]). Measurements are taken at 10 Hz (i.e., 10 measurements per second) while participants perform behavioral tasks. Their neurochemical activity is then estimated and correlated with their performance on the task.

This technological achievement involves the translation of hardware commonly used in pre-clinical settings to the human neurosurgery operating room, but also critical is the novel application of statistical learning tools. This is because the technique does not directly measure neurotransmitter concentration (the measure of scientific interest) but rather records high-dimensional electrical signals that arise from measuring current changes in response to manipulation of the voltage potential across a recording electrode. These signals then must be "converted" into neurochemical concentration estimates via a predictive model. To train this model, one can use an in vitro training set where the true concentrations of the neurotransmitters (i.e., the outcome) are known, and the covariates (i.e., the electrical measurements) are measured ([Kishida et al., 2016]; [Moran, et al., 2018]). In practice, it is impossible to generate calibration data sets that exactly replicate the same settings as those present in brain tissue. The use of FSCV in humans requires the use of prediction models that generalize from controlled laboratory settings to data collected in the operating room where less experimental control is afforded. Producing accurate predictions in the brain therefore requires that models trained on in vitro training sets generalize 1) to different electrodes and 2) to data collected in the brain.

The original work to implement FSCV in humans ([Kishida et al., 2016]; [Moran, et al., 2018]) report that training models on datasets that combine training sets from multiple electrodes substantially improves the cross-electrode generalizability and accuracy of neurotransmitter estimates. This forces one to train an algorithm on data from multiple sources (multiple electrodes in vitro) and make predictions on data collected in a different setting (a different electrode in vivo). As validation in vivo is not possible, we sought to optimize neurochemical estimate accuracy by improving cross-electrode generalizability of the models. We used 15 in vitro datasets, each generated on a different electrode. We then developed algorithms to train models on all datasets save a held out validation set. This presents a cross-study problem, where each "study" is a training set generated on a different electrode. In an attempt to improve the quality of estimates of neurotransmitter concentrations, we drew upon the multi-study learning literature.

1.2. *Dataset Structure.* The data studied here are from 15 studies (electrodes) each comprised of roughly 20,000 observations. Covariates for each observation of the form vector called CV ("cyclic voltammogram") which can also be viewed as a single functional covariate. The covariates are electrical measurements (current measured in nA) collected at 1000 discrete voltage potentials. In each observation, the outcome is a measurement of chemical (neurotransmitter) concentration (nM). This dataset exhibited considerable between-study heterogeneity in the distribution of the covariates (Figure 15) and in the conditional distribution, of the outcome conditional on the covariates, (Figure 16). Neuroscientists have analyzed voltammetry data based not only on the raw covariates themselves ([Rodeberg, et al., 2017]) a but also on a numerical estimate of the derivative of the current with respect to voltage potential index, in an effort to provide between-study standardization ([Kishida et al., 2016]; [Moran, et al., 2018]). For this reason we present results from both here. The data are described in greater depth in supplementary materials (Section C.1).

DRAFT

1.3. *Formal Problem Statement.* We propose methods for training models using multiple studies (electrodes) to improve their predictive performance. We explore the setting in which we have K training studies with common outcomes and covariates, and we aim to make predictions on data collected in a separate study (study $K+1$), or so called "multi-study learning." We begin with a set of training studies, $\{\mathbb{S}_1, ..., \mathbb{S}_K\}$, $\mathbb{S}_k = [\mathbf{y}_k | \mathbb{X}_k]$, where we denote the outcome of the kth study as $\mathbf{y}_k$ and its matrix of covariates $\mathbb{X}_k$. The studies have sample sizes, $n_1,\ n_2, ...,\ n_K$. We assume the data from each study are generated independently *across* studies, where the covariates from study k are drawn from distribution $f_{\mathbb{X}_k}(\mathbb{X}_k)$, and the outcome from $f_{\mathbf{y}_k}(\mathbf{y}_k)$. We aim to make predictions based upon the design matrix of the target study, $\mathbb{X}_{K+1}$.

We do not assume the marginal distributions of $\mathbb{X}_k$, $\mathbf{y}_k$ or the conditional distribution of $\mathbf{y}_k|\mathbb{X}_k$ are common between studies. We specifically deal with between-study heterogeneity in both the distribution of the covariates, $f(\mathbb{X}_k)$, and the conditional distribution of the outcome given the covariates, $f(\mathbf{y}_k \mid \mathbb{X}_k)$, (i.e., the true model coefficients, $\boldsymbol{\beta}_k$).

1.4. *Earlier Methods.* The present work proposes weighting and ensembling schemes that build directly upon recent work which we review briefly. Connections with other related concepts are found in the Discussion.

1.4.1. *Merging.* As a baseline, we will consider the standard strategy of training a single model $\hat{\mathbf{Y}}_{Merged}(\cdot)$ on data from all studies combined into a single "Merged dataset." $\mathbb{S}_{Merged} = [\mathbb{S}_1, \mathbb{S}_2, ..., \mathbb{S}_K]^T$.

1.4.2. *Trained-on-Observed-Studies (TOS) Ensemble.* Recent work has shown that in settings where there is high between-study heterogeneity, ensembling classifiers trained on each of the individual studies can improve the predictive performance in multi-study settings above that of Merging ([Guan et al., 2019]; [Patil and Parmigiani, 2018]). Denote a classifier trained on data from the kth study by $\hat{\mathbf{Y}}_k(\cdot)$ and the predictions made with this classifier on the covariates of study $K+1$ by $\hat{\mathbf{Y}}_k(\mathbb{X}_{k+1})$. If one assumes a linear model and uses the OLS estimator, for example, then $\hat{\mathbf{Y}}_k(\mathbb{X}_{k+1}) = \mathbb{X}_{k+1}\hat{\boldsymbol{\beta}}_k$, where $\hat{\boldsymbol{\beta}}_k = (\mathbb{X}_k^T\mathbb{X}_k)^{-1}\mathbb{X}_k^T\mathbf{y}_k$.

The Trained-on-Observed-Studies (TOS) classifier $\hat{\mathbf{Y}}_{TOS}(\cdot)$ trains a single model on each study and then ensembles the resulting classifiers via

$$\hat{\mathbf{Y}}_{TOS}(\mathbb{X}_{K+1}) = \sum_{k=1}^{K} w_k \hat{\mathbf{Y}}_k(\mathbb{X}_{K+1})$$

where w_k are weights.

1.4.3. *Stacking Weights.* This ensembling paradigm can be implemented with a variety of weighting schemes ([Ramchandran, et al., 2019]; [Patil and Parmigiani, 2018]). Stacking weights ([Breiman, 1996]), for example, are generated from the coefficient estimates of a regression of the predictions from each constituent model against the true outcome labels of the training studies. This upweights predictions made from models that exhibit strong cross-study predictive performance. It is common to regress $\mathbb{X}_S$ on $\mathbf{y}_S$ using, for example, a non-negative least squares (NNLS) or Ridge regression, where

$$\mathbf{y}_S = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_K \end{bmatrix}_{N\times 1} \qquad \mathbb{X}_S = \begin{bmatrix} \hat{\mathbf{Y}}_{\mathbf{1}}(\mathbb{X}_1) & \hat{\mathbf{Y}}_{\mathbf{2}}(\mathbb{X}_1) & ... & \hat{\mathbf{Y}}_{\mathbf{K}}(\mathbb{X}_1) \\ \hat{\mathbf{Y}}_{\mathbf{1}}(\mathbb{X}_2) & \hat{\mathbf{Y}}_{\mathbf{2}}(\mathbb{X}_2) & ... & \hat{\mathbf{Y}}_{\mathbf{K}}(\mathbb{X}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{Y}}_{\mathbf{1}}(\mathbb{X}_K) & \hat{\mathbf{Y}}_{\mathbf{2}}(\mathbb{X}_K) & ... & \hat{\mathbf{Y}}_{\mathbf{K}}(\mathbb{X}_K) \end{bmatrix}_{N\times K}$$

DRAFT

and $\hat{\mathbf{Y}}_{\mathbf{k}}(\mathbb{X}_j)$ are the predictions on the design matrix of study j using the model trained on study k; $\mathbf{y}_k$ are the labels from study k. Assuming an intercept is included and the model is fit with NNLS, the weights $\hat{w}$ are generated from the following constrained optimization problem

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \|\boldsymbol{y}_S - \mathbb{X}_S^* \boldsymbol{w}\|_2^2 \qquad \text{subject to: } w_k \geq 0$$

where $\mathbb{X}_S^* = [\mathbb{1}\mathbb{X}_S]$. We use stacking in several cases, and include implementation details in the supplementary materials (Section: A.1).

1.4.4. *Hierarchical Resampling.* We also leverage the bootstrap and bootstrap aggregation ("bagging") literature. [Davison, et al., 1997] (pp.100-102) described a *randomized cluster bootstrap* or "Strategy 1", where both clusters and observations within a cluster are sampled *with* replacement; and the *cluster bootstrap* "Strategy 2" where only the first stage is sampled *with* replacement. Also relevant is Bagging or "**B**ootstrap **agg**regation," first proposed by Leo Breiman ([Breiman, 1996]), which implements bootstrap resampling in prediction settings. Training one or more learners on each of the bootstrap samples and then ensembling these resulting models can sometimes enhance performance, often through a variance reduction.

1.5. *Heuristic Summary and Motivation for Proposed Methods.* Our methods are motivated by the empirical observation that ensembling (TOS) and Merging each outperform each other at times, but that it may be difficult to predict when one will be superior. We thus sought to combine the methods by ensembling classifiers that were trained on combinations of the different datasets. This is achieved through a hierarchical resampling scheme that first samples a bag of studies (a list of study indices from which to subsample), then assigns slots to each study by drawing random proportions, and finally randomly selects observations from the selected studies to fill the slots. This results in synthetic studies ("Study Straps") that are combinations of a subset of the K observed training studies, with varying degrees of similarity to the originals. We then use these Study Straps as the units for training constituent models in an ensemble.

This hierarchical resampling scheme is a form of the generalized bootstrap. The application of the resampling approach implemented here in an ensembling setting is in essence a hierarchical bagging or cluster-based bagging approach to ensembling ([Li, et al., 2019]).

We analytically explore fundamental properties of the Study Strap framework and demonstrate that the TOS Ensemble and Merging approaches are special cases of the Study Strap. By identifying a unifying framework for these opposing strategies, we show that a continuum exists between them. We argue that leveraging this continuum to accommodate the level of within- and between-study heterogeneity can produce gains in predictive performance.

In our application, as well as others, the design matrix of the target study is available prior to observation of the labels, and at the time training takes place. This information can be leveraged to give greater weight to studies or study straps that have covariate distributions, or profiles, that are more similar to the target study. To implement this, we propose the Covariate Profile Similarity (CPS) Weighting scheme. Heuristically, this assumes that predictors will perform better on a given dataset if it has been trained on data that have a comparable distribution of covariates. We propose an ensembling method that tries to favor the generation of pseudo-studies that are similar, by some metric, to a target study.

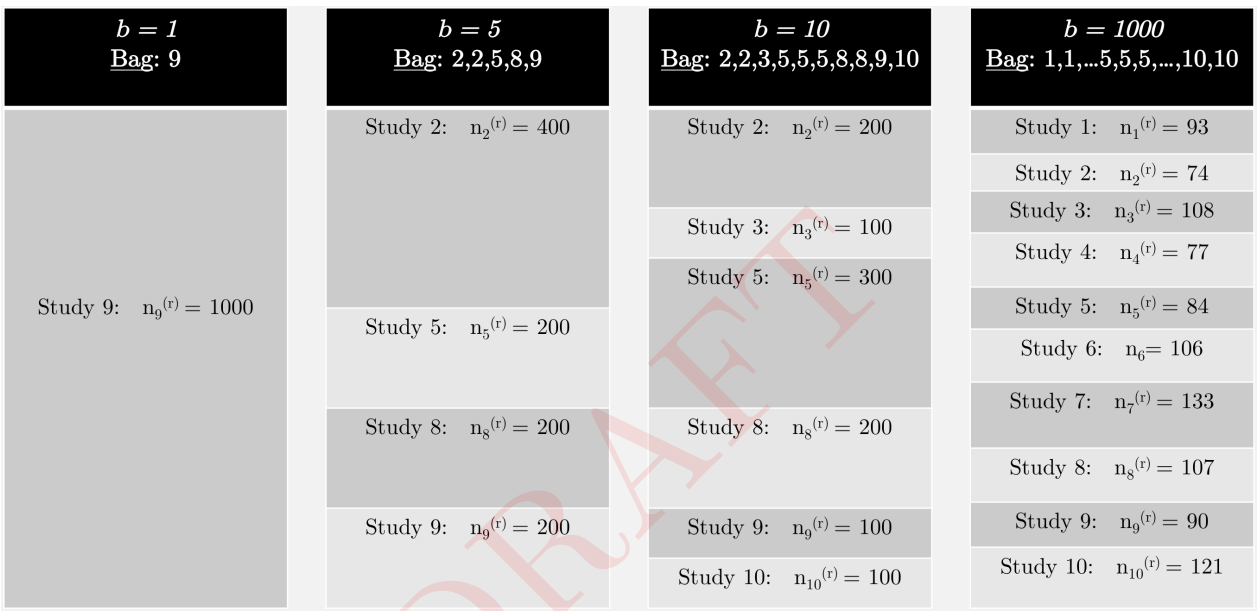
1.6. *Objectives and Outline.* This paper seeks to provide an in-depth introduction to Covariate Profile Similarity Weighting, the Study Strap, and the Covariate-Matched Study Strap. We describe the algorithms (Section 2), explore their properties analytically (Section 3), and provide empirical results via applications to simulated data (Section 4) and the electrode measurement dataset (Section 5) ([Kishida et al., 2016]).

DRAFT

Algorithm 1 Ensembling with Covariate Profile Similarity Weighting

Require: $\mathcal{S}(\cdot,\cdot)$, similarity function, $\mathbb{S}_1, ..., \mathbb{S}_K$, $\mathbb{X}_{K+1}$, $\boldsymbol{v}$, initial weights prior to CPS weighting

for each k in $[K]$ **do**
  $\boldsymbol{s}_k \leftarrow \mathcal{S}(\mathbb{X}_{K+1}, \mathbb{X}_k)$ *(Calculate similarity metric; store in vector $\boldsymbol{s}$)*
  Train $\hat{\mathbf{Y}}_{\mathbf{k}}(\cdot)$ on $\mathbb{S}_k$ *(Fit model on k^{th} study)*
  $\hat{\mathbb{Y}}_k \leftarrow \hat{\mathbf{Y}}_{\mathbf{k}}(\mathbb{X}_{K+1})$ *(Predict on test study; store in matrix $\hat{\mathbb{Y}}$)*
end for
return $\hat{\mathbb{Y}}, \boldsymbol{s}$ *(Return predictions matrix & similarity metrics vector)*
$\boldsymbol{w} \leftarrow \boldsymbol{s}\boldsymbol{v}/\sum_{k=1}^{K} s_k v_k$ *(Calculate weights)*
$\hat{\boldsymbol{y}}^* \leftarrow \hat{\mathbb{Y}}\boldsymbol{w}$ *(Take weighted average of prediction vectors)*
return $\hat{\boldsymbol{y}}^*$



DRAFT

Fig 1: Illustration of bag size tuning parameter on toy data example for the $r^{(r)}$ study strap. Example with $K = 10, n = n_1 = n_2 = ... = n_K = 1000$. Black portion shows bag size, b, and example study bags.

2. Algorithms.

2.1. *Ensembling with Covariate Profile Similarity Weighting.* We propose Covariate Profile Similarity (CPS) Weighting, a weighting scheme that upweights classifiers trained on datasets with average feature profiles that are similar to that of the target study, $\mathbb{S}_{K+1}$. For a given study or study strap k, we define a similarity metric, $s_k = \mathcal{S}(\mathbb{X}_{K+1}, \mathbb{X}_k) \in \mathbb{R}^+$ such that larger values correspond to greater similarity. An example metric may simply be the inverse of the ℓ_2 distance between the sample means of the covariates in the two studeis considered: $\mathcal{S}(\mathbb{X}_k, \mathbb{X}_{K+1}) = \|\bar{\mathbf{x}}_{\mathbf{K+1}} - \bar{\mathbf{x}}_{\mathbf{k}}\|_2^{-1}$. The similarity metric can then be used to adjust the weights for the predictions from the kth study to obtaining final weights $w_k = s_k v_k / \sum_{k=1}^{K} s_k v_k$.

2.2. *Ensembling with the Study Strap.* We propose an ensembling framework based on a hierarchical resampling scheme that generates pseudo-study replicates called *study straps*, and then ensembles the models trained on these replicates. To generate the rth study strap $\mathbb{S}^{(r)}$, we sample

Algorithm 2 Ensembling with the Study Strap

Require: bag size b, number of unique pseudo-studies R, observed studies $\mathbb{S}_1, ..., \mathbb{S}_K$, target study predictors $\mathbb{X}_{K+1}$, weight function $\mathcal{W}(\cdot)$

for each r in $[R]$ **do**
  $\mathbf{A}^{(r)} \leftarrow Multinomial_K\left(b, \frac{1}{K}\mathbb{1}_K\right)$ *(Study Counts)*
  for each k $\in$ [K] **do**
    $n_k^{(r)} \leftarrow \text{round}(n_k A_k^{(r)}/b)$
    Sample $n_k^{(r)}$ rows from $\mathbb{S}_k$ without replacement
    Add sampled rows from $\mathbb{S}_k$ to $\mathbb{S}^{(r)}$
  end for
  Train $\hat{\mathbf{Y}}^{(r)}(\cdot)$ on study strap $\mathbb{S}^{(r)}$ *(Fit model on r^{th} study strap)*
  $\hat{\mathbb{Y}}_r \leftarrow \hat{\mathbf{Y}}^{(r)}(\mathbb{X}_{K+1})$ *(Predict on test study; store in matrix $\hat{\mathbb{Y}}$)*
end for
return $\hat{\mathbb{Y}}$ *(Return matrix of predictions)*
$\boldsymbol{w} \leftarrow \mathcal{W}(\cdot)$ *(Calculate weights)*
$\hat{\boldsymbol{y}}^* \leftarrow \hat{\mathbb{Y}}\boldsymbol{w}$ *(Weighted average of study strap predictions)*
return $\hat{\boldsymbol{y}}^*$

counts $\mathbf{A}^{(r)} = (A_1^{(r)}, \ldots, A_K^{(r)})$ of study labels from a $Multinomial_K\left(b, \frac{1}{K}\mathbb{1}_K\right)$. We then determine, by normalizing these counts, the proportions $A_k^{(r)}/b$ of the original n_k observations that are to be drawn into the pseudo-study. This is shown in Figure 1. In the second, or observation-sampling step, we sample $n_k^{(r)}$ observations (with or without replacement) from each training study, rounding $n_k A_k^{(r)}/b$ to the nearest integer. We refer to b as the "bag size". It controls, stochastically, the number of studies from which we draw observations. A small bag size leads to fewer studies being represented with more observations from each study. A large bag size allows for more studies to be represented in a given study strap with fewer observations from each study. Although the Study Strap can employ observation-level sampling with or without replacement, we will argue below that sampling without replacement is preferable and therefore the standard Study Strap samples without replacement. The sample size of each pesudo-study replicate, $n^{(r)} = \sum_k n_k(r)$ is random, and is controlled by the bag size (through $\mathbf{A}^{(r)}$) and the sample sizes of the original studies. When $n_k = n$ for all k's, then $n^{(r)} = n$. We denote a classifier trained on the rth "study strap" as $\hat{\mathbf{Y}}^{(r)}(\cdot)$. To avoid redundancy in the predictions we require that each study strap is unique. We generate a total of R unique pseudo-study replicates. We later show that Merging and TOS are special cases of the Study Strap that arise from limiting values of b. As such, b can be used to adapt to the degree of between-study heterogeneity. We explore this property in Section 3. Study Strap ensemble learning proceeds by building a collection of studies in this fashion, training one or more single-study learners on each, and ensembling predictions with a simple average or another weighting scheme (e.g., stacking, CPS weighting or both). We leave this general in the algorithm description and denote the weighting function simply as, $\mathcal{W}(\cdot)$.

2.3. *Ensembling with the Covariate Matched Study Strap.* While generating study straps has the potential to create powerful training sets, it may also require large numbers of straps before enough studies with high similarity to the target comprise the training set. To address this issue, we developed an adaptive "accept/reject" (AR) approach to embedding covariate similarity into the study strap resampling scheme. AR generates study straps as in Algorithm 1, but only trains classifiers on study straps that meet an increasingly selective threshold for covariate similarity. Each accepted study strap updates the threshold for the similarity metric of the current strap. The

DRAFT

algorithm runs until a certain number of successive study straps are generated without acceptance for model training. Due to the stochasticity of this sampling scheme, we propose iterating through multiple accept/reject paths and ensembling all accepted classifiers from all paths. This reduces between-path variability in performance.

Algorithm 3 Ensembling with the Covariate Matched Study Strap (Accept Reject)

Require: b, $n^{(r)}$, η, convergence criteria, $\mathcal{S}(\cdot,\cdot)$, similarity metric, $\mathcal{W}(\cdot)$, weight function, paths, p, $\mathbb{S}_1, ..., \mathbb{S}_K$, $\mathbb{X}_{K+1}$

count $\leftarrow 0$ *(count tracks number of consecutive unaccepted straps)*
$r \leftarrow 1$ *(r counts number of accepted straps)*
$\tau \leftarrow -\infty$ *(τ: similarity threshold; $\tau \leftarrow 0$ ensures 1th strap is accepted)*
for path in p **do**
  while count $< \eta$ **do**
    count $\leftarrow$ count $+ 1$
    Generate $\mathbb{S}^{(r)}$ *(Random candidate study strap as above (2))*
    $_{(r)} \leftarrow \mathcal{S}(\mathbb{X}_{K+1}, \mathbb{X}^{(r)})$ *(Calculate similarity metric)*
    if $s_{(r)} > \tau$ **then**
      $\tau \leftarrow s_{(r)}$ *(Update threshold to similarity metric)*
      Train $\hat{\mathbf{Y}}^{(r)}(\cdot)$ on $\mathbb{S}^{(r)}$ *(Fit model on accepted study strap)*
      $\hat{\mathbb{Y}}_r \leftarrow \hat{\mathbf{Y}}^{(\mathbf{r})}(\mathbb{X}_{K+1})$ *(Predict on test study; store in $\hat{\mathbb{Y}}$)*
      count $\leftarrow 0$ *(Accepted study strap restarts counter)*
      $r \leftarrow r + 1$
    end if
  end while
  $\tau \leftarrow 0$ *(Path completion restarts acceptance threshold)*
end for
return $\hat{\mathbb{Y}}$ *(Return matrix of predictions)*
$\boldsymbol{w} \leftarrow \mathcal{W}(\cdot)$ *(Calculate weights)*
$\hat{\boldsymbol{y}}^* \leftarrow \hat{\mathbb{Y}}\boldsymbol{w}$ *(Weighted average of accepted study straps)*
return $\hat{\boldsymbol{y}}^*$

DRAFT

As was the case for the Study Strap, we can weight predictions based upon, for example, stacking, CPS, or simple average weights (i.e., $\boldsymbol{w} = \frac{1}{R}\mathbb{1}$).

2.4. *`studyStrap` Package.* We provide the `studyStrap` package, which implements the standard methods Merging, TOS, and stacking. Additionally it implements our proposed methods, the Study Strap, Covariate-Profile Similarity Weighting, and the Covariate-Matched Study Strap. Embedded within the `caret` environment, our package allows users to fit models with a broad range of methods (i.e., all available through `caret`) and fit multiple single-study learners per study. Our prediction functions automatically apply average weights, stacking weights, custom user-specified weighting functions, or one of many standard built in covariate-profile similarity weighting functions. Our package can be downloaded from `github` repository:
https://github.com/gloewing/studyStrap

3. Analytical Results. We present analytical results to demonstrate that the Study Strap provides a flexible and useful multi-study prediction framework that encompasses not only the methods proposed above but also earlier methods. Specifically, we show that the TOS Ensemble and the Merging Algorithm are special cases of the Study Strap. This is visually depicted in the supplementary materials (Figure 1).

First, observe that under standard (non-hierarchical) bagging, the number of times an observation is represented in a given bootstrap sample follows the Binomial distribution $Bin(N, \frac{1}{N})$ where $N = \sum_k n_k$ is the total sample size. In the Study Strap, we implement a hierarchical resampling scheme. From the multinomial specification of the vector of counts $\mathbf{A}^{(r)}$, the count $A_k^{(r)}$ corresponding to the kth study is *marginally* distributed as a $Bin(b, \frac{1}{K})$. If one samples observations with replacement, then the number of times the ith observation of the kth study is included in the study strap is distributed as $C_{k,i}^{(r)}|A_k^{(r)} \sim Bin(n_k^{(r)}, \frac{1}{n_k})$ where $n_k^{(r)} = n_k A_k^{(r)}/b$. If one samples without replacement, then $U_{ik}^{(r)}$ is indicator that the ith observation of the kth study is represented in the rth study strap, where $U_{ik}^{(r)}|A_k^{(r)} \sim Ber(A_k^{(r)}/b)$

We point out that under standard (non-hierarchical) bagging without replacement (or bagging with jackknife resampling) of the Merged dataset, the probability that an observation is represented in a given bootstrap sample is s/N, where N is the total sample size of the dataset and s is the size of the subsample.

PROPOSITION 1. *The Trained-on-Observed-Studies Ensemble is a Study Strap with bag size $b = 1$, $R = K$, and sampling without replacement.*

PROOF. First, observe that for the rth study strap, all of the observations in $\mathbb{S}^{(r)}$ are sampled from a single study, say $\mathbb{S}_k$ since we draw the rth study bag from a $Multinomial_K\left(1, \frac{1}{K}\mathbb{1}_K\right)$. Therefore $A_k(r) = 1$ for the sampled k and 0 otherwise. Then $n_k^{(r)} = n_k$ observations are sampled from the kth study. When the study strap is sampled without replacement, all observations are sampled so the rth study strap and the kth study are identical (i.e., $\mathbb{S}^{(r)} = \mathbb{S}_k$). It then trivially follows that the corresponding classifiers and predictions are equivalent: $\hat{\mathbf{Y}}^{(r)} = \hat{\mathbf{Y}}_k$ and $\hat{\mathbf{Y}}^{(r)}(\mathbb{X}_{K+1}) = \hat{\mathbf{Y}}_k(\mathbb{X}_{K+1})$.

Since we require each study strap to be unique, (i.e., $\mathbb{S}^{(i)} \neq \mathbb{S}^{(j)}\ \forall\ i \neq j \in \{1, 2, ..., R\}$, there are exactly K distinct bags (i.e., $R = K$) and corresponding study straps ($\{\mathbb{S}_1, \mathbb{S}_2, ..., \mathbb{S}_K\} = \{\mathbb{S}^{(1)}, \mathbb{S}^{(2)}, ..., \mathbb{S}^{(K)}\}$). It therefore directly follows that the final ensembles are equivalent. That is, the Trained-on-Observed-Studies Ensemble $\hat{\mathbf{Y}}_{TOS}(\cdot)$ is equivalent to the Study Strap classifier, $\hat{\mathbf{Y}}_{SS}(\cdot)$. □

DRAFT

The above proof stipulates that we cannot have multiple identical classifiers trained in an ensemble in order to achieve an exact equivalence between a special case of the Study Strap algorithm and the Trained-on-Observed-Studies Ensemble algorithm approaches. Although we feel this stipulation is principled, it is trivial to show that the vector of predictions generated from a Study Strap with no limitations on the number of identical classifiers in an ensemble would converge in probability to that produced by the Trained-on-Observed-Studies Ensemble (i.e., as $R \to \infty$), through appealing to the Weak Law of Large Numbers:

$$\frac{1}{R}\sum_{r=1}^{R} \hat{\mathbf{Y}}^{(r)}(\mathbb{X}_{K+1}) \xrightarrow{p} \frac{1}{K}\sum_{k=1}^{K} \hat{\mathbf{Y}}_k(\mathbb{X}_{K+1})$$

We now move to an important connection between the Merging algorithm and the Study Strap.

PROPOSITION 2. *The study strap with $b = \bar{n}$ is a delete $N - b$ jackknife (non-hierarchical) bagging of the Merged dataset.*

PROOF. Let $U_{ik}^{(r)}$ be the indicator of whether the ith observation in the kth study is represented in the rth study strap. First, recall that a delete-d jackknife sampling of a sample of size n^* entails sampling $n - d$ observations *without* replacement from the n^* observations. Thus, a delete $N - b$

jackknife entails sampling $N - (N - b) = b$ observations from the the Merged dataset, which has sample size N.

Next, recall that under standard (non-hierarchical) bagging without replacement (or bagging with jackknife resampling) of the Merged dataset, the indicator Z_i that observation i is represented in a given bootstrap sample is distributed as $Ber(s/N)$ where s is the number of observations sampled with replacement from the datatset. Now if we let $s = b = \bar{n}$, we have $b/N = \bar{n}/\sum_{k=1}^{K} n_k = \sum_{k=1}^{K} n_k/K \sum_{k=1}^{K} n_k = 1/K$. Thus that $Z_i \sim Ber(b/N) \equiv Ber(1/K)$.

Now, let us show for a study strap with parameters, $b = \bar{n}$ that the probability that the ith observation in the kth study is represented in the rth bag is marginally distributed as $Ber(1/K)$. First, recall that in generating the rth study strap, we sample $n_k^{(r)} = (n_k A_k^{(r)}/b)$ observations from the kth study, where $A_k^{(r)} \sim Bin(b, 1/K)$. Now the probability that the ith observation in the kth study is represented in a given study strap follows approximately the conditional distribution $U_{ik}^{(r)}|A_k^{(r)} \sim Ber(A_k^{(r)}/b)$. Where the approximation is up to the rounding error to ensure an integer sample size for the rth study strap. Therefore, the marginal probability that the ith observation in the kth study is represented in the rth study strap can be expressed as:

$$
\begin{aligned}
\mathbb{P}\left(U_{ik}^{(r)} = 1\right) &= \sum_{a=0}^{b} \mathbb{P}\left(U_{ik}^{(r)} = 1 | A_k^{(r)} = a\right) \mathbb{P}\left(A_k^{(r)} = a\right) \\
&= \sum_{a=0}^{b} \binom{a}{b} \binom{b}{a} \left(\frac{1}{K}\right)^a \left(1 - \frac{1}{K}\right)^{b-a} \\
\text{(Binomial identity)} \qquad &= \sum_{a=1}^{b} \binom{b-1}{a-1} \left(\frac{1}{K}\right)^a \left(1 - \frac{1}{K}\right)^{b-a} \\
&= \frac{1}{K} \sum_{a^*=0}^{b^*} \binom{b^*}{a^*} \left(\frac{1}{K}\right)^{a^*} \left(1 - \frac{1}{K}\right)^{b^*-a^*} \\
&= \frac{1}{K} \; \forall \; b \in \{1, 2, ..., N\}
\end{aligned}
$$

□

We can achieve a similar result when we sample observations without replacement, however, this requires the extra, and perhaps unrealistic, assumption that $n = n_1 = n_2 = ... = n_k$.

Proposition 3. *The study strap with $b = N$ is a standard (non-hierarchical) bagging of the Merged dataset when $n = n_1 = n_2 = ... = n_k$.*

Proof. Recall the Study Strap, is a hierarchical resampling scheme, where the proportion of observations sampled from the kth study, *marginally*, is proportional to $A_k^{(r)} \sim Bin(b, \frac{1}{K})$. The number of times the ith observation of the kth study is included in a bootstrap sample is distributed as $C_{k,i}^{(r)}|A_k^{(r)} \sim Bin(n_k A_k^{(r)}/b, \frac{1}{n_k})$. By standard probability results, the marginal distribution of $C_{k,i}^{(r)} \sim Bin(b, \frac{1}{K*n_k}) \equiv Bin(b, \frac{1}{N})$, where the equivalence follows from the assumption that $n = n_1 = n_2 = ... = n_k$. Thus when $b = N$, we have that the probability that a given observation in the kth study is represented in a bag follows the distribution $Bin(N, \frac{1}{N})$, a standard bagging scheme.

□

It then directly follows that under the same conditions, (i.e., $n = n_1 = n_2 = ... = n_k$), but with $b = n$ and $n^{(r)} = n$, the Study Strap is a standard (non-hierarchical) bagging of the Merged

DRAFT

dataset with a smaller sample size. We achieve a similar result when we sample *without* replacement, but instead the marginal probability of an observation being selected into a study strap follows a $Ber(b/N)$. Then the study strap with $b = \bar{n}$, is simply a delete-d jackknife bagging scheme where $d = N - \bar{n}$.

Taken together, the TOS is a Study Strap with bag size, $b = 1$, sample size, $n^{(r)} = \bar{n}$ and sampling without replacement. The Study Strap with the same parameters except where $b = n_k$ is a delete-d jacknife bagging scheme of the Merged dataset. Framed in this manner we see how the parameter b, varying between 1 and $\bar{n}$, can be thought of as providing a tuning mechanism that generates a continuum between the Merging algorithm and the TOS.

The utility of these proofs is three-fold. First, by framing the TOS and the Merging Algorithm as special cases of the Study Strap, we propose a unifying framework for multi-study prediction. Second, we provide a theoretical foundation for why the bag size provides a tuning parameter that can be used to account for varying levels of heterogeneity. With greater between-study heterogeneity, the TOS is often advantageous compared to the Merging algorithm ([Guan et al., 2019]); in such a setting, our proposed framework would then anticipate a smaller bag size to be optimal as the bag size $b = 1$ is identical to that of the TOS. Conversely, the framework predicts lower between-study heterogeneity would be better accommodated by a larger bag size (i.e., moving closer a subsample of the Merged Dataset). A hold-one-study out cross validation provides a principled tuning procedure to empirically identify the optimal bag size.

The third contribution of these proofs to the Study Strap framework is that they provide analytical rationale for the choice of sampling *with* or *without* replacement in the secondary (observation-level) step. If the studies have variable sample sizes and one wishes to ensure that each observation has the same probability of being selected into a study strap, then sampling *without* replacement will ensure this because this property is "robust" against variation in sample sizes between the studies (save some small approximation error from rounding). This then essentially "weights" the contribution of each study to the Study Strap ensemble in a manner proportional to the corresponding sample sizes of each training study. On the other hand, if one wishes to more evenly draw observations from each study despite variable sample sizes, then sampling *with* replacement is justified. This is because observations that come from studies with smaller sample sizes will have a greater marginal probability of being selected in a given study strap than observations from studies with larger sample sizes. Recall that in Proposition 3, we required that $n = n_1 = ... = n_k$ to achieve an equivalence between the Merging and a special case of the Study Strap. On the other hand, when one samples *without* replacement at the observation-level step, this readily accommodates variable sample sizes across the studies as we did not need to assume $n = n_1 = ... = n_k$ to achieve the desired equivalence.

4. Simulations.

4.1. *Simulation Framework.* To probe the performance of our proposed methods, we generated 100 iterations, each with 16 training studies ($K = 16$) one test set, under various conditions. We simulated studies with 20 covariates ($p = 20$) and sample sizes of $n_k = 400$ for all studies. We arbitrarily set 10 of these true model coefficients to be exactly 0, so as to simulate the sparse setting we encountered in the neuroscience data. We chose $p = 20$, $n_k = 400$, to replicate the ratio of $\frac{p}{n_k}$ in the neurosience data ($\frac{p}{n_k} \approx \frac{1000}{20000} = \frac{20}{400} = \frac{1}{20}$). We noticed considerable variation in the conditional distribution of $\mathbf{y}_k|\mathbb{X}_k$ in the neuroscience dataset (Figure 16) and thus we tried to emulate this in simulations. The conditional distribution of our outcome followed $\mathbf{y}_k|\mathbb{X}_k \sim N(\mathbb{X}_k\boldsymbol{\beta}_k, 1)$. We visually illustrate the general simulation framework in figure 9*a*. We randomly generated the pth model coefficient from the kth study as $\beta_{k,p} \sim N(\mu_{\boldsymbol{\beta}_p}, \sigma^2_{\boldsymbol{\beta}})$. To explore the impact of varying degrees of between-study heterogeneity in the true model, we simulated datasets where the variance ($\sigma^2_{\boldsymbol{\beta}}$)

of the hyperdistribution from which we drew the true model coefficients ranged over four levels: $\sigma^2_{\boldsymbol{\beta}} = 0.05,\ 0.25,\ 1,\ 3$. What constitutes high or low between-study heterogeneity in the true model coefficients is dependent on the study context and likely requires domain specific knowledge. In the absence of any meaningful measure to contextual the degree of heterogeneity in simulation, we based our range off of the performance of the Merging Algorithm. As shown above each figure in 8, the performance of the Merging Algorithm range from nearly perfect performance (RMSE is bounded below by 1 since $Var\left(\mathbf{y}_k|\mathbb{X}_k\right) = 1$), to roughy 150 times that, suggesting that we achieved a wide range of scenarios that captured "easy" problems all the way up to arguably unrealistically "hard" problems. We describe further parameter choices in the supplement (Supplement B.1).

We also tested the impact of study clusters on the performance of our proposed methods. We opted to explore this because the data application that motivated the current work appeared to exhibit study clusters (Figure 4). Clusters were modeled by generating studies with similar true model coefficients and similar distributions of the covariates across studies within a cluster. The true model coefficients differed across studies within a cluster by a degree of noise that was proportional to the between-study variability in model coefficients. If $\boldsymbol{\beta}^*_c$ is a mean vector of true model coefficients for studies in the cth cluster, then the true model coefficients for the jth study in the cth cluster is $\boldsymbol{\beta}_c = \boldsymbol{\beta}^*_c + \boldsymbol{\epsilon}_{\boldsymbol{\tau}_{c,j}}$ where $\boldsymbol{\epsilon}_{\boldsymbol{\tau}_{c,j}}$ is drawn from $Unif(-\sigma^2_{\boldsymbol{\beta}}/5,\ \sigma^2_{\boldsymbol{\beta}}/5)$. Noise was added within clusters to ensure the studies within a cluster did not have the same model coefficients or covariate distributions. We simulated studies that had the following number of clusters $c = 0,\ 4$ ($c = 0$ indicates that each study was generated from a distinct distribution). The number of training and test studies was selected to allow for even studies/cluster numbers. The simulation scheme for the clusters is visually depicted in Figure 9*b*.

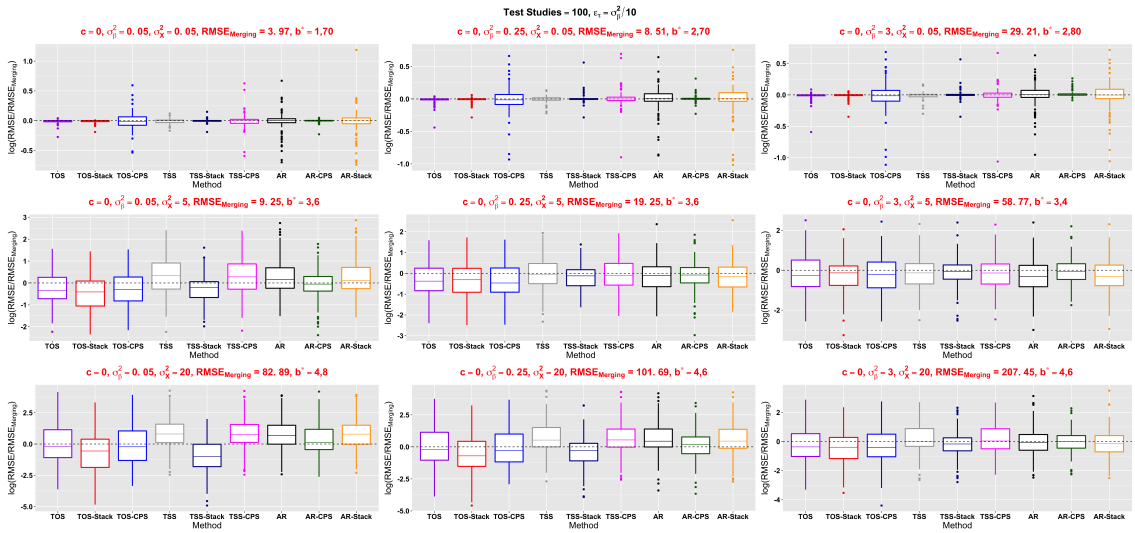
We opted to use a LASSO as our single-study learner for two reasons. First, we wanted to emulate the sparse setting of the FSCV data as well as many multi-study settings. Given the sparsity of the simulations, we then encountered the challenge that a generic similarity measure comparing covariate profile similarity would equally weight all covariates, including the covariates that did not impact the conditional distribution of $f(\mathbf{y}_k|\mathbb{X}_k)$. As such we opted to weight the covariate-wise similarity by a function of the corresponding coefficient estimates as a measure of variable importance. For this reason, we wanted to use a single-study learner that set some coefficients exactly to zero. Additionally, the LASSO is a common method used in the human FSCV literature. Specifically, we used the similarity measure: $\mathcal{S}_{K+1,(r)} = \left\| \left(\bar{\mathbf{x}}^{(r)} - \bar{\mathbf{x}}_{K+1}\right) \odot g(\hat{\boldsymbol{\beta}}) \right\|^2_2$, where $\odot$ indicates the Hadamard Product and where the pth element

$$g(\hat{\boldsymbol{\beta}})_k = \frac{1}{K}\left(\sum_{k=1}^{K} \frac{|\hat{\beta}_{k,p}|}{\hat{\sigma}^2_{k,p}} * \frac{1}{K}\sum_{k=1}^{K} \hat{\sigma}^2_{k,p}\right)$$

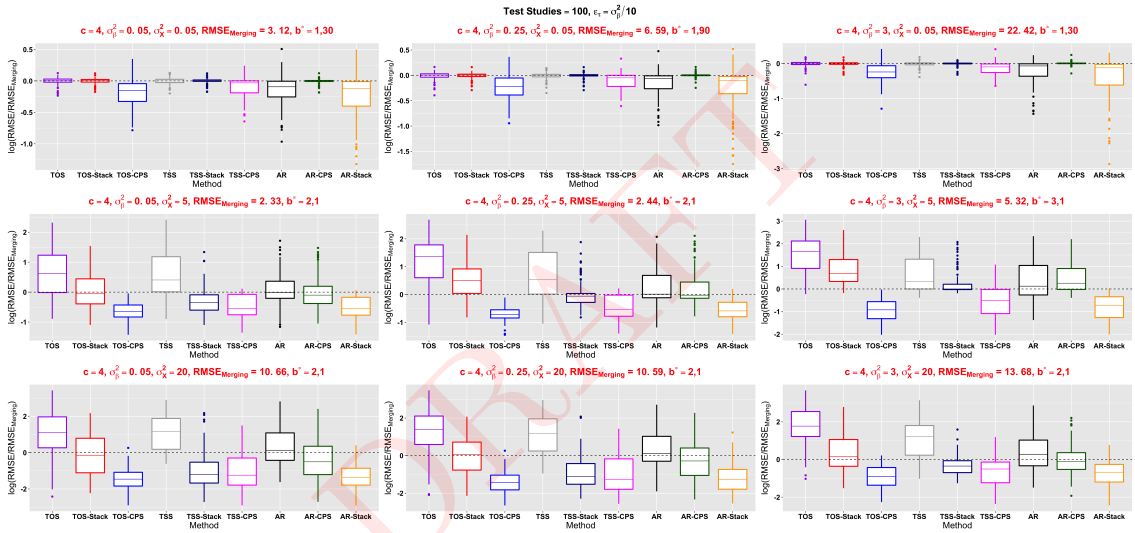
where we estimated the variance of $\hat{\beta}_{k,p}$, $\hat{\sigma}^2_{k,p}$, using a bootstrap with 500 iterations. The motivation behind including $\hat{\sigma}^2_{k,p}$ was to ensure that noisier studies that had greater variance for the coefficient estimates were penalized.

4.2. *Simulation Results.* We present the results of our simulations in Figure 8 and in Table 2. Figure 8 is on the log scale to allow for easier comparison of the methods. Versions on the original scale are presented in the supplement (Figure 9).

DRAFT



(a) No clusters



(b) Clusters

DRAFT

Fig 2: *Simulation results (on log scale) across different cluster sizes and between-study variability (σ_β) in $\boldsymbol{\beta}$. "$RMSE_{Merging}$" provides the raw mean RMSE of the Merging Algorithm for each simulation. Each observation in a plot is the out-of-study-RMSE from a single test study (i.e., each whisker plot is comprised of 100 points). b* indicates the selected bag size (from tuning) for the Trained-on-Study-Straps Ensemble and AR, respectively.*

TABLE 1

Simulation Results ($RMSE/RMSE_{Merging}$) without clusters. Bold indicates superior performance; no bolded entry in a row indicates the Merging Algorithm was superior. The top row of the methods titles indicates ensembling architecture and the bottom indicates weighting.

	Observed Studies			**Study Straps**			**Accept/Reject**		
σ^2_β (σ^2_X)	**Avg**	**CPS**	**STA**	**Avg**	**CPS**	**STA**	**Avg**	**CPS**	**STA**
0.05 (0.05)	**98.76**	100.56	98.95	98.94	100.39	99.68	99.66	100.92	99.70
0.25	**98.28**	100.66	98.92	99.64	102.09	100.82	101.36	101.27	100.64
1.00	**97.92**	100.52	98.67	98.81	100.21	100.72	101.00	102.77	100.94
3.00	**98.18**	101.04	98.98	98.67	100.32	100.86	101.13	101.48	101.66
0.05 (5.00)	105.19	99.30	**84.74**	219.88	212.99	93.76	199.23	201.55	113.50
0.25	108.29	103.86	109.48	138.45	137.46	**99.73**	120.01	122.23	115.95
1.00	121.24	116.66	112.84	125.43	124.83	106.04	108.97	111.04	111.29
3.00	135.72	128.05	123.02	124.95	124.45	129.70	115.74	115.40	124.39
0.05 (20.00)	360.45	301.22	209.33	607.49	578.26	**95.48**	500.63	500.83	422.97
0.25	314.00	276.41	165.67	533.43	507.98	162.63	461.94	465.96	226.79
1.00	184.47	169.93	126.02	296.91	289.39	148.80	281.57	276.38	200.85
3.00	155.26	142.04	121.43	213.76	204.58	140.30	185.40	182.55	147.10

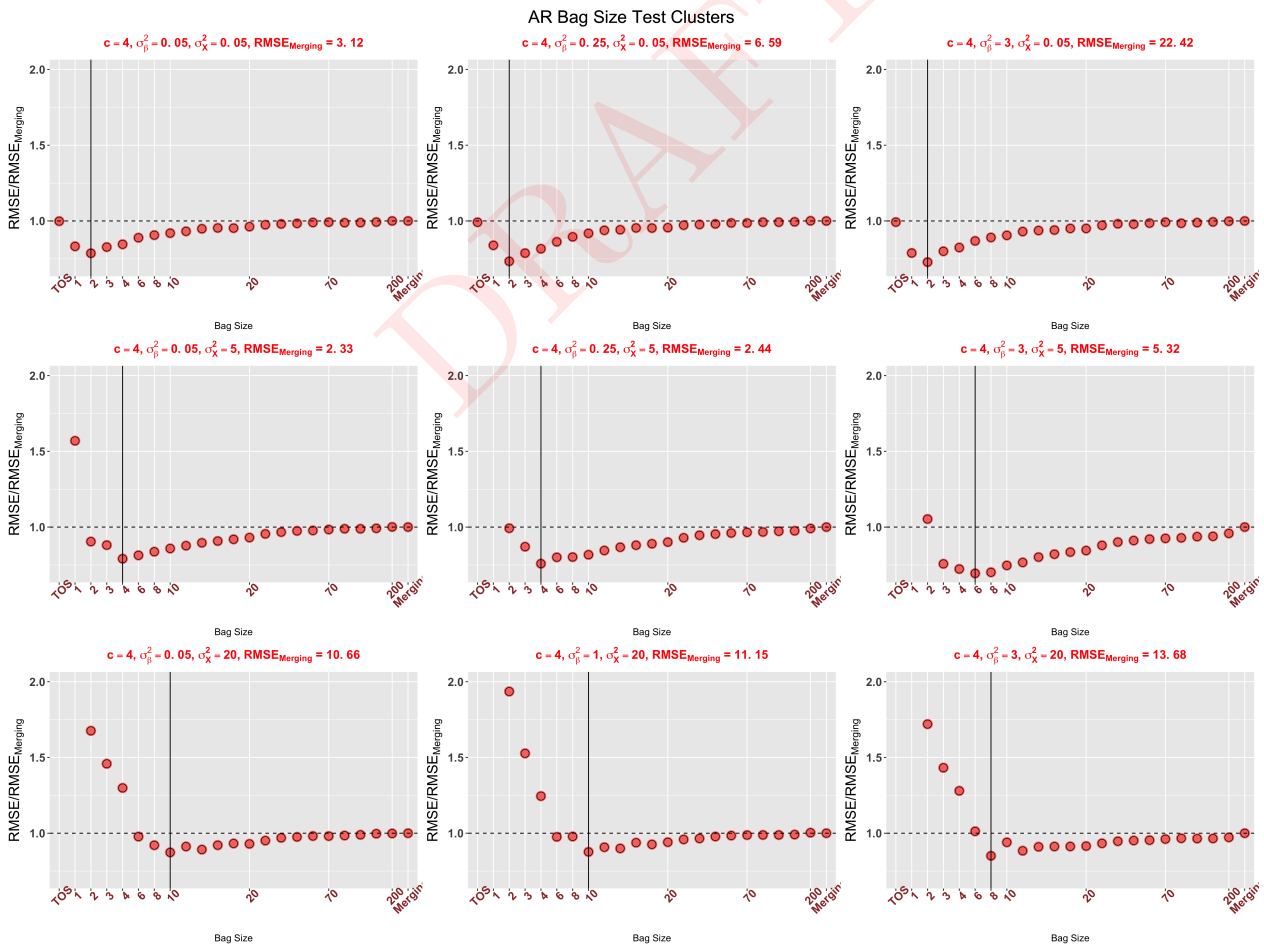


Fig 3: Optimal bag size rises with increasing σ^2_β. AR algorithm bag size curve: RMSE is on test study and standardized by performance of the Merging Algorithm.Vertical line indicates optimal bag size.

DRAFT

TABLE 2

Simulation Results ($RMSE/RMSE_{Merging}$) with clusters. Bold indicates superior performance; no bolded entry in a row indicates the Merging Algorithm was superior. The top row of the methods titles indicates ensembling architecture and the bottom indicates weighting.

	Observed Studies			**Study Straps**			**Accept/Reject**		
$\boldsymbol{\sigma^2_\beta}$ ($\boldsymbol{\sigma^2_X}$)	**Avg**	**CPS**	**STA**	**Avg**	**CPS**	**STA**	**Avg**	**CPS**	**STA**
0.05 (0.05)	99.86	85.36	99.72	100.09	91.49	99.91	88.50	**83.20**	99.43
0.25	99.14	82.17	99.15	99.32	89.91	99.36	87.81	**81.96**	100.25
1.00	99.26	**81.30**	99.38	99.01	88.52	99.37	89.04	82.70	100.68
3.00	99.25	80.96	99.34	98.51	88.27	99.30	84.34	**76.84**	100.72
0.05 (5.00)	249.40	**55.33**	126.16	252.50	66.23	79.92	129.08	**64.46**	115.39
0.25	429.57	**50.57**	209.95	277.69	67.20	112.96	171.13	61.42	159.29
1.00	582.50	**42.07**	284.23	286.62	72.35	134.45	183.37	53.69	200.28
3.00	592.96	**42.70**	292.49	259.56	68.73	149.05	201.08	52.59	209.21
0.05 (20.00)	510.08	**28.07**	168.63	465.12	61.10	69.09	247.79	36.40	135.17
0.25	601.44	**28.42**	174.46	496.59	61.38	68.20	209.33	**41.52**	130.42
1.00	749.50	**35.47**	207.99	500.03	62.03	93.81	231.30	45.17	159.31
3.00	818.70	**47.40**	223.60	470.57	70.08	77.90	246.55	55.87	151.13

DRAFT

As expected, the average performance of all methods monotonically deteriorates as between-study variability in true model coefficients ($\sigma_{\boldsymbol{\beta}}$) increases. Similarly, when there is clustering in the studies, performance improves; this is expected since increasing the number of studies per cluster functionally reduces the level of between-study variability in true model coefficients. Indeed, the degree of between-study variability within a cluster pales in comparison to that of the between-study variability across clusters.

Consistent with past results, the TOS is superior to Merging in a majority of the test studies when there are no clusters ($c = 0$) and there is moderate to high heterogeneity in model coefficients ($\sigma^2_\beta > 0.05$).

CPS weighting confers substantial benefit when there is clustering in the data. The Trained-on-Study-Straps Ensemble produces variable benefits above the TOS when there is clustering, but shows the most promise when combined with CPS weights.

The AR algorithm with average weights is superior to other forms of ensembling (TOS and SS) when there is clustering. Interestingly, it is only when σ^2_β is low, that we see clear benefits above Merging.

When tuning the study strap and accept/reject step algorithms, we used a hold-one-study-out cross validation scheme within the *training* studies. We generated bag size performance curves but unlike the tuning step, we generated these post-hoc where the RMSEs are from the test sets. Here we include one for the accept/reject algorithm (Figure 3) with clusters here to emphasize the utility of bag size tuning: as σ^2_X grows, so does the optimal bag size. We emphasize this relationship further in the supplement (figure 14). Moreover, the performance curve underscores the continuum between the TOS and the Merging Algorithm generated through varying the bag size, b. The performance of high bag sizes is nearly identical to that of the Merging Algorithm and likewise the performance of low bag sizes approaches that of the TOS as anticipated by the Study Strap framework. The bag size that was used in the accept/reject algorithm and the study strap are indicated on each plot in figure 8. The tuning step did not always choose the optimal bag size, but weighted predictions appear less sensitive to correct bag size specification.

For brevity, we only include the AR bag size for the clusters here, but have included all figures in the supplement including with the performance of weighting (Supplementary Materials B). Importantly, weighting with CPS or stacking appears to be most beneficial for lower bag sizes; as the bag size becomes large and approaches the Merging Algorithm, the performance of average, CPS and stacking weights appear to converge to the performance of Merging. Moreover, the relationship between optimal bag size and σ_β^2 is far less clear for the Trained-on-Study-Straps Ensemble as well as for the AR without clusters. The relationship between between-study heterogeneity and optimal bag size appears complex and warrants further study.

We point out that more sophisticated bag size tuning schemes may be an important area of future research: in almost every case the tuned bag size for the AR and Trained-on-Study-Straps Ensemble algorithms were suboptimal. This could be because of cross validation bias. Additionally, this may have arisen because we did not correct for the ratio of the b to K when moving from the hold-one-study-out cross validation scheme (where $K_{train} = 15$ since one study is held out) to the actual training setting (where $K = 16$).

Importantly, the benefit of either CPS or Stacking weighting schemes diminishes as the bag size increases and approaches the Merging Algorithm (Figure 3). This is likely because as b rises, the degree of variability between study straps diminishes since study straps will be more likely to be composed of roughly equal proportions from all studies.

Taken together, these simulations demonstrate that CPS weights and the Study Strap framework confers substantial benefit when there are clusters of data that share similar marginal covariate distributions and similar conditional distributions of the outcome given the covariates. Moreover, the simulations broadly demonstrate the benefit of the bag size tuning parameter, b, as the improvement in predictive performance of the Trained-on-Study-Straps Ensemble and AR varied considerably between bag sizes (Figure 3). Importantly, the simulations offer empirical evidence that the bag size tuning parameter creates a continuum between the TOS and Merging Algorithm approaches.

5. Data Application: Neurochemical measurements. We now move to the primary motivation for the above described methods. The purpose of this work was to improve the cross-electrode generalizability of models trained on in vitro data. The scientific motivation for this multi-study approach was to develop methods that would generate more reliable neurochemical concentration estimates from measurements made in the brain. The need for this arises since electrodes used to generate in vitro training sets differed from those used for brain recordings and training sets were generated purely from in vitro data. Specifically we sought to improve the predictive performance of models trained on a set of in vitro training sets each generated on a different electrode (i.e., studies), and validated on a training set generated on another electrode.

5.1. *Modeling and Methods Application.* We applied our proposed methods to these data using Principal Component Regression (PCR) ([Rodeberg, et al., 2017]) as the single study learner since it is the standard in the field. We point out, however, that unlike most applications of PCR in FSCV, we selected the number of principal components to retain using cross validation as opposed to using a measure of the proportion of variance explained. While regularization methods have also been proposed ([Kishida et al., 2016]; [Moran, et al., 2018]), we found that PCR produced superior cross-study predictive performance (Figure 18). While we did not explicitly account for the functional nature of the data, we point out that we implemented functional data analytic methods such as functional regression and functional Principal Components Regression (fPCR) using basis splines and found that it exhibited inferior performance in this application. We tuned the number of principal components to keep in the case of each held-out-study (electrode) separately. This was based upon a hold-one-study out cross validation scheme excluding the test study with the

DRAFT

merging approach (i.e., each test study was not used in their respective tuning stage for training or validation). The number of components selected for each held-out-study was held constant across all of the methods explored. The final methods were validated with a hold-one-study-out validation scheme to estimate out-of-study prediction error. Since the outcome is chemical concentration, we imposed the non-negativity constraint: $\hat{y}_i^+ = \hat{y}_i \mathbb{1}_{\hat{\mathbf{y}}_i > 0}$. We implemented this constraint so that predictions would be scientifically coherent in any case, although it only slightly impacted the net predictions/performance.

The vector of covariates was high dimensional and the vector of true model coefficients was sparse. For this reason, we opted to engineer our features to summarize average differences in the covariates between the studies. Let $\mathbb{X}_k$ be the design matrix of the kth study. As we describe in greater detail (Supplement C.2), we engineered features via a dimension reduction step where for each study $\boldsymbol{\nu}_k = g(\mathbb{X}_k)$, $\boldsymbol{\nu}_k \in \mathbb{R}^8$ is a vector features engineered to provide a summary of the covariate profile of the kth study. We then used as a similarity measure $\mathcal{S}_{K+1,k} = ||\boldsymbol{\nu}_{K+1} - \boldsymbol{\nu}_k||_2^2$ (Figure 17). We used this distance measure taken on the raw covariates when fitting models on both the original covariates and a numerical estimate of the derivative of the covaraites. We estimated the derivative of the covariates using the `diff()` function in R, since the measurements were taken at evenly spaced intervals (in time and in voltage potential intervals).

Given the inherently stochastic nature of our methods, we noticed some between-seed variability in results for the accept/reject step method. For that reason, we ran the methods on 10 separate seeds and present the results averaged across seeds. We show the variability in average performance between seeds in the supplement (Figure 20).

As the FSCV data were large, memory considerations even on a powerful cluster were a constant challenge. For that reason, we were unable to systematically implement stacking in all of our analyses given the additional computational cost. We did explore the use of stacking weights and found that it provided no additional benefit. Given the additional computational cost, we opted not to implement it. We have included a figure in the supplement exploring the performance of stacking weights in the AR algorithm (Figure 19).

For the bag size tuning curve, we ran all bag sizes on the accept/reject algorithim on a single path and with a subset of data ($n_k = 2500$) due to the high computational cost associated with using the full dataset. We tuned our bag sizes $b = K = 14$ for the raw and $b = 75$ for the derivative analyses.

5.2. *Results.* As mentioned above, we noticed clustering in the average covariates of the studies and provided a motivation for emulating this in the simulations.

DRAFT

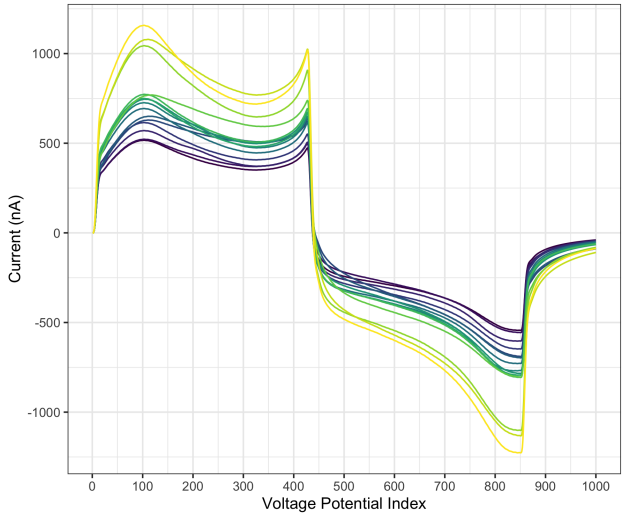
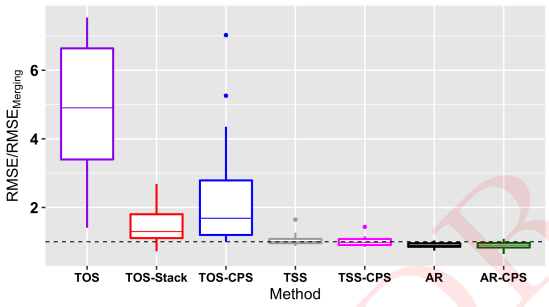
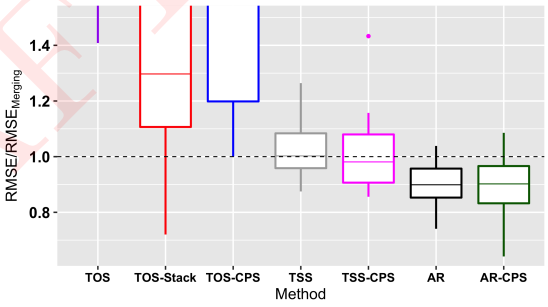


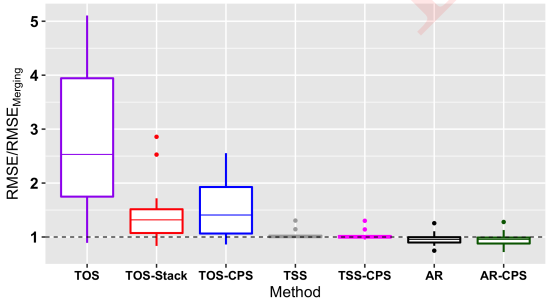
Fig 4: Studies exhibit clustering in average covariates. X-axis indicates covariate number and y-axis is the covariate value. Each line plots the average covariates of the entire study (electrode). The average magnitude of the covariates of each study scales the color in a continuous fashion. This demonstrates the clustering of the studies.



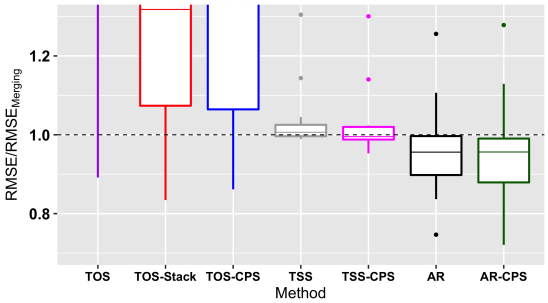
(a) Raw: $RMSE_{Merging} = 416.53$



(b) Raw zoomed in.



(c) Derivative: $RMSE_{Merging} = 359.64$



(d) Derivative zoomed in.

Fig 5: Predictive performance of methods. TOS: Trained-on-Original-Studies with average weights; TSS: Trained-on-Study-Straps with average weights; AR: Accept/Reject with average weights. -CPS indicates CPS weights and -Stack indicates stacking weights. Dotted line shows the performance of the Merging Algorithm for each held-out study. Average RMSE of using the Merging Algorithm is indicated in (a) and (c)

DRAFT

TABLE 3

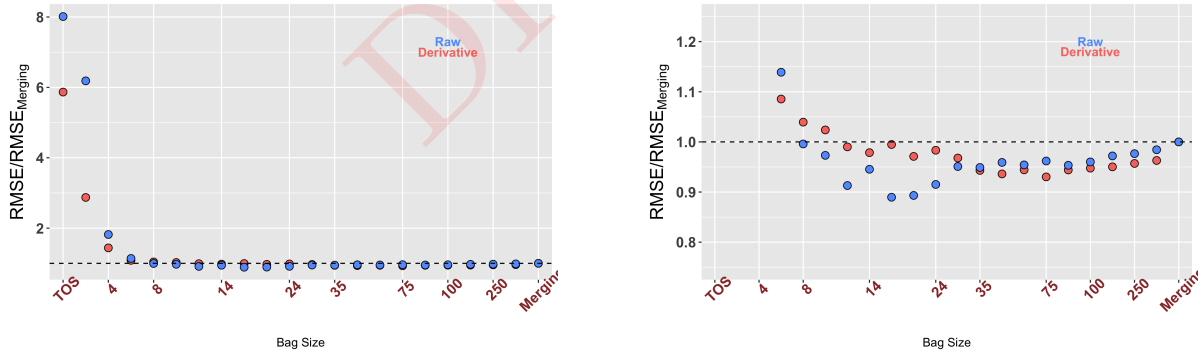
Neuroscience Data Results ($RMSE/RMSE_{Merging}$) for number of clusters (c) and between-study variability ($\sigma_{\boldsymbol{\beta}}$) in $\boldsymbol{\beta}$. Bold indicates superior performance. The top row of the methods titles indicates ensembling architecture and the bottom indicates weighting.

Covariate	Observed Studies			Study Strap		Accept/Reject	
Processing	Avg	CPS	Stack	Avg	CPS	Avg	CPS
Raw	4.788	2.435	1.449	1.056	1.014	0.893	**0.885**
Derivative	2.808	1.506	1.427	1.036	1.027	0.958	**0.951**

The results above demonstrate the strength of the Covariate Profile Similarity weighting scheme. While the Trained-on-Observed-Studies Ensemble with equal weights, and the weighted versions of it (CPS weights and Stacking weights) perform dramatically worse than Merging, the degree of improvement from the simple average weights is substantial. The CPS weights produced a 53.9% and 60.7% reduction in RMSE relative to simple average weights for the raw and derivative respectively. While these weights are inferior to stacking, the covariate profile similarity measure allows for the integration of an accept/reject step within the Study Strap framework. Indeed, the accept/reject step exhibited the strongest performance among all the methods examined. Importantly, this demonstrates that even when Merging is superior to the TOS, that ensembling approaches can out-compete Merging in some scenarios.

Importantly, the standardization of the covariates that the derivative pre-processing provides (Figure 15) shifts the optimal bag size towards the Merging Algorithm. As apparent in figure 6, the optimal bag size lies between roughly 12 and 24 for the raw data, and between 35 and 85 for the derivative. The preprocessing thus pushes the optimal bag-size towards the Merging Algorithm. Thus, the present data provides a within-dataset demonstration of the utility of the bag size: as between-study heterogeneity changes, the optimal bag size necessary to accommodate the heterogeneity also shifts.

DRAFT



(a) (b) Zoomed in.

Fig 6: Derivative shifts the optimal bag size towards the Merging Algorithm (towards higher b). RMSEs are standardized to the RMSE of the $Merging_{Raw}$ and $Merging_{Derivative}$ respectively. The dotted line indicates average performance of the $Merging$.

Consistent with the observation that the optimal bag size is shifted towards the Merging Algorithm in the derivative case, is that the benefit of the accept/reject step is reduced. This echos the observation in the simulations that as the bag size increases, the benefit of the SS, AR or any weighting scheme diminishes (Figure 10, Figure 12). The derivative provides a between-study

standardization of the covariates and improves the performance of all methods substantially. This enhancement in performance across-the-board may also contribute to the discrepancy in performance improvement between the raw and derivative analyses.

6. Discussion. Motivated by challenges of transportability of prediction models in emerging neuroscience technologies, we propose a general multi-study learning framework including a resampling scheme and complementary weighting procedure. These methods produced substantial improvements in performance when there is clustering of studies. Importantly, in these settings, Merging was superior to the Trained-on-Observed-Studies Ensemble (with or without stacking weights). However, when we ensembled and applied Covariate Profile Similarity weighting and/or the accept/reject algorithm, these vastly outperformed Merging, highlighting that ensembling approaches can sometimes outperform Merging when between study heterogeneity is low. It appears that a Covariate Profile Similarity approach, whether applied through CPS weights or the accept/reject algorithm, is most helpful when the "borders" of the studies are murky. That is, when there is substantial clustering of the studies, one might be better off delineating between clusters instead of between the original observed studies, themselves. When the studies do not exhibit clustering, it appears that the standard TOS or Merging techniques are superior.

By expressing the TOS and the Merging Algorithm as special cases of the Study Strap, we propose a unifying framework for multi-study prediction. We suggest a theoretical foundation for why the bag size can help adjust for varying levels of heterogeneity. Using a numerical estimate of the derivative for feature engineering appears to produce cross-study standardization and shifts the optimal bag-size to greater values (i.e., towards the Merging Algorithm) compared to the optimal bag-size of the raw covariates. This example serves to demonstrate the utility of the bag size tuning parameter in accounting for between-study heterogeneity. We emphasize that the benefit of the Covariate-Matched Study Strap (AR) appears to rely on ensembling at its core. For example, using the final study strap accepted prior to convergence as the sole learner performed worse than ensembling all accepted study straps in the neuroscience data. This is the case even though the last accepted replicate has a covariate profile more similar to the target study than all previously accepted study straps. Interestingly, including a burn-in step (eliminating early accepted study straps) did not appear to provide any systematic benefit and could actually be detrimental depending on the degree of burn in. We point out that we used 10 paths in our accept/reject step algorithm to be conservative, but we noticed little benefit after 5 paths. Moreover, the average performance does not change from running one path, but instead served to reduce between-run variability in performance.

We explored a variety of similarity measures in the in the initial stages of this work, although we present the results from only one for the sake of brevity and consistency. Matrix correlation packages and canonical correlation analysis techniques are out-of-the-box methods that exhibited promise in the simulations we explored. Our `studyStrap` packagae, provides over 20 of such generic similarity measures that are automatically computed for any of the ensembling methods. Context-specific similarity measures (as we developed for the neuroscience application) may be preferable if subject-matter knowledge can guide their development. But even in the absence of any straight-forward method of constructing such metrics, more general measures, such as that used in the simulations, may be viable in a range of scenarios.

The Study Strap framework allows for a variety of weighting schemes including stacking, CPS and out-of-study-bag weighting. It not only introduces flexibility in how study replicates and ensembles are formed, but also in how stacking can be executed. For example, we have proposed a number of stacking approaches that differ in the manner in which one constructs the design matrix and outcome vector used in the stacking regression. We outline a couple in the supplement that are enabled by the Study Strap framework and may be suited to different scenarios (Sections A.2 and A.3). Although this was not the focus of the paper, we point this out as a potential area of future

DRAFT

research.

Importantly, the FSCV results demonstrate that despite the statistical challenges present in the application of FSCV in humans, substantial improvements in cross-electrode generalizability are feasible. Statistical research into this topic is critical as FSCV in humans is a new technique and has, in some domains, unprecedented capacity to provide insight into the brain, behavior and cognition. Moreover, FSCV in humans marks an important victory for statistics in science: novel applications of statistical learning methods were instrumental in its successful implementation in humans. What is more, these statistical methods have enabled researchers to use FSCV to measure multiple neurotransmitters simultaneously (), a feat thought not to be possible with this technology, even in preclinical settings ([Moran, et al., 2018];).

[17] consider heterogeneity in multi-study settings, with the aim of identifying a subset of studies within which comparison of prediction algorithms can be reliably carried out, and comparing algorithms using a merged subset of the studies. They address both problems by integrating clustering of studies and model comparison through a Bayesian model for the array of cross-study validation statistics. The combination of study-strap and covariate similarity weighting is a natural evolution of this approach to multi-study ensembles. While implementation is not Bayesian, and clustering is not explicit, the resampling and reweighting jointly will make is to that the out-of-study predictions will be based on the subset of studies most similar in both covariate similarity and covariate-outcome relations.

6.0.1. *Covariate Shift.* The Covariate Shift literature contains a rich array of prediction methods for settings in which there exist differences in the distribution of the data between training and test sets. Investigators distinguish between three situations: 1) "virtual drift" (also sometimes referred to as simply "covariate shift") occurs when differences in the distribution of the covariates exist (i.e., $f[\mathbb{X}_{train}] \neq f[\mathbb{X}_{test}]$) but the conditional distribution remains constant (i.e., $f[\mathbf{Y}_{train}|\mathbb{X}_{train}] = f[\mathbf{Y}_{test}|\mathbb{X}_{test}]$); 2) "real drifts" refer to contexts in which the conditional probability, $f[\mathbf{Y}_t|\mathbb{X}_t]$), changes over time independent of changes in the marginal distribution of the covaraites, $f[\mathbb{X}_t]$; 3) "hybrid drifts" present when both $f[\mathbf{Y}_t|\mathbb{X}_t]$ and $f[\mathbb{X}_t]$ change over time ([Yang, et al., 2019]).

In the present work, we focus on a multi-study analogue of the hybrid drift literature. The multi-study framework does not posit, however, any meaningful ordering (temporal or otherwise) among the studies and focuses on data that are often from different sources. This is in contrast to much of the concept/hybrid drift literature that deal with streams of data (often time series) that change over time (in one of the manners described above) and are typically from one source, such as environmental or energy data (Yang et al., 2019; Raza et al., 2019; Karnick et al., 2008; Ditzler et al., 2010), (i.e., $f(\mathbb{X}_t)$ and/or $f(\mathbf{Y}_t|\mathbb{X}_t)$ change over time). As such, many concept drift methods employ approaches that are inherently tailored to the temporal nature of the data streams, such as moving averages ([Raza, et al., 2008]), detection systems to identify the presence ([Yang, et al., 2019]) or speed ([Minku, et al., 2009]) of a shift at a given instance, and ensemble methods that add, remove, or reweight classifiers across time ([Raza, et al., 2008]; [Minku, et al., 2009]). While virtual drift work often does not assume changes over time, its methods usually center on reweighting observations ([Sugiyama, et al., 2008]; ([Shimodaira, 2000]) and assume that $f(\mathbf{Y}|\mathbb{X})$ remains constant between training and test sets. These differences likely explain why there has been little implementation of covariate shift methods in multi-study learning.

6.1. *Limitations.* Our methods and analyses are not without limitations. To begin, this work does not provide an explicit analytical expression for prediction error (RMSE) as a function of the bag size parameter, b. As such, tuning the bag size parameter, b, must be conducted empirically with cross validation. Indeed, our simulations and data application reveal that the bag size parameter is helpful in accommodating between-study heterogeneity, but do not reveal the exact nature of het-

DRAFT

erogeneity for which it benefits. While it appears to help account for heterogeneity in the marginal distribution of the covaraites, σ_X^2) in the simulations with clusters and the neuroscience data, it is clear that the relationship is neither linear nor monotonic and the contours of the relationship appears to vary considerably between simulation settings. Finally, although the accept/reject step showed the most promise in the neuroscience data, it is also the most computationally expensive. More efficient algorithms that achieve a similar goal may allow the application of this technique in a wider range of settings. Future work may seek to achieve this through an optimization step as opposed to random re-sampling.

As mentioned above, we did not account for the functional nature of the neuroscience data explicitly because we found that the functional methods that we implemented (e.g., fPCR with basis splines) produced inferior performance. In the future, investigators may wish to experiment with other functional approaches (e.g., wavelet regression) as they may produce gains in predictive accuracy. Moreover, we did not explicitly account for the temporal nature of the neuroscience data as this was not the goal of the present work. We point out that no standard methods in the FSCV field, to our knowledge, account for the time-series nature of the data in estimating neurochemical concentration ([Rodeberg, et al., 2017]; [Kishida et al., 2016]; [Moran, et al., 2018]). Nevertheless, future work might consider implementing methods that account for this feature of the data structure as it could be exploited for gains in performance.

Finally, although we have shown progress in enhancing between-electrode generalizability of the models, we cannot verify the generalizability from in vitro models to in vivo, as this would require an additional gold standard measure of neurochemical concentartion in the human brain for which assays does not currently exist. We point out that such a gold standard does not exist even in animal models where a greater array of invasive methods are available. It is our hope that by enhancing cross-electrode generalizability that generalization to the brain is improved.

6.2. *Conclusions.* Taken together, the Study Strap, Covariate Profile Similarity Weighting and the Covariate-Matched Study Strap are flexible ensembling and weighting schemes that can improve predictive performance in multi-study settings. We hope the methodology will benefit the statistics, neuroscience and computational psychiatry ([Montague, et al., 2012];) communities. We have provided a user-friendly R package that can be used by investigators that use FSCV as well as other analysts who encounter multi-study prediction settings. We argue that the present work contributes to the multi-study literature both in its proposed methods and theoretically through the general Study Strap framework.

ACKNOWLEDGEMENTS

SUPPLEMENTARY MATERIAL

A.1. Stacking Strategy. We used non-negative least squares with an intercept to generate weights. That is, we used the weights from

$$\underset{\boldsymbol{w}_{K+1\times 1}}{\text{minimize}} \quad \|\boldsymbol{y}_S - \mathbb{X}_S^* \boldsymbol{w}\|_2^2$$
$$\text{subject to: w} \geq 0$$

where $\mathbb{X}_S^* = [\mathbb{1}\mathbb{X}_S]$. We found that standardizing the weights led to a degradation of performance

DRAFT

and so we proceeded without standardizing the coefficient estimates. Thus, the final predictions are:

$$\hat{\mathbf{Y}}_{Stacking} = \hat{w}_0 * \mathbb{1} + \sum_{k=1}^{K} \hat{w}_k \hat{\mathbf{Y}}_{\mathbf{k}}(\mathbb{X}_{K+1})$$

A.2. Study Strap Stacking Strategy. This is the Study Strap (or AR) analogue of the stacking that was implemented for the TOS. This regresses $\mathbb{X}_S$ on $\mathbf{y}_S$, where

$$\mathbf{y}_S = \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \vdots \\ \mathbf{y}^{(R)} \end{bmatrix}_{N^* \times 1} \qquad \mathbb{X}_S = \begin{bmatrix} \hat{\mathbf{Y}}^{(\mathbf{1})}(\mathbb{X}^{(1)}) & \hat{\mathbf{Y}}^{(\mathbf{2})}(\mathbb{X}^{(1)}) & \dots & \hat{\mathbf{Y}}^{(\mathbf{R})}(\mathbb{X}^{(1)}) \\ \hat{\mathbf{Y}}^{(\mathbf{1})}(\mathbb{X}^{(2)}) & \hat{\mathbf{Y}}^{(\mathbf{2})}(\mathbb{X}^{(2)}) & \dots & \hat{\mathbf{Y}}^{(\mathbf{R})}(\mathbb{X}^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{Y}}^{(\mathbf{1})}(\mathbb{X}^{(R)}) & \hat{\mathbf{Y}}^{(\mathbf{2})}(\mathbb{X}^{(R)}) & \dots & \hat{\mathbf{Y}}^{(\mathbf{K})}(\mathbb{X}^{(R)}) \end{bmatrix}_{N^* \times R}$$

and $\hat{\mathbf{Y}}^{(\mathbf{r})}(\mathbb{X}^{(j)})$ are the predictions on the design matrix of study strap j using the model trained on study strap r; $\mathbf{y}^{(j)}$ are the labels from study strap r. $N^* = \sum_{r=1}^{R} n^{(r)}$ The stacking procedure then proceeds as above.

A.3. Study Strap Stacking Strategy 2. Below is an additional way one could construct the stacking regression for Study Strap (or AR). This regresses $\mathbb{X}_S$ on $\mathbf{y}_S$, where

$$\mathbf{y}_S = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_K \end{bmatrix}_{N \times 1} \qquad \mathbb{X}_S = \begin{bmatrix} \hat{\mathbf{Y}}^{(\mathbf{1})}(\mathbb{X}_1) & \hat{\mathbf{Y}}^{(\mathbf{2})}(\mathbb{X}_1) & \dots & \hat{\mathbf{Y}}^{(\mathbf{R})}(\mathbb{X}_1) \\ \hat{\mathbf{Y}}^{(\mathbf{1})}(\mathbb{X}_2) & \hat{\mathbf{Y}}^{(\mathbf{2})}(\mathbb{X}_2) & \dots & \hat{\mathbf{Y}}^{(\mathbf{R})}(\mathbb{X}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{Y}}^{(\mathbf{1})}(\mathbb{X}_K) & \hat{\mathbf{Y}}^{(\mathbf{2})}(\mathbb{X}_K) & \dots & \hat{\mathbf{Y}}^{(\mathbf{K})}(\mathbb{X}_K) \end{bmatrix}_{N \times K}$$

and $\hat{\mathbf{Y}}^{(\mathbf{r})}(\mathbb{X}_k)$ are the predictions on the design matrix of training study k using the model trained on study strap r; $\mathbf{y}_k$ are the labels from training study k. The stacking procedure then proceeds as above.

One could similarly construct the stacking regression to weight the TOS or standard Trained-on-Study-Straps Ensemble using the design matrices of the accepted study straps in the AR algorithm (and the models from the TOS or standard TOS respectively). The above variations on standard stacking highlight the flexibility introduced by the Study Strap framework.

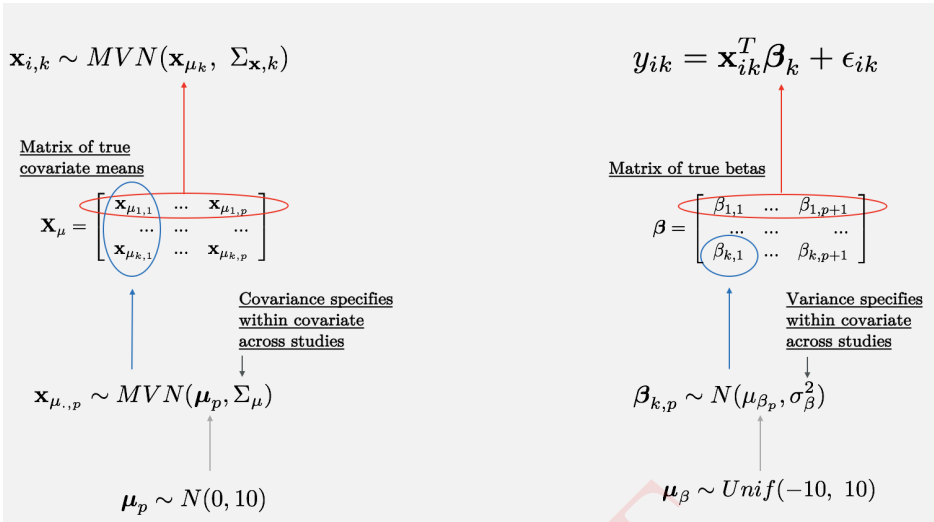
APPENDIX B: SIMULATIONS SUPPLEMENTARY MATERIALS

B.1. Simulation Parameters. Here we describe additiomal details related to the simulations. When tuning the λ hyperparameter for the LASSO, we used a hold-one-study out cross validation scheme and used the same λ across studies and methods within a given simulation iteration. In other words, we used the same λ for all methods (i.e., Merging, TOS, SS, AR) so that this did not influence our between-methods comparisons. We tested 43 values of λ between 0.0001 and 5. When tuning the study strap for the bag size, we implemented 150 study straps at each of 21 bag sizes between $b = 1$ and $b = N$. We used a hold-one-study-out cross validation scheme to tune the bag size. When generating the study strap bag size performance curve we used 250 study straps per bag size. When testing the final Study Strap with the tuned b, we used 500 study straps. We used less study straps during tuning stages in order to avoid the computational cost (since we were doing a hold-one-study-out cross validation scheme, it was computationally intensive).

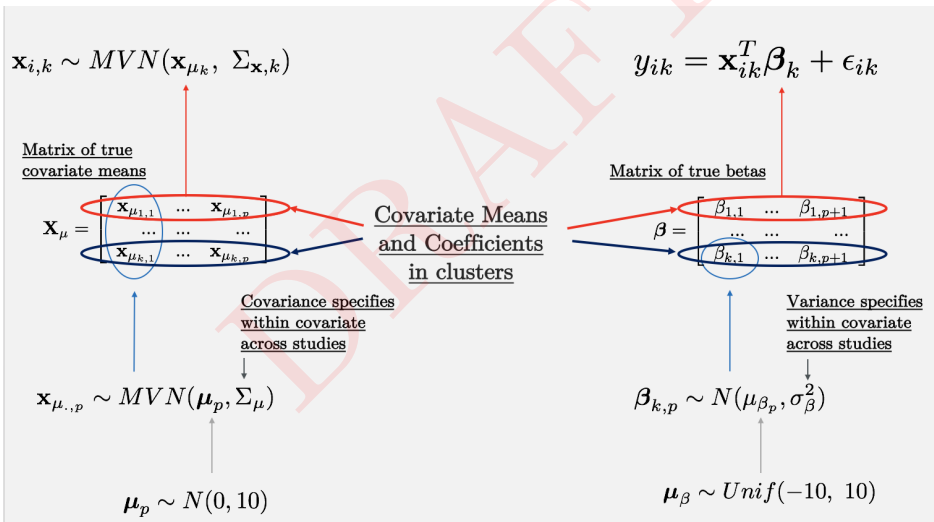
When we were tuning the AR bag size, we used a convergence criteria of 10000 consecutive study straps without an acceptance. We averaged across 3 paths. We used the same parameters for the

DRAFT

AR bag size performance curve. For the final AR implementation with the tuned b, we used 5 paths and a convergence criteria of 100000 consecutive study straps without an acceptance. We felt these parameters struck the balance of being computationally feasible and also sufficient for our purposes.



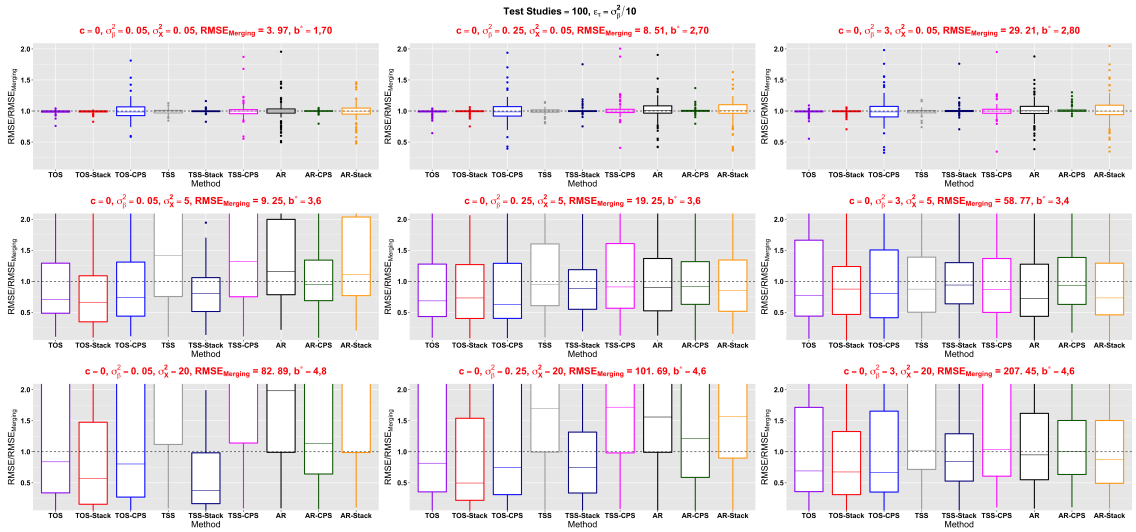
(a)



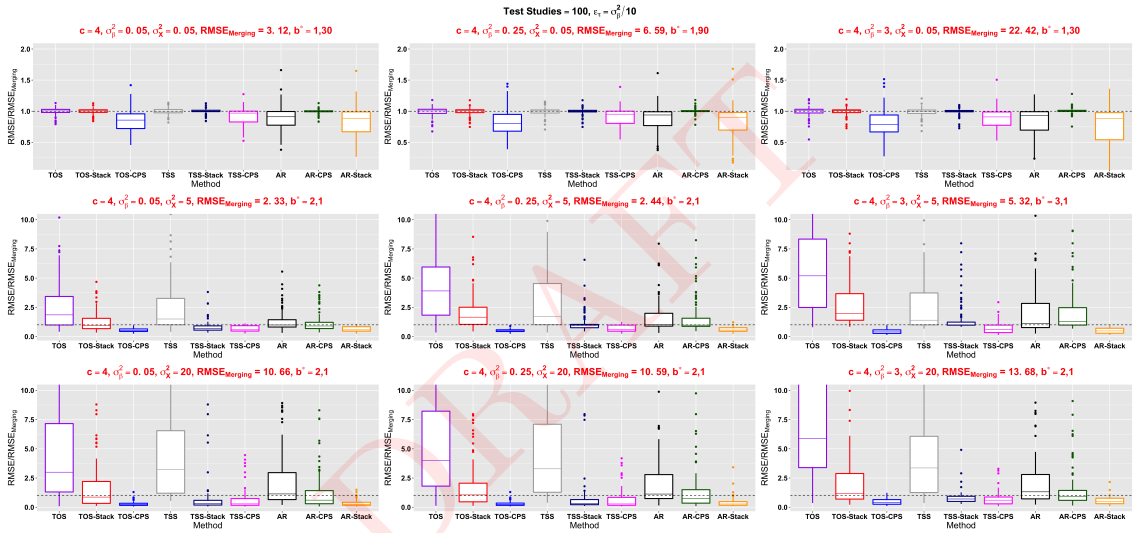
(b)

DRAFT

Fig 7: (A) General simulation framework. (B) Simulation scheme for clusters.



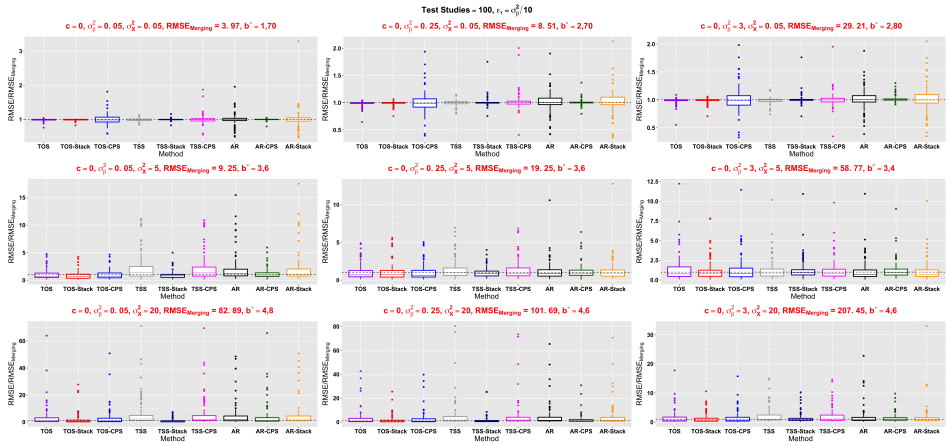
(a) No clusters



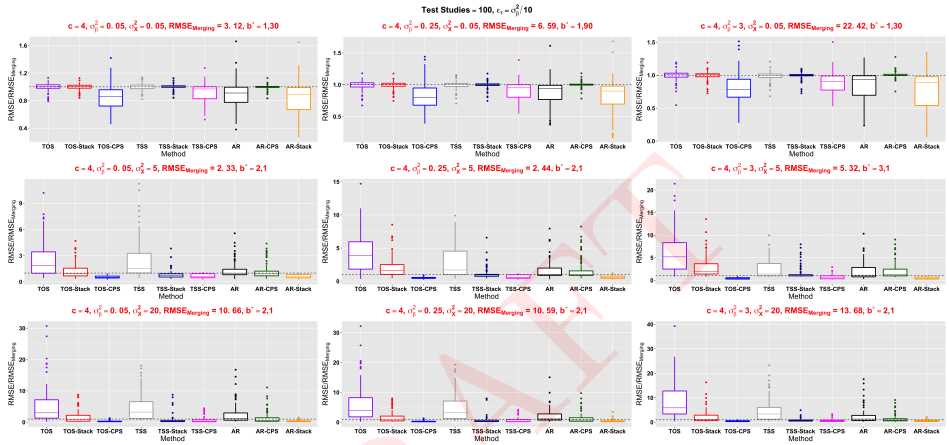
(b) Clusters

DRAFT

Fig 8: Simulation results on original scale (zoomed-in) across different cluster sizes and between-study variability ($\sigma_{\boldsymbol{\beta}}$) in $\boldsymbol{\beta}$. Each observation in a plot is the out-of-study-RMSE from a single test study (i.e., each whisker plot is comprised of 100 points). b^* indicates the selected bag size (from tuning) for the Trained-on-Study-Straps Ensemble and AR, respectively.



(a) No Clusters.



(b) Clusters

Fig 9: Simulation results on original scale across different cluster sizes and between-study variability ($\sigma_{\boldsymbol{\beta}}$) in $\boldsymbol{\beta}$. Each observation in a plot is the out-of-study-RMSE from a single test study (i.e., each whisker plot is comprised of 100 points). b^* indicates the selected bag size (from tuning) for the Trained-on-Study-Straps Ensemble and AR, respectively.

DRAFT

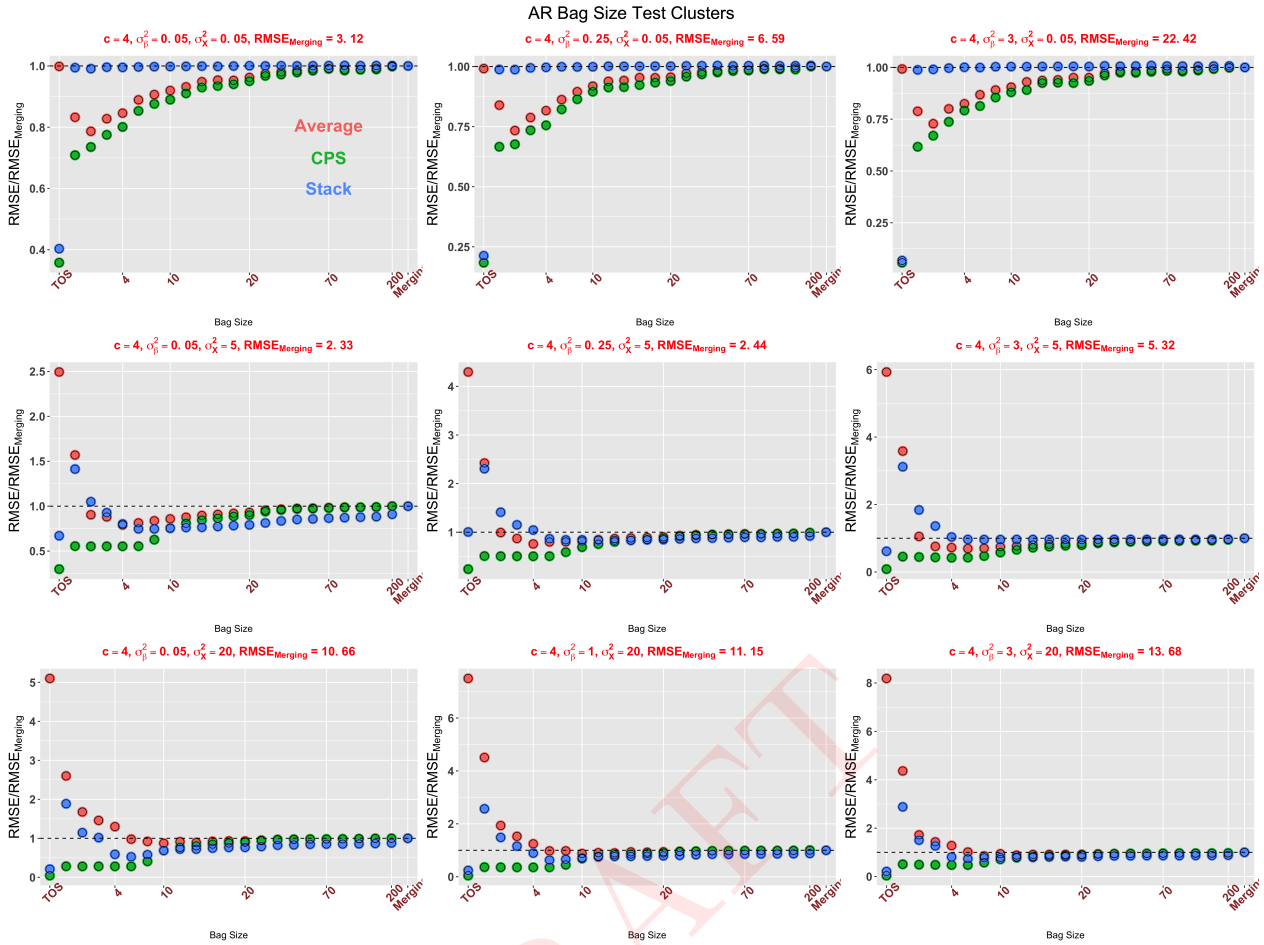


Fig 10: AR Bag size curve with clusters. Performance (RMSE) from test set and standardized to performance with Merging.

DRAFT

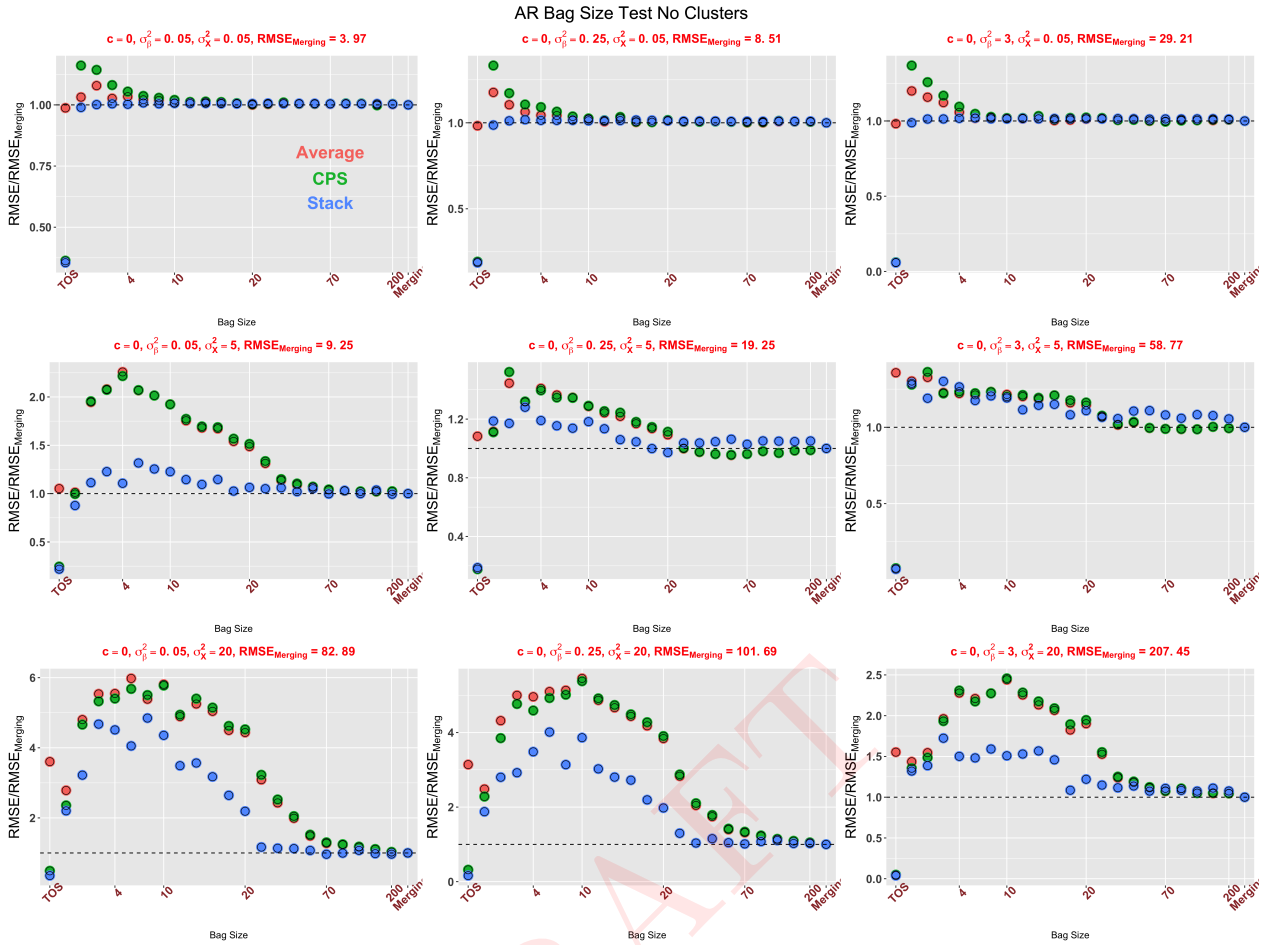


Fig 11: AR Bag size curve without clusters. Performance (RMSE) from test set and standardized to performance with Merging.

DRAFT

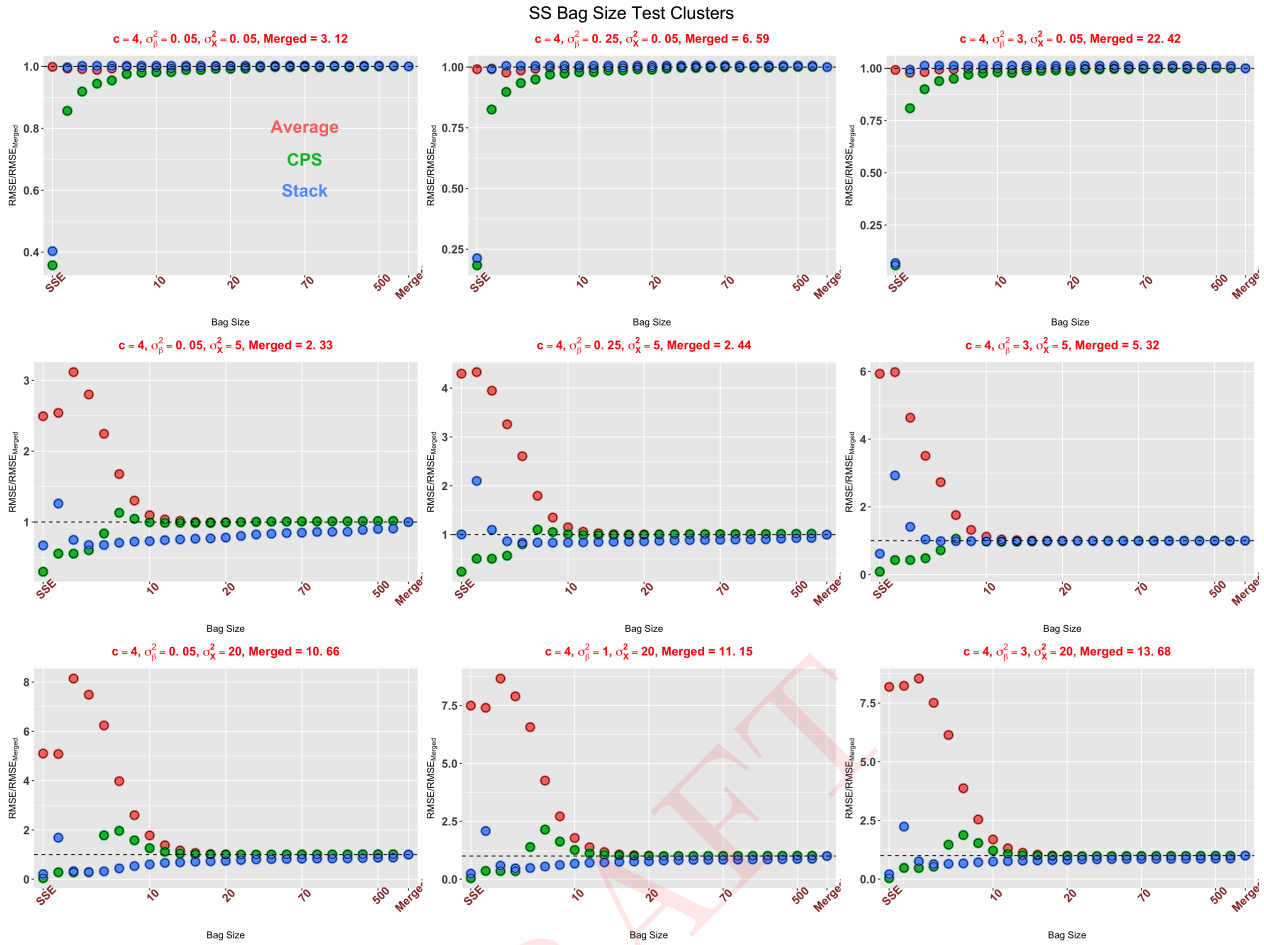


Fig 12: Trained-on-Study-Straps Ensemble Bag size curve with clusters. Performance (RMSE) from test set and standardized to performance with Merging.

DRAFT

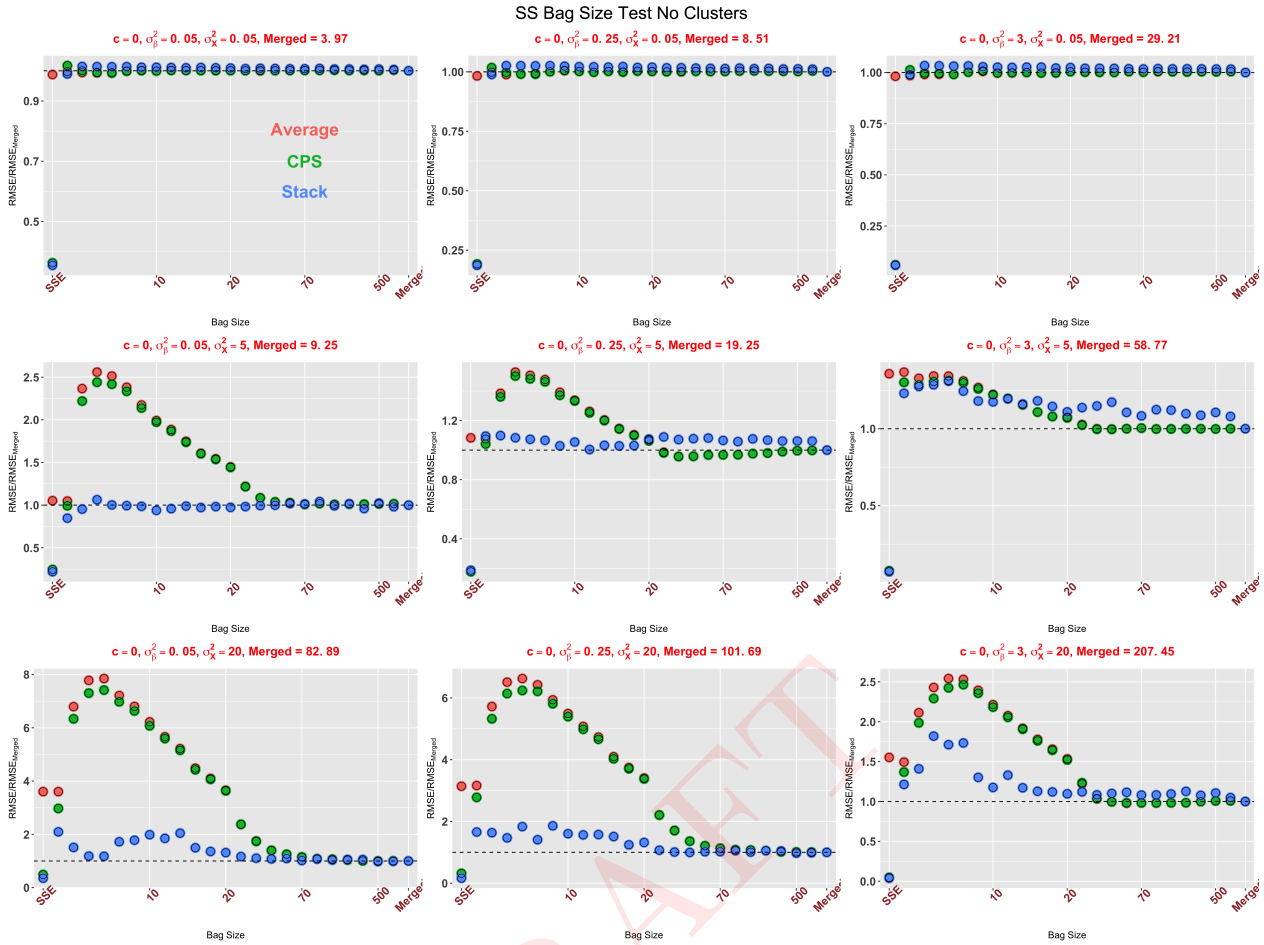


Fig 13: Trained-on-Study-Straps Ensemble Bag size curve without clusters. Performance (RMSE) from test set and standardized to performance with Merging.

DRAFT

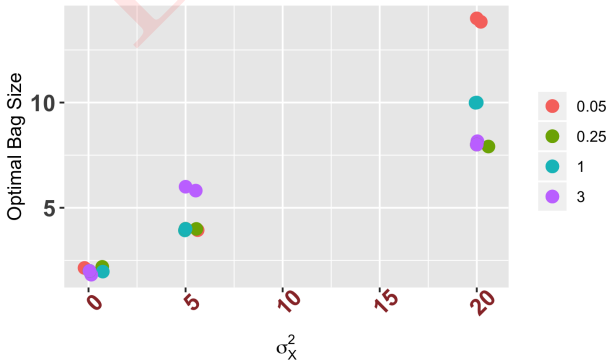


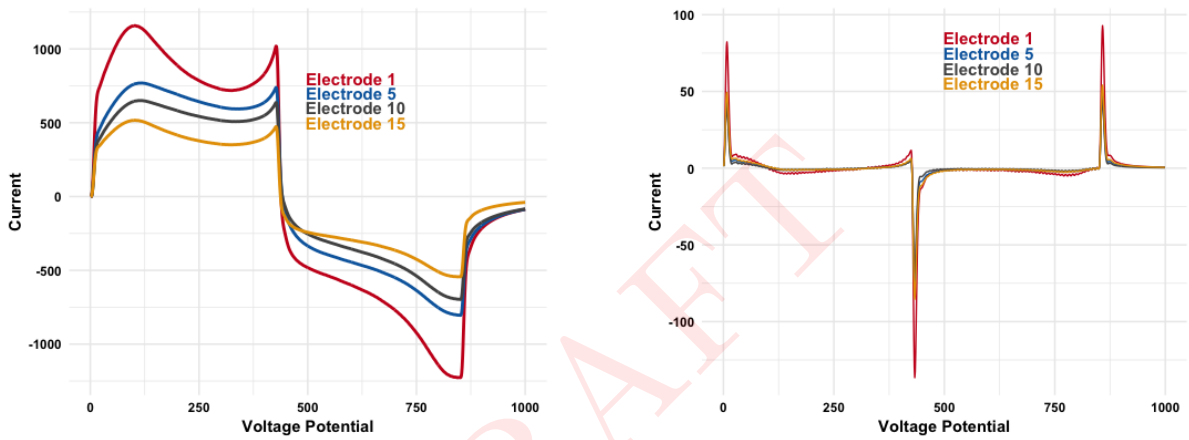
Fig 14: AR algorithm bag size curve: Optimal bag size scales with σ^2_X. Color indicates σ^2_β. Minor noise was added to points for clarity.

B.2. Extra Simulation Figures.

APPENDIX C: NEUROSCIENCE DATA

C.1. Data Description. The data is collected by exposing an electrode in a flow cell (in vitro) to different prepared concentrations (c_j) of a neurotransmitter, dopamine, a naturally occurring brain chemical (where the concentration is known to the investigator). Many measurements are taken at a given concentration (these different measurements are referred to as "Replicate" below). While measurements are taken across time, time is ignored in the training set (i.e., each observation is treated as independent). The rows of the dataset then correspond to a measurement (observation) at a given concentration, at a given time point. Furthermore, these data are paired with a vector of labels of known neurotransmitter concentrations ($[DA]_i$). The structure of the data for the kth electrode is presented in Table 4.

A sample of the average covariates of four of the 15 electrodes is presented in Figure for both the raw and derivative pre-processing versions. The figure highlights the standardization in covariates that the derivative provides.



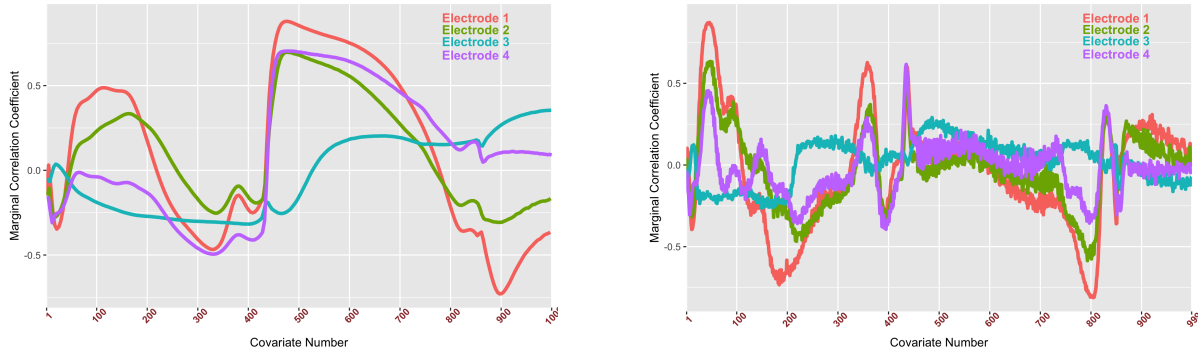
(a) Raw covariates.

(b) A numerical estiamte of the derivative of the average covariates of 4 studies.

Fig 15: Between-electrode variability: a comparison of average CVs (the covariates) of four different electrodes/studies. The x-axis is the covariate index (i.e., the jth Voltage Potential is the jth covariate) and the y-axis is the magnitude of the corresponding covariate. Data are presented here since the 1000 covariates can also be viewed as a single functional covariate (they are all on the same scale and units).

C.2. Similarity Metric. In order to compare the covariates of two studies, we engineered features that summarize the covariate profile of a given observation. These features are the coordinates of the inflection points of the CVs (Figure 17b): $\boldsymbol{\nu}$ contains theses coordinates collapsed into a single vector (i.e., $\boldsymbol{\nu} \in \mathbb{R}^8$). Then to compare, the ith and j^{th} studies, we calculate the distance between the average $\boldsymbol{\nu}$ of each study via the similarity metric, $s_{ij} = ||\boldsymbol{\nu}_i - \boldsymbol{\nu}_j||_2^2$ (Figure 17a). Studies/electrodes differed not only in the average magnitude of the covariates (i.e., the height of the figures), but also the covariates that were concentration-sensitive (i.e., which coefficients, $\boldsymbol{\beta}$ were non-zero). This measure was designed to account for this by measuring distance in terms the covariate index and height.

DRAFT



(a) Raw. (b) Derivative.

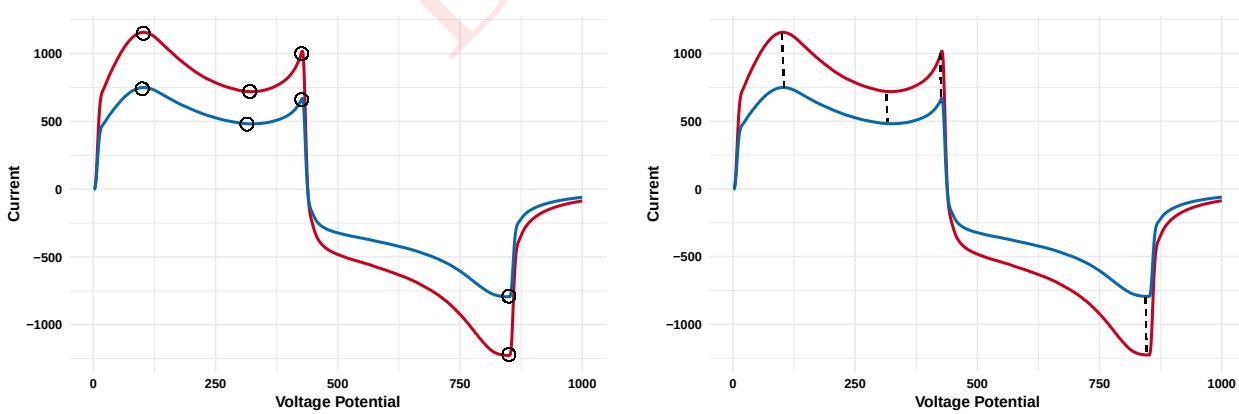
Fig 16: Marginal correlation coefficient estimates ($\hat{\rho}_{k,p}$) from four example studies for each covariate ($p = 1000$).

[DA]	Replicate	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\cdots$	$\boldsymbol{x}_{1000}$
c_1	1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,1000}$
c_1	1	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,1000}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
c_1	700	$\cdots$	$\cdots$	$\cdots$	$\cdots$
c_2	1	$\cdots$	$\cdots$	$\cdots$	$\cdots$
c_2	2	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
c_2	700	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
c_{30}	700	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,1000}$

TABLE 4

Structure of the data for the kth electrode/study

DRAFT



(a) (b)

Fig 17: (A) Inflection points of the average of the two studies. (B) Similarity metric: sum of the squared lengths of the diagonal dotted lines.

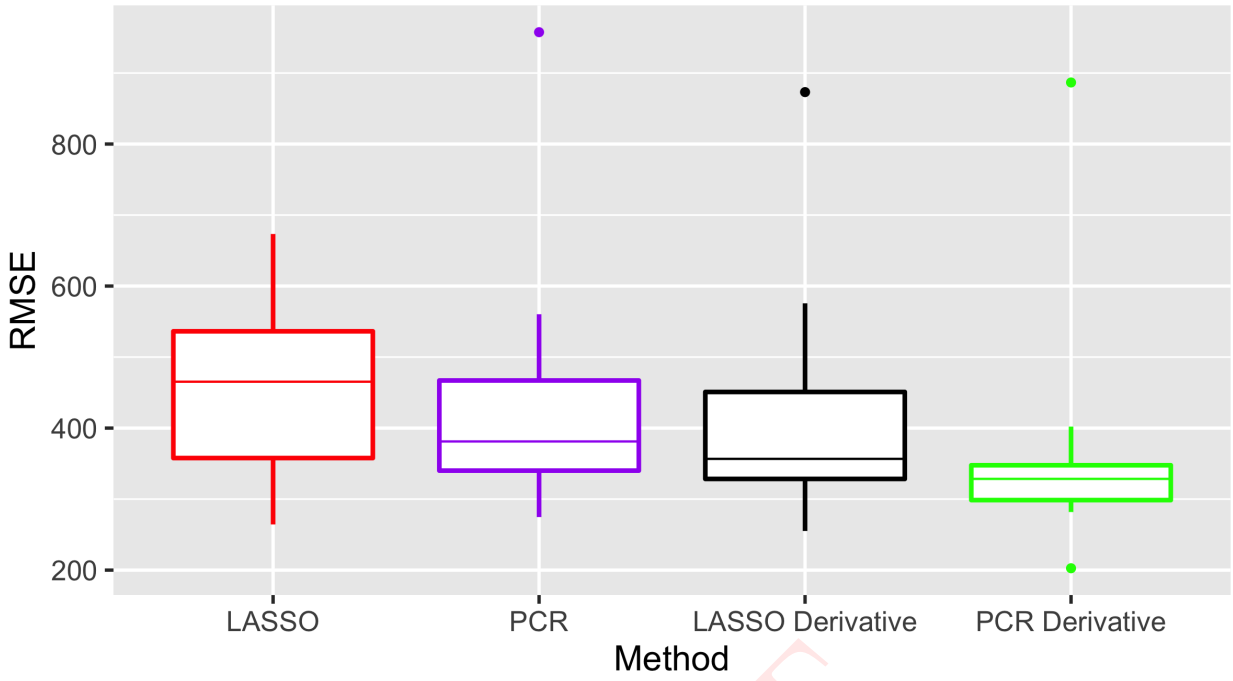
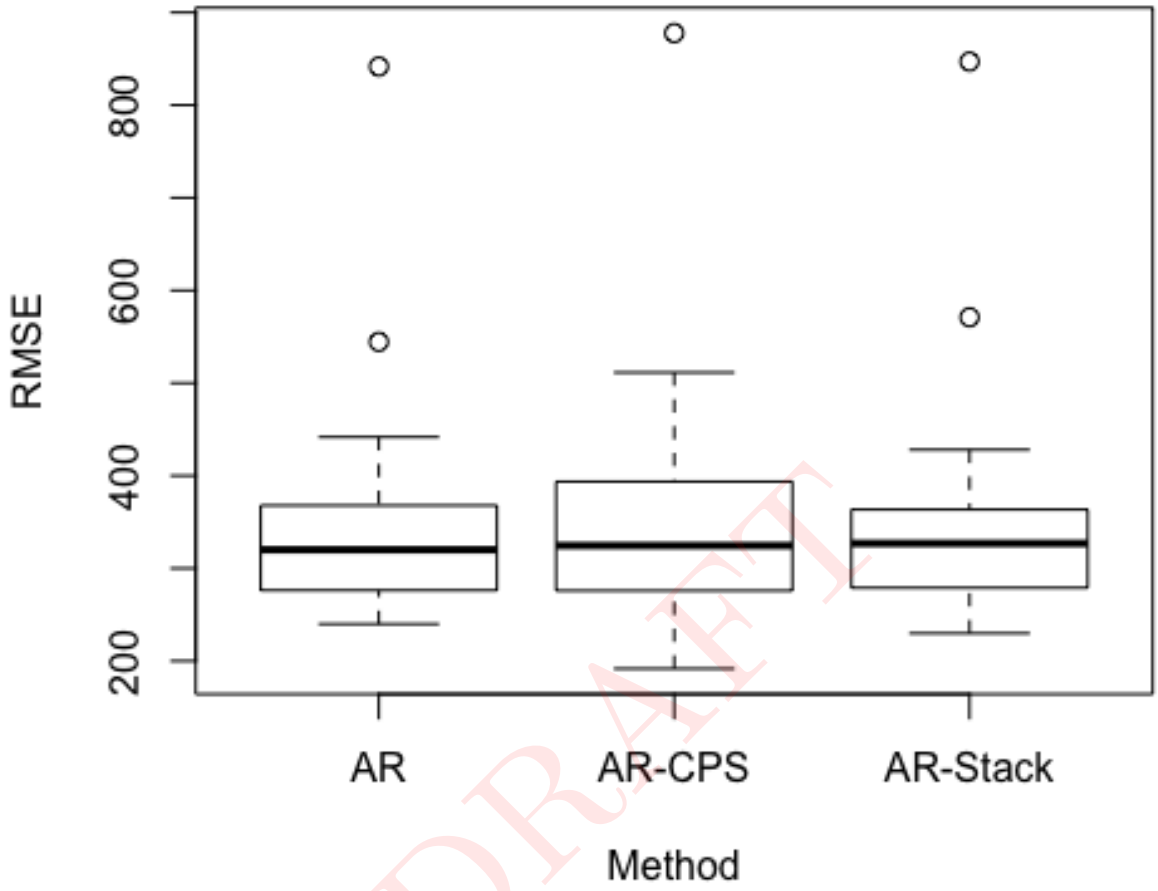


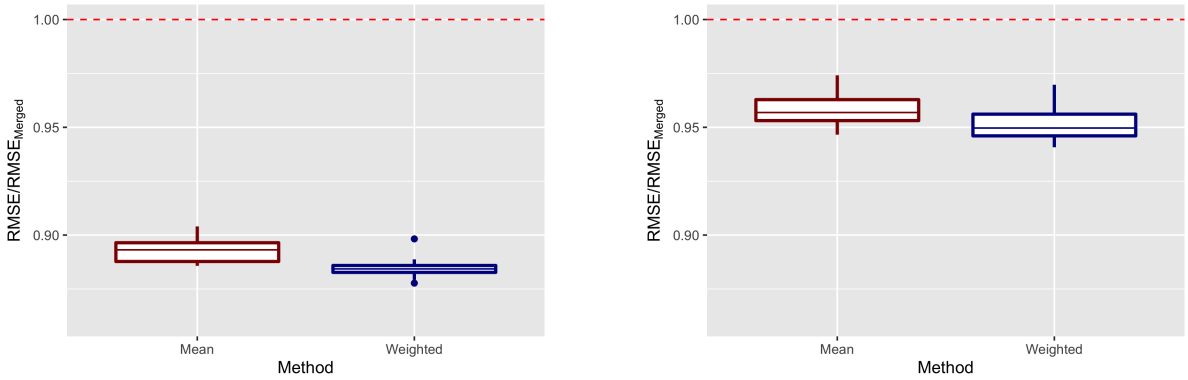
Fig 18: Predictive performance of Merging approaches for both the raw and derivative data preprocessing. LASSO and PCR hyperparameters both tuned via hold-one-study-out CV.

DRAFT



DRAFT

Fig 19: Predictive performance of AR, AR-CPS and AR-Stack on the raw data, $b = 14$.



(a) Raw. (b) Derivative.

Fig 20: Between-seed variability in average performance (relative to the Merging Algorithm) Each point is the mean $RMSE/RMSE_{Merging}$ for a single seed, averaged across all held-out-studies.

().

REFERENCES

[1] Breiman, L. (1996). Bagging Predictors. *Machine Learning* **24** 123–140.
[2] Breiman, L. (1996). Stacked Regressions. *Machine Learning* **24** 49–64.
[3] Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.
[4] Ditzler, G., Polikar, R. (2010). An ensemble based incremental learning framework for concept drift and class imbalance. *2010 Int. Joint Conf. on Neural Nets* 1–8.
[5] Engel, A., Moll, C., Fried, I., Ojemann., G. (2018). Invasive Recordings from the Human Brain: Clinical Insights and Beyond. *Nat. Neurosc. Rev.* **6** 35–47.
[6] Guan, Z., Parmigiani, G. and Patil, P. (2019). Merging versus Ensembling in Multi-Study Machine Learning: Theoretical Insight from Random Effects. **arXiv:1905.07382**
[7] Karnick, M., Muhlbaier, M., and Polikar, R. (2008). Incremental learning in non-stationary environments with concept drift using a multiple classifier based approach. *19th Int. Conf. on Pattern Recog.* **343** 1–4.
[8] Kishida K., Sandberg S., Lohrenz T., Comair Y., Sáez I., et al. (2011). Sub-Second Dopamine Detection in Human Striatum. *PLOS ONE* 6(8): e23291.
[9] Kishida, K., Saez, I., Lohrenz, T., Witcher, M., Laxton, A., Tatter, S., White, J., Ellis, T., Phillips, P., and Montague, R. (2016). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *PNAS* **113** 200–205.
[10] Klaes, C. (2018). Invasive Brain-Computer Interfaces and Neural Recordings From Humans. *Handbook of Behav. Neurosc.* **28** 527–539.
[11] Li, L., Girguis, M., Lurmann,F., Wu, J., Urman, R., Rappaport, E., Ritz, B., Franklin, M., Breton, C., Gilliland, F., Habre, R. (2019). Cluster-based bagging of constrained mixed-effects models for high spatiotemporal resolution nitrogen oxides prediction over large regions. *Environment International* **128** 310–323.
[12] Minku, L., White, A., Yao, X. (2009). The Impact of Diversity on On-line Ensemble Learning in the Presence of Concept Drift. *IEEE Transac. on Knowledge & Data Engin.* **22** 730–742.
[13] Moran, R., Kishida, K., Lohrenz, T., Saez, I., Laxton, A., Witcher, M., Tatter, S., White, J., Ellis, T., Phillips, P., Dayan, P., Montague, R. (2018). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Neuropsychopharmacology* **43** 1425-1435.
[14] Montague, P. R., Dolan, R. J., Friston, K. J., Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences* **16** 72-80.
[15] Patil, P. and Parmigiani, G. (2018). Training replicable predictors in multiple studies. *PNAS* **115** 2578–2583.
[16] Platt, M. and Pearson, J. (2016). Dopamine: Context and counterfactuals. *PNAS* **113** 22–23.
[17] Trippa, L., Waldron, L., Huttenhower, C, Parmigiani, G. (2015). Bayesian nonparametric cross-study validation of prediction methods *Annals of Applied Statistics* **9** 402-428.
[18] Ramchandran, M., Patil, P., and Parmigiani, G. (2020). Tree-weighting for multi-study ensemble learners. *Pacific Symposium on Biocomputing 25:451-462*
[19] Raza, H., Rathee, D., Zhou, S., Cecotti, H., Prasad, G. *(2008). Covariate shift estimation based adaptive*

DRAFT

ensemble learning for handling non-stationarity in motor imagery related EEG-based brain-computer interface. Neurocomputing ***343*** *154–166.*

[20] Rodeberg, N., Sandberg, S., Johnson, J., Phillips, P., and Wightman, R. *(2017). Hitchhiker's Guide to Voltammetry: Acute and Chronic Electrodes for in Vivo Fast-Scan Cyclic Voltammetry. ACS Chem Neurosci.* ***8*** *221–234.*

[21] Shimodaira, H. *(2000). Improving predictive inference under covariate shift by weighting the log-likelihood function . J. of Stat. Planning and Inf.* ***90*** *227–244.*

[22] Sugiyama, M., Nakajima, H., Kashima, P., Von Buenau, M., Kawanabe, and M. Sugiyama *(2008). Direct importance estimation with model selection and its application to covariate shift adaptation. Adv. in Neural Info. Process. Syst* ***20*** *1433–1440.*

[23] Volkow, N., Wise, R., and Baler, R. *(2017). The dopamine motive system: implications for drug and food addiction. Nat Rev Neurosci.* ***18*** *741–752.*

[24] Yang, Z., Al-Dahidi, S., Baraldi, P., Zio, E., Montelatici, L. *(2019). A Novel Concept Drift Detection Method for Incremental Learning in Nonstationary Environments. IEEE Trans. on Neural Net. Learn. Syst.* ***DOI: 10.1109/TNNLS.2019.2900956***

Address of the First and Second authors
Usually a few lines long
E-mail: gloewinger@g.harvard.edu
kkishida@wakehealth.edu

Address of the Third author
Usually a few lines long
Usually a few lines long
E-mail: gp@jimmy.harvard.edu
URL: http://www.foo.com

DRAFT