UC Irvine
UC Irvine Electronic Theses and Dissertations

Title
Enhancing the Discovery of Neural Representations: Integrating Task-Relevant Dimensionality Reduction and Domain Adaptation

Permalink
https://escholarship.org/uc/item/6mg841f9

Author
Orouji, Seyedmehdi

Publication Date
2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Enhancing the Discovery of Neural Representations: Integrating Task-Relevant Dimensionality Reduction and Domain Adaptation

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
In Cognitive Sciences

by

Seyedmehdi Orouji

Dissertation Committee:
Associate Professor Megan Peters, Chair
Professor Emily Grossman
Professor Jeff Krichmar
Professor Ramesh Srinivasan

2024

© 2024 Seyedmehdi Orouji

DEDICATION

To

My parents, family, and friends
for their support, encouragement, and love

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge and give my warmest gratitude to my advisor Professor Megan Peters for being an outstanding mentor throughout my doctoral journey. Her dedicated support, insightful advice, and her commitment to my educational development have been invaluable for me. Without her support and expertise, this dissertation would not have been possible.

Also, I would like to express my appreciation to my committee members, Professor. Emily Grossman, Professor. Jeff Krichmar, and Professor. Ramesh Srinivasan. Their valuable feedback and guidance have been a significant help in writing this dissertation.

I would also like to thank the Department of Cognitive Science at the University of California-Irvine, for funding my studies and providing me with this great opportunity to expand my knowledge during my PhD program.

Chapter 1 of this dissertation is a reprint of the material as it appears in " “Task-relevant autoencoding” enhances machine learning for human neuroscience," posted to arXiv. The coauthors listed in this publication are Seyedmehdi Orouji, Vincent Taschereau-Dumouchel, Aurelio Cortese, Brian Odegaard, Cody Cushing, Mouslim Cherkaoui, Mitsuo Kawato, Hakwan Lau, & Megan A. K. Peters. Megan A. K. Peters directed and supervised research which forms the basis for the dissertation.

Chapter 2 of this dissertation is a reprint of the material as it appears in " Domain adaptation in small-scale and heterogeneous biological datasets," posted to arXiv. The coauthors listed in this publication are Seyedmehdi Orouji, Martin C. Liu, Tal Korem, Megan A. K. Peters. Megan A. K. Peters directed and supervised research which forms the basis for the dissertation.

VITA

Seyedmehdi Orouji

2013-15 M.S. in Electrical and Computer Engineering, University of Missouri

2016-18 Electrical Engineer, Exocytronics

2018-19 M.S. Bioengineering, University of California, Riverside

2019-24 Ph.D. Cognitive Neuroscience, University of California, Irvine

FIELD OF STUDY

Cognitive Sciences

PUBLICATIONS

Extracting task-relevant low dimensional representations under data sparsity. Proceedings of the Cognitive Computational Neuroscience Annual Meeting, 2022.

"Task-relevant autoencoding" enhances machine learning for human neuroscience. arXiv, 2022.

Domain adaptation in small-scale and heterogeneous biological datasets. arXiv 2024

ABSTRACT OF THE DISSERTATION

Enhancing the Discovery of Neural Representations: Integrating Task-Relevant Dimensionality Reduction and Domain Adaptation

By

Seyedmehdi Orouji

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2024

Professor Megan Peters, Chair

In human neuroscience, machine learning models can be used to discover lower-dimensional neural representations relevant to behavior. However, these models often require large datasets and can be overfit with the small sample sizes typical in neuroimaging. To address this, we developed the Task-Relevant Autoencoder via Classifier Enhancement (TRACE) to extract behaviorally relevant representations. When tested against standard autoencoders and principal component analysis, TRACE showed up to 12% increased classification accuracy and 56% improvement in discovering task-relevant representations using fMRI data from ventral temporal cortex (VTC) of 59 subjects, highlighting its potential for behavioral data.

Machine learning models applications also extend to predictive modeling and pattern discovery in modern biology. However, these models often fail to generalize across different datasets due to statistical differences. This issue also exists in neuroscience, where data are collected across various laboratories using different experimental setups.

Domain adaptation can align statistical distributions across datasets, enabling model transfer and mitigating overfitting issues. In the second chapter we discussed domain adaptation in the context of small-scale, heterogeneous biological data, outlining its benefits, challenges, and key methodologies. We advocate for integrating domain adaptation techniques into computational biology, with further customized developments.

Building on these insights, we used DA for understanding brain region interactions during visual processing. We examine the ventral temporal cortex (VTC) and prefrontal cortex (PFC) using Domain Adaptive Task-Relevant Autoencoding via Classifier Enhancement (DATRACE) to explore shared neural representations. DATRACE leverages domain adaptation techniques within an encoder-decoder architecture to predict voxel activities from a shared latent space, in order to ensure relevance for object recognition tasks. Preliminary results indicate that shared representations capture similar object categories in both VTC and PFC. We computed the representational dissimilarity matrix (RDM) of the shared representation between VTC and PFC and contrasted that to the RDM obtained from the low dimensional representation of VTC. Our results suggest that the nature of the information shared with PFC is very similar to those encoded in VTC. Additionally, feature perturbation analysis suggests the need for further studies to reveal the semantic interpretations of shared dimensions in these brain regions. This integrated approach underscores the potential of advanced machine learning techniques in both neuroscience and biology.

INTRODUCTION OF THE DISSERTATION

In the field of computational neuroscience, researchers are interested in the understanding of the human brain by utilizing advanced computational techniques. Recently, there has been an increasing interest among cognitive neuroscientists in using machine learning and domain adaptation in order to develop more accurate predictive models that are effective when trained on small-scale neuroimaging data[1]. Our study utilizes deep learning techniques to explore the encoding process within and between brain regions in the visual cortex.

In the first chapter we discuss the implementation of the Task-Relevant Autoencoder via Classifier Enhancement (TRACE) model[2]. TRACE is designed to remove noise and distill task-relevant information from data while dealing with the challenge of small datasets that are common in functional magnetic resonance imaging (fMRI) studies. TRACE's architecture consists of a simple autoencoder with minimal hidden layers which is specifically tailored to extract task-relevant features from neural representations. Our model not only allows the discovery of a latent low-dimensional representation of neural data but also increases the decoding accuracy even at the native voxel space. This demonstrates the potential of our method in denoising and extracting task-relevant information in fMRI data[2].

In the second chapter of this study we pay particular attention to domain adaptation (DA) and its utility in dealing with small biological datasets. In the field of computational biology, the goal is to uncover generalizable biological truths rather than finding mere statistical correlations[3]. However, the process of collecting and labeling biological data is often expensive, and time-consuming. This results in numerous small datasets gathered from different sources under varying conditions[4], such as the Autism Brain Imaging Dataset (ABIDE)[1], where data collected across multiple sites. This introduces significant challenges in data aggregation due to differences in experimental conditions. These variations create different data domains with distinct statistical distributions which pose many challenges in data curation[5,6].

Domain adaptation (DA) is a powerful tool to address the variation issue across different but related datasets[7–9]. DA was initially developed in the field of computer science to increase model performance by enabling cross-dataset information utilization. DA facilitates the transfer of knowledge from a well-labeled “source domain” to an unlabeled or poorly labeled “target domain”[7,9–11]. This is achieved by aligning the statistical distributions of the source and target domains, allowing a model trained on the source domain to accurately predict labels in the target domain [12]. In the context of brain imaging, we can take advantage of this “byoproduct” of DA in order to find a shared representation between different brain regions by treating them as different domains. This alignment helps in understanding the interconnected activity between these brain regions but also enhances our ability to interpret complex neural interactions involved in visual cognition.

In the third chapter we analyze the interactions between the ventral temporal cortex (VTC) and prefrontal cortex (PFC), which are essential for processing visual information. We propose the Domain Adaptive Task Relevant to Autoencoding via Classifier Enhancement (DATRACE) model in order to investigate shared neural representations that play roles in object recognition. We hope that this approach enhances our ability to explore and interpret the complex information transfer between brain regions.

Chapter 1. “Task-relevant autoencoding” enhances machine learning for human neuroscience

Authors: Seyedmehdi Orouji[1], Vincent Taschereau-Dumouchel[2-3], Aurelio Cortese[4], Brian Odegaard[5], Cody Cushing[6], Mouslim Cherkaoui[6], Mitsuo Kawato[4], Hakwan Lau[7], & Megan A. K. Peters[1,8]

1 Department of Cognitive Sciences, University of California, Irvine, Irvine, California, USA 92697
2 Department of Psychiatry and Addictology, Université de Montréal, Montreal, Canada, H3C 3J7.
3 Centre de recherche de l’institut universitaire en santé mentale de Montréal, Montréal, Canada.
4 ATR Computational Neuroscience Laboratories, Kyoto, Japan 619-0288
5 Department of Psychology, University of Florida, Gainesville, FL USA 32603
6 Department of Psychology, University of California Los Angeles, Los Angeles, 90095, USA
7 RIKEN Center for Brain Science, Tokyo, Japan
8 Center for the Neurobiology of Learning and Memory, University of California, Irvine, Irvine, California, USA 92697

Abstract

In human neuroscience, machine learning can help reveal lower-dimensional neural representations relevant to subjects' behavior. However, state-of-the-art models typically require large datasets to train, so are prone to overfitting on human neuroimaging data that often possess few samples but many input dimensions. Here, we capitalized on the fact that the features we seek in human neuroscience are precisely those relevant to subjects' behavior. We thus developed a Task-Relevant Autoencoder via Classifier Enhancement (TRACE), and tested its ability to extract behaviorally-relevant, separable representations compared to a standard autoencoder, a variational autoencoder, and principal component analysis for two severely truncated machine learning datasets. We then evaluated all models on fMRI data from 59 subjects who observed animals and objects. TRACE outperformed all models nearly unilaterally, showing up to 12% increased classification accuracy and up to 56% improvement in discovering "cleaner", task-relevant representations. These results showcase TRACE's potential for a wide variety of data related to human behavior.

Keywords: human neuroscience, machine learning, dimensionality reduction, task-relevant representation, fMRI, MVPA, autoencoder

1. Introduction

In studying the human brain and human behavior, we often use machine learning methods to home in on the (ideally lower-dimensional[13–16]) representations contained in multivariate, feature-rich datasets. These data typically contain noisy, irrelevant signals [17–19] that we would like to filter out using methods such as multivariate decoders [20–23], various types of autoencoders, generative adversarial networks like InfoGAN [24], or even principal components analysis (PCA) [25–27]. However, state-of-the-art machine learning methods typically require very large datasets to train while data for individual human subjects collected with methods such as functional magnetic resonance imaging (fMRI) [17–19] are often severely limited in sample size [28,29] (i.e., have very few training exemplars compared to the dimension of data). Consequently, these methods are susceptible to overfitting on such neuroimaging data, reducing their predictive power and utility [30–32]. What's more, parametric methods (such as PCA), which may better avoid the need for large training sets, by definition require rigid assumptions regarding the nature of the dimensionality reduction process and thus are limited *a priori* to insights consistent with these parametric assumptions. Thus, we are in need of a nonparametric method that can reveal the *low-dimensional, task-relevant* representations in a given brain region using *exemplar-poor but input-dimension-rich* datasets.

Here, we sought to capitalize on a unique property of many human neuroimaging datasets, which is that the features we wish to identify can be conceptualized based on whether they are relevant for the subject's behavior.

We drew inspiration from previous successes with classifier-enhanced autoencoders [33–36] to develop the Task-Relevant Autoencoder via Classifier Enhancement (TRACE) model. TRACE's architecture is purposely simple to limit overfitting to small datasets, consisting of a fully-connected autoencoder with only one hidden layer on each of the encoding and decoding arms and a logistic regression classifier attached to the bottleneck layer (**Figure 1.1**).

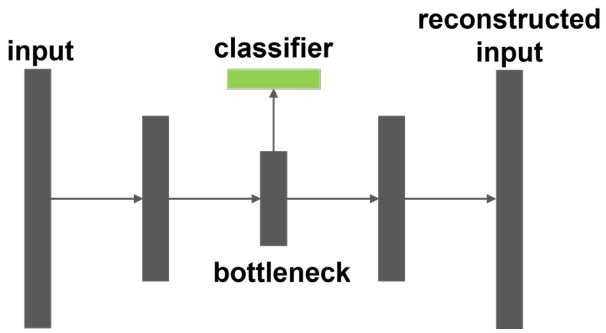

Figure 1.1. A cartoon representation of the TRACE network architecture. Each gray rectangle represents a layer of the autoencoder, consisting of fully connected units. The input layer is connected to the bottleneck layer via one hidden encoding layer, and again to the reconstruction layer via one hidden decoding layer. A classifier is attached to the bottleneck and contributes to the objective optimization function.

We developed four quantitative metrics to assess TRACE's performance at different bottleneck dimensionalities (compression levels), and then comprehensively benchmarked TRACE under conditions of severe data sparsity using the MNIST [37] and Fashion MNIST [38] datasets, two of the most popular machine learning datasets. We then applied TRACE to a neuroimaging (fMRI) dataset of subjects who viewed and categorized animals and objects while blood oxygen level dependent (BOLD) signal was collected from ventral temporal cortex (VTC) in a single, 1-hour session. By constraining the

dimensionality reduction process to specifically prioritize features that were relevant to the participants' behavioral task, we show that TRACE can extract both quantitatively and qualitatively 'cleaner' representations at both reduced dimensions and in the original input dimensionality, showing up to threefold improvement in decoding accuracy and separation of class-specific patterns. These results demonstrate our method can distill highly separable, low dimensional neural representations even with sparse and noisy data. TRACE may thus show promise on a broad variety of behaviorally-relevant neuroimaging datasets.

2. Results

We quantified the performance of the Task-Relevant Autoencoder via Classifier Enhancement (TRACE) model against that of a standard autoencoder (AE), a Variational Autoencoder (VAE), and using principal component analysis (PCA) via (1) *reconstruction fidelity,* (2) *reconstruction classifier accuracy*, (3) *bottleneck classifier accuracy,* and (4) *reconstruction class specificity* (see **Methods Section 4.4**) ("class" here refers to the class of the input image, e.g. "9" or "shoe" or "cat"). We assessed these metrics as a function of different bottleneck dimensionalities (i.e., compression levels), first on the MNIST and Fashion MNIST datasets under increasing data sparsity and then on a previously-collected fMRI dataset of ventral temporal cortex (VTC) (i.e., voxel activations while 59 human subjects viewed 40 classes of animals and objects). We also performed additional investigation at each dataset's 'optimal' bottleneck dimensionality (where reconstruction class specificity is maximized) to characterize each model's behavior.

2.1 Benchmarking TRACE's advantages, including under increasing data sparsity

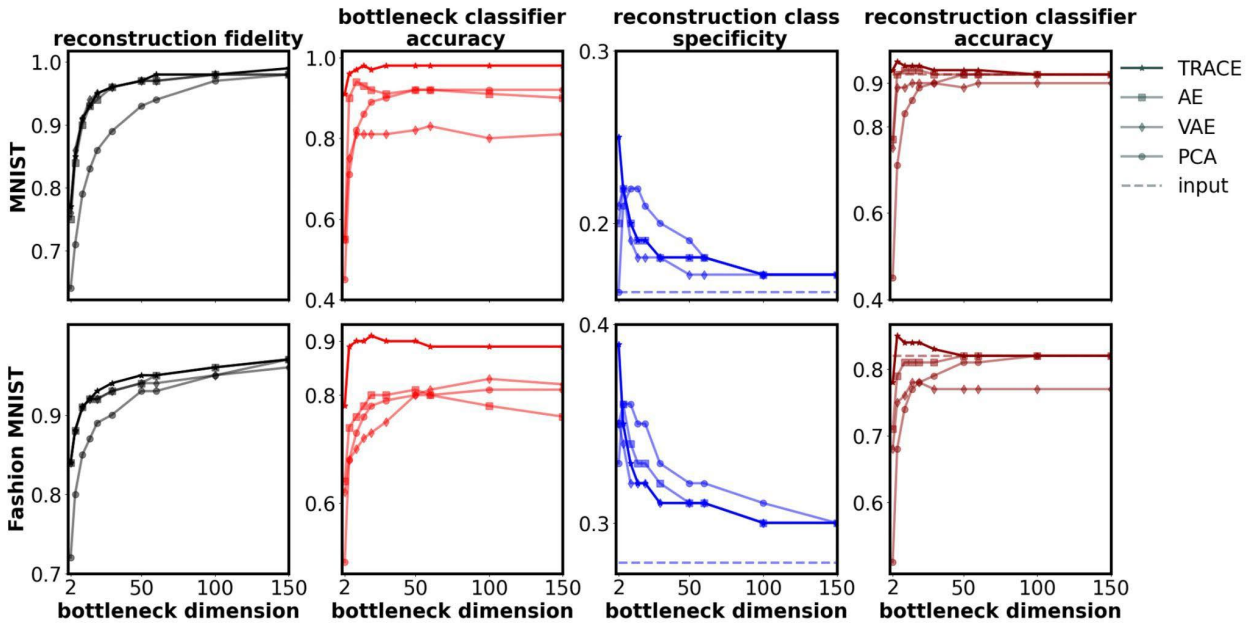

Figure 1.2. Quantitative comparison between TRACE and other models (AE, VAE, and PCA) on the four outcome metrics, for the two benchmark datasets (MNIST & Fashion MNIST) for bottleneck dimensionalities between 2 and 150. All metrics show superiority of TRACE over other models (except higher dimensionalities in reconstruction class specificity). TRACE is shown by the darker line while other models are shown by lighter lines. The black, red, blue, and dark red lines show the reconstruction fidelity, bottleneck classifier accuracy, reconstruction class specificity, and reconstruction classifier accuracy, respectively (see **Methods**). The dashed blue and dark red lines show the input class specificity and input classifier accuracy respectively. Outcome metrics for all bottleneck dimensionalities tested (2-784) are shown in **Figure S3**; locations of peaks for all four metrics are shown in **Table S1.** The chance levels of bottleneck and reconstruction classifier accuracy are both 10% (not shown in the plot).

We first examined *reconstruction fidelity* (black, **Figure 1.2**), i.e. the mean Pearson correlation of the inputs and corresponding reconstructions. High reconstruction fidelity

assures us that the discovered features in the bottleneck provide a reasonable representation of this high-dimensional information – i.e., that the autoencoder portion of the models can be successfully trained. Notably, TRACE's reconstruction fidelity performed in a similar fashion despite the fact that the contribution of the reconstruction part of the loss function (mean square error; MSE) for TRACE was smaller than for AE and VAE (i.e., the objective function in TRACE is the sum of reconstruction loss (L_R) and classification loss functions (L_{CE}); see **Methods Section 4.4.1**).

Next, we examined *bottleneck classifier accuracy* (bright red, **Figure 1.2**), i.e. the accuracy of a separate classifier trained with bottleneck features as input after the training of all models. Bottleneck classifier accuracy was much higher for TRACE than for other models even at very low bottleneck dimensionalities. As bottleneck dimensionality grew, this metric asymptotically equalizes to at least ~10% better than all other models in the MNIST and Fashion MNIST datasets. Notably, though, in both datasets, at all bottleneck dimensionalities tested, TRACE bottleneck classifier accuracy is *always* higher than that for other models. Although by attaching a cross-entropy loss function to the bottleneck of the network one can expect to naturally discover features that increase the classification accuracy, crucially this achievement is gained without losing the ability to reconstruct the input in the decoder part of the network. In other words, the lower dimensional representations learned by TRACE are not only more suitable for classification purposes but also can be used just as effectively to reconstruct the input.

The third metric we examined was *reconstruction class specificity* (blue, **Figure 1.2**), i.e. the average within-class correlation of the reconstructed inputs minus the average between-class correlation. This metric quantifies the degree of separation between class clusters in reconstruction feature space as a measure of reconstruction representations' categorical 'purity'. Reconstruction class specificity peaks at bottleneck dimensionality d=2 for TRACE for both MNIST and Fashion MNIST. As with the other metrics, TRACE outperformed other models at optimal bottleneck dimensionality d=2.

Fourth, we examined *reconstruction classifier accuracy* (dark red, **Figure 1.2**), i.e. the accuracy of a separate logistic classifier trained to discriminate classes using reconstructed data. Reconstruction classifier accuracy quantifies the task-relevance of the information extracted through the compression process, and also provides a direct benchmark against which to compare to the noisiness of representations in the original input space (see below). Reconstruction classifier accuracy for both MNIST and Fashion MNIST peaked at bottleneck dimensionality d=5 for TRACE, and was consistently higher for TRACE over other models. Interestingly, that this metric peaks at higher bottleneck dimensionalities than reconstruction class specificity suggests that the performance of a classifier trained on these high-dimensional reconstructions may not meaningfully reflect the maximum compression that TRACE can achieve without loss of overall performance.

A final – and critical – test of TRACE would examine its ability to not only distill task-relevant information into low-dimensional representations but also 'push' such distilled insights back into the native space of the input. This would be especially important if one

wished to use TRACE to de-noise fMRI data to discover multivoxel patterns representing a target concept or category to be used with noninvasive intervention strategies such as decoded neurofeedback (DecNef) [39–42] (or to simply investigate those activity patterns in native space). Although iterative sparse logistic regression and support vector machine classification have been demonstrated as successful at identifying such patterns when trained on the native input data [39,43,44], we wanted to see whether TRACE would be able to denoise the data such that an even cleaner target pattern would become discoverable. Specifically, if TRACE is successful at actively removing task-irrelevant noise rather than simply passively averaging across it (as is done with a standard category-based logistic regression) or removing it through iterative sparsity approaches (iterative sparse logistic regression), then we should observe two patterns. First, reconstruction classifier accuracy should approach or exceed classification accuracy of an identical logistic regression classifier trained on the native inputs. Second, reconstruction class specificity should behave similarly, approaching and then exceeding input class specificity. This behavior makes reconstruction class specificity an ideal metric for defining the 'optimal bottleneck dimensionality' if one's goal is to optimally distill representations in native space.

To evaluate this behavior, we (a) trained an additional logistic regression classifier on each of the datasets to classify the native input, and (b) computed class specificity directly from the raw input data for all three datasets. We then compared the outcomes to the reconstruction classifier accuracy and reconstruction class specificity computed as a function of bottleneck dimensionality.

Results revealed that, for MNIST, reconstruction classifier accuracy (solid dark red, **Figure 1.2**) exceeded input classifier accuracy (dashed dark red line) immediately (at d=2) for TRACE but not until d=10 for AE; it never exceeded the input for other models. For Fashion MNIST, this occurred at d=5 for TRACE and only at much higher dimensionality – if at all – for the other models tested. These results show that TRACE provides not only superior compression but also superior denoising even in comparison to the direct inputs. TRACE's denoising capability can be particularly useful in DecNef [39–41,45–51] studies as it can minimize the task-irrelevant information of exemplars even in the anatomical and functional brain space.

Results for reconstruction class specificity followed a different pattern, but still favored TRACE: reconstruction class specificity (solid blue line, **Figure 1.2**) exceeded input class specificity (dashed blue line) at most bottleneck dimensionalities for all models, but was higher for TRACE at the optimal bottleneck dimensionality (d=2). These results show that TRACE can provide a powerful method for not only distilling low-dimensional representations, but also in pushing those cleaner representations back into the structure and dimensionality of the raw input space. That is, a structurally identical logistic classifier with the same number of parameters can exhibit better performance using the reconstructed inputs than using the original inputs themselves.

Note that conducting statistical tests of the results from **Figure 1.2** is not feasible since the results reported here come from the training of cross-validated models on the entire dataset at each dimensionality of the bottleneck.

2.1.1 Comprehensive comparison across metrics as a function of increasing data sparsity

We next sought to select a single bottleneck dimensionality for TRACE to explore its benefits over AE, VAE, and PCA under increasing data sparsity. For this purpose, we selected the maximal value of reconstruction class specificity because this metric provides the best balance between task-relevant information extraction and compression, both for analyzing low-dimensional representations and patterns in the original input dimensionality (e.g., for use with real-time DecNef [39–41,45–51]).

Reconstruction class specificity peaked at d=2 for both MNIST and Fashion MNIST, so we can first examine TRACE's superiority at this dimensionality when maximal data is available (n = 60,000 training samples for both datasets). Since the goal is to compress information as much as possible without losing information, we chose d=2 to conduct the rest of the analysis given TRACE peaks at a bottleneck dimensionality lower than AE or other models (i.e. d=2). Additionally, at TRACE's peak (d=2), TRACE shows superior performance compared to other models' performance at dimensionalities where those other models peak (e.g., TRACE's reconstruction class specificity is higher at TRACE's peak [d=2] than AE's reconstruction class specificity is at AE's peak [d=5])**.** Here, we see that TRACE's superior extraction of task-relevant information comes at no loss in reconstruction fidelity over AE (**Figure 1.2**; **Table S2**). Further explorations, described below, were therefore done at bottleneck dimensionality d=2.

To examine how TRACE fared versus the other models under increasing data sparsity, we trained each model after removing 10, 30, 50, 70, 90, 95, and 98 percent of the training data. Training examples at each level of sparsity for all models remained the same. We then used the conventional 10,000 held-out test set on the trained models and calculated all four metrics for all levels of data sparsity.

TRACE was much more robust to increasing data sparsity than other models (**Figure 1.3**). Specifically, TRACE's performance was much better even when only 2% of the data (1200 samples) remained available for training. (The test set remained fixed at the same 10,000 standard test set used for these datasets.) We note that the fMRI dataset we use below has a similar samples-to-input-dimensions ratio as the 98% truncated MNIST and Fashion MNIST datasets (~1.6 for MNIST and Fashion MNIST, and ~1.5 for this fMRI dataset). At this level of data reduction (i.e., 98% truncation) and bottleneck dimensionality d=2, we performed 50 jack-knife replications to select 2 percent of exemplars in MNIST and Fashion MNIST for training, and reported the mean values (calculated within the standard test set) of the 50 independent training sets for all metrics. As shown in **Figure 1.3**, TRACE continued to demonstrate superior performance even at the most extreme level of data truncation (i.e., 98% truncation). TRACE nearly uniformly swept other models across all performance metrics. To confirm TRACE's superior performance, we ran four one-way repeated measures ANOVAs at 98% truncation – one for each of the outcome metrics – with factor model (4 levels). We then followed each omnibus ANOVA with planned contrasts comparing TRACE to each other model in a pairwise fashion. This analysis revealed a main effect of model for all four outcome

metrics, and that TRACE was statistically superior to all other models in all 12 pairwise comparisons; see **Table S3** for all statistics).

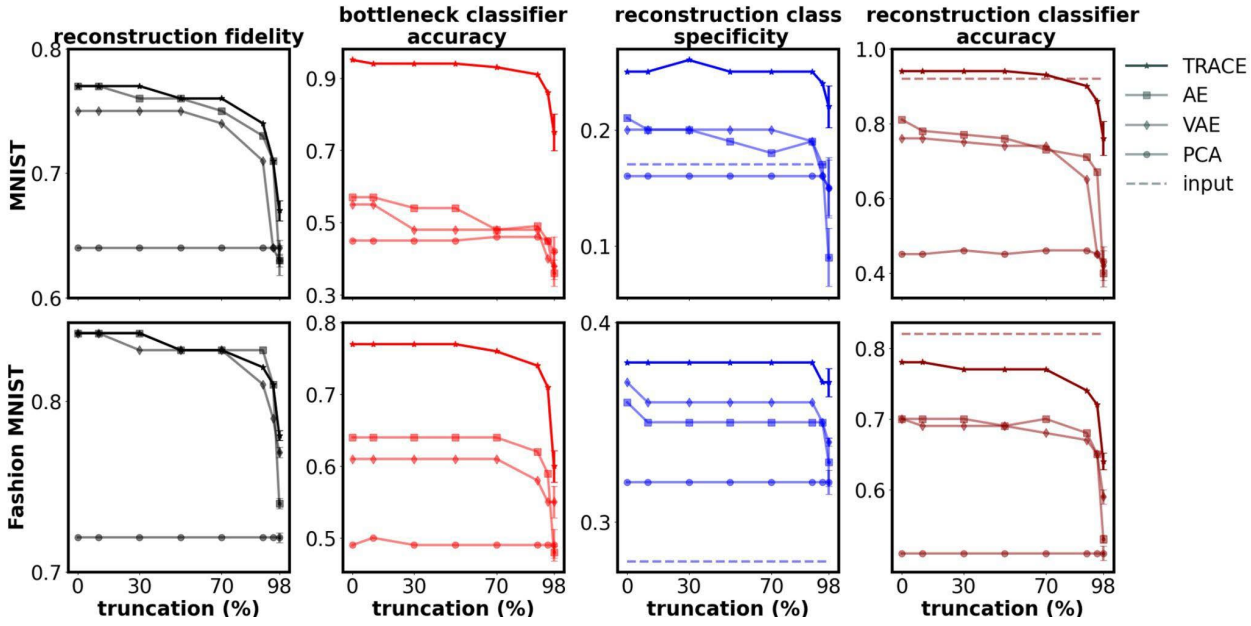

Figure 1.3. Performance of TRACE and other models as a function of sample size for the optimal bottleneck dimension of d=2. At 98% truncation level, we used 50 independent jack-knife resamplings to truncate 98 percent of exemplars and reported the means and standard deviations of the metrics (calculated on the standard test set) for MNIST and Fashion MNIST. Error bars show the standard deviation of results across the 50 jack-knife resamplings at 98% data truncation. Small variations in the metrics are likely due to random initialization of weights and use of GPUs in fitting the models.

At maximal data reduction (98% truncation) and bottleneck dimensionality d=2, we then performed additional explorations of both bottleneck representations and reconstructions. First, we visualized bottleneck representations by plotting the activities of the two

bottleneck features against each other for each of the 10 classes in each dataset for TRACE versus the other models (**Figure 1.4**). The results are striking: TRACE showed superior task-relevant representations especially for MNIST, i.e. a clear qualitative advantage in clustering performance showing distinct clusters for different classes in stark contrast to the other models' class clusters, which are heavily overlapping. Although this difference in clustering ability was less apparent for Fashion MNIST, TRACE's clusters do appear visually more tightly bound.

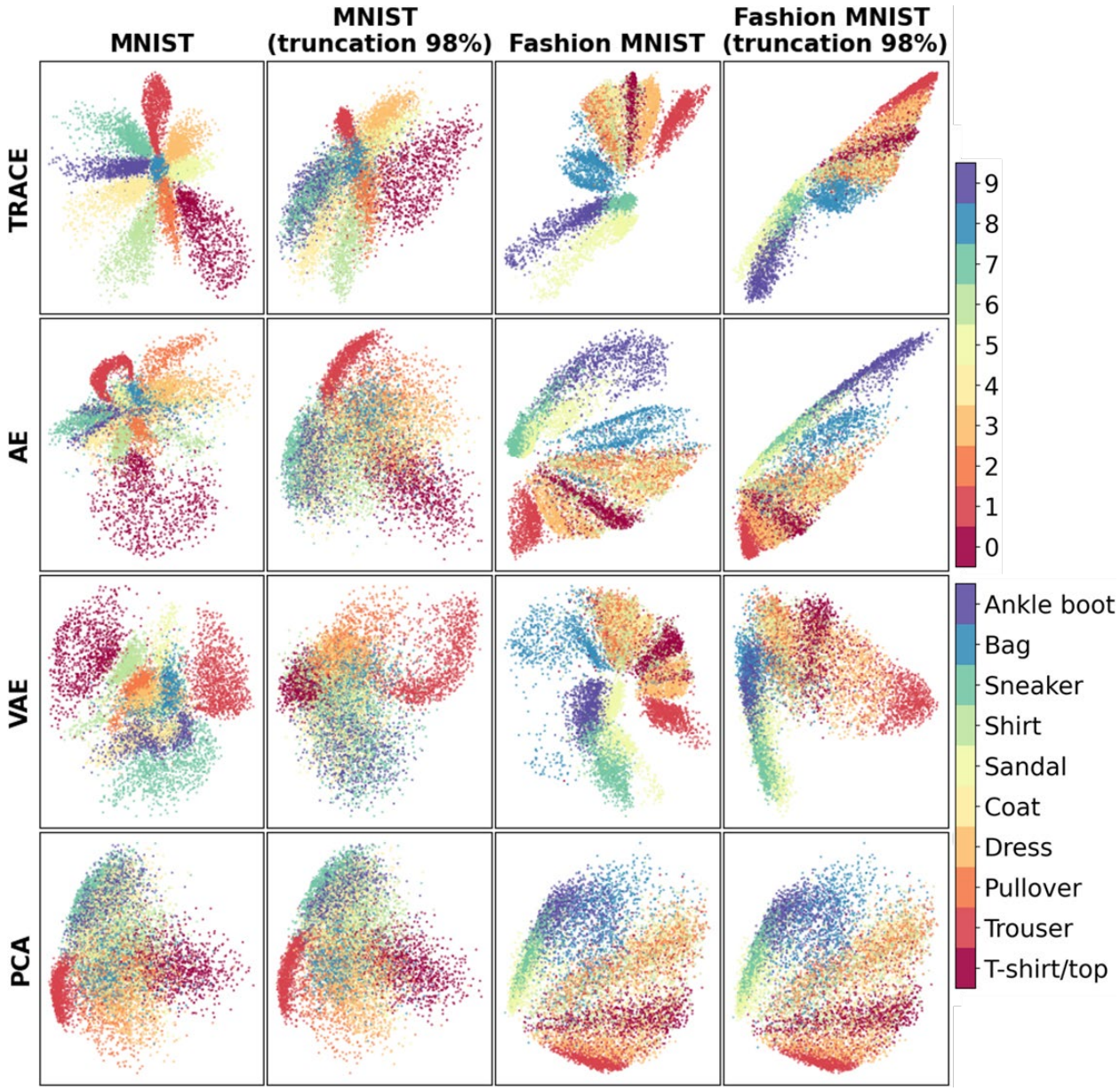

Figure 1.4. Visualization of bottleneck features for MNIST and Fashion MNIST datasets using TRACE, AE, VAE, and PCA. When trained on the full dataset, TRACE shows clear superiority in creating distinctive clusters in the bottleneck for different classes for MNIST dataset in comparison to other models The distinction is less clear but still apparent in the Fashion MNIST dataset. This pattern persists even at the 98% truncation level (trained on only 2% of the data), again showing the robustness of TRACE.

We next turned to examining the reconstructions (still at bottleneck d=2). We first examined the MNIST reconstructions for several different exemplars of the same categories (e.g., several different “3” and “6” exemplars). TRACE’s superiority is clear to the naked eye: the reconstructions of particular “3” and “6” exemplars from TRACE are much more “three-like” and “six-like” than reconstructions from other models especially at the 98% truncation level (**Figure 1.5**). (Recall that this qualitative superiority does not come at any quantitative cost to the reconstructions). Similar findings held for Fashion MNIST (e.g., sandal and shirt), although the visual result is less striking. These patterns held even when only 2% of the data was available for training.

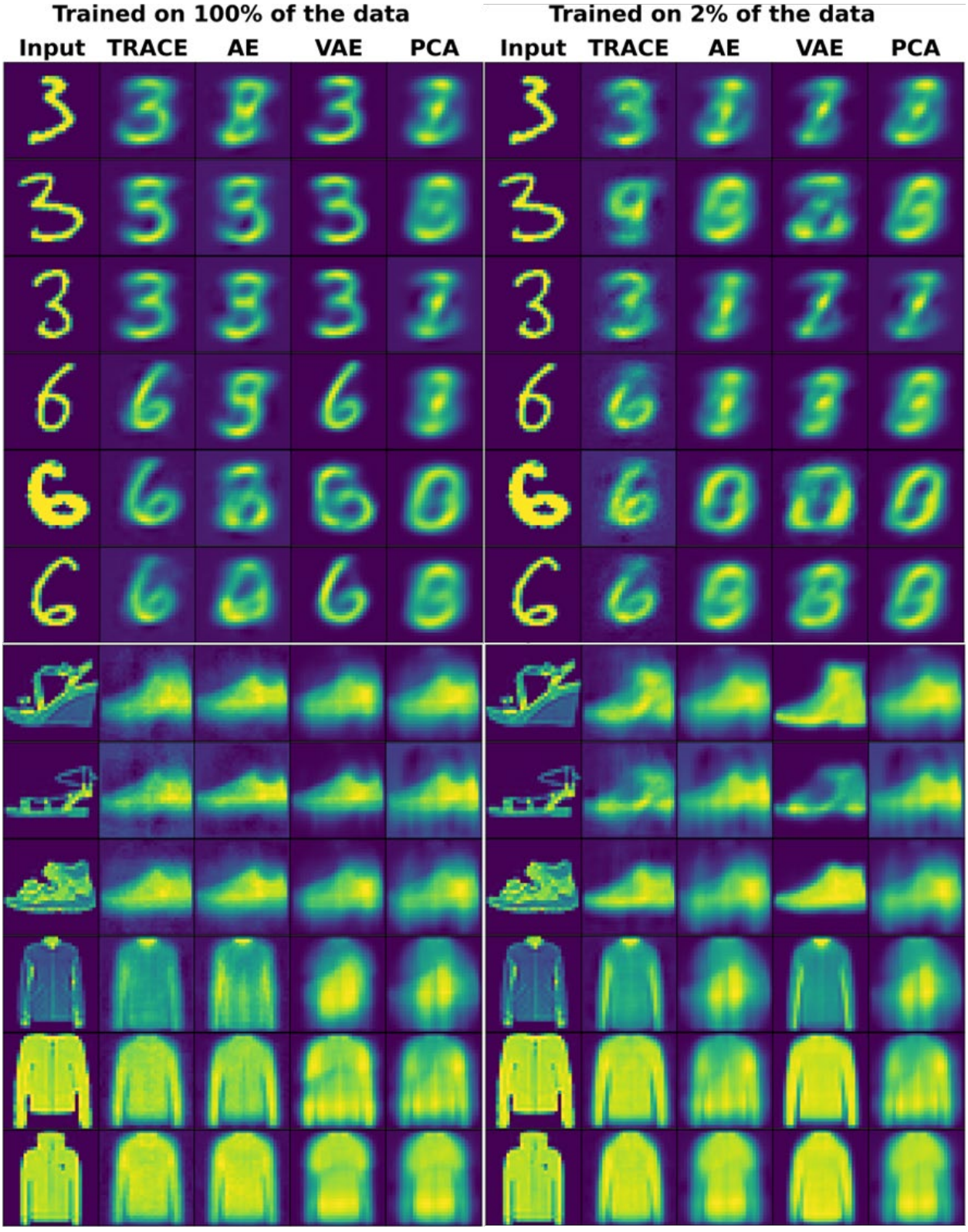

Figure 1.5. Visualization of reconstructions for MNIST and Fashion MNIST datasets using TRACE, AE, VAE, and PCA. The reconstruction of three representative instances

of numbers “three” and “six” in MNIST dataset and three instances of classes “sandal” and “shirt” in the fashion MNIST dataset when there are two features in the bottleneck shows the same pattern. TRACE shows a more clear and *canonical* reconstruction of the inputs across several exemplars from the same category.

We next wanted to quantitatively investigate the distributions of within-class versus between-class clusters, both in the bottleneck and the reconstructions. This approach will facilitate evaluation of the fMRI dataset since visual inspection in fMRI data is not possible in the same sense as for MNIST and Fashion MNIST given that optimal bottleneck dimensionality is larger than 2 (**see Results Section 2.2**). We computed the effect size (Cohen’s d) separating clusters in both the bottleneck and reconstructions using pairwise within- versus between-class Euclidean distances. Whether trained on all of the data or 98% truncated, Cohen’s d was always larger for TRACE than for other models (**Table 1.1**).

		MNIST	**Fashion MNIST**	**MNIST (98% truncation)**	**Fashion MNIST (98% truncation)**
Bottleneck	**TRACE**	1.63 ± 0.21	1.65 ± 0.5	1.36 ± 0.41	1.45 ± 0.35
	AE	1.1 ± 0.39	1.5 ± 0.43	1.01 ± 0.41	1.26 ± 0.44
	VAE	1.32 ± 0.55	1.61 ± 0.51	0.8 ± 0.31	1.48 ± 0.44
	PCA	1.06 ± 0.54	1.25 ± 0.59	1.06 ± 0.54	1.25 ± 0.59

Reconstruction	**TRACE**	1.58 ± 0.21	1.6 ± 0.69	1.51 ± 0.63	1.54 ± 0.68
	AE	1.2 ± 0.29	1.48 ± 0.62	1.02 ± 0.44	1.41 ± 0.66
	VAE	1.27 ± 0.28	1.42 ± 0.61	0.68 ± 0.38	1.37 ± 0.65
	PCA	1.06 ± 0.54	1.25 ± 0.59	1.06 ± 0.54	1.25 ± 0.59

Table 1.1 Cohen's d measures of effect size comparing within-class versus between-class Euclidean distances in the bottleneck and reconstructions for TRACE, AE, VAE, and PCA.

2.2 TRACE's performance on a real fMRI dataset

Given TRACE's apparent superiority over AE, VAE, and PCA even under extreme data sparsity, we next sought to evaluate TRACE using a real-world fMRI dataset, since ultimately our goal is to learn about neural representations. Thus, we used the same metrics as we used to evaluate TRACE on MNIST and Fashion MNIST on an fMRI dataset consisting of 59 individuals who each viewed 3600 exemplars of 40 classes of animals and objects (90 exemplars per class) while BOLD signal from ventral temporal cortex (VTC) was obtained. The number of voxels in VTC for each individual was different; however, the average of voxels for the 59 subjects was 2382 ± 303.

Excitingly, the fMRI dataset showed the same patterns in our four quantitative metrics as the MNIST and Fashion MNIST datasets almost across the board. First, reconstruction fidelity was actually slightly higher for AE over TRACE and VAE at higher dimensions, although this is likely due to the fact that reconstructing the input is the only objective of

the AE network; however, note that the numerical difference between TRACE and AE is very small, and that both are outperforming VAE. PCA also showed higher reconstruction fidelity than all other models starting around d=500, which is also expected since as the number of principal components increases, the PCA model can explain the variance of the input data almost perfectly.

Reconstruction classifier accuracy asymptoted at bottleneck dimensionality around d=250 for all models, but again TRACE showed higher reconstruction classifier accuracy than AE, VAE, and PCA at all bottleneck dimensionalities tested. TRACE also showed higher bottleneck classifier accuracy at all bottleneck dimensionalities in comparison to other models.

TRACE outperformed other models in reconstruction class specificity as well, showing that even in the native space of the input – i.e., voxel patterns of activity in ventral temporal cortex – TRACE not only successfully distills lower-dimensional representations of task-relevant data, but also faithfully projects them back into original, high-dimensional voxel space. Reconstruction class specificity peaked at bottleneck dimensionality d=30, and then fell again. The same was not true for other models, for which reconstruction class specificity rose but then asymptoted. Crucially, though, reconstruction class specificity was also always higher for TRACE than for other models, much exceeding input class specificity (**Figure 1.6**, solid and dashed blue lines, respectively). This capacity to distill a task-relevant, low-dimensional representation and put it back in brain space could potentially have great value for studies in which such multivoxel patterns are the target of

DecNef [39–42] or other investigations which require anatomically-related representations. We discuss this possibility in greater detail in the **Discussion**, below.

Finally, TRACE's reconstruction classifier accuracy even surpassed the input classifier accuracy for bottleneck dimensionalities higher than d=60 (dashed dark red line, **Figure 1.6**) which again suggests that the reconstructed version in the original input space contains more task-relevant information.

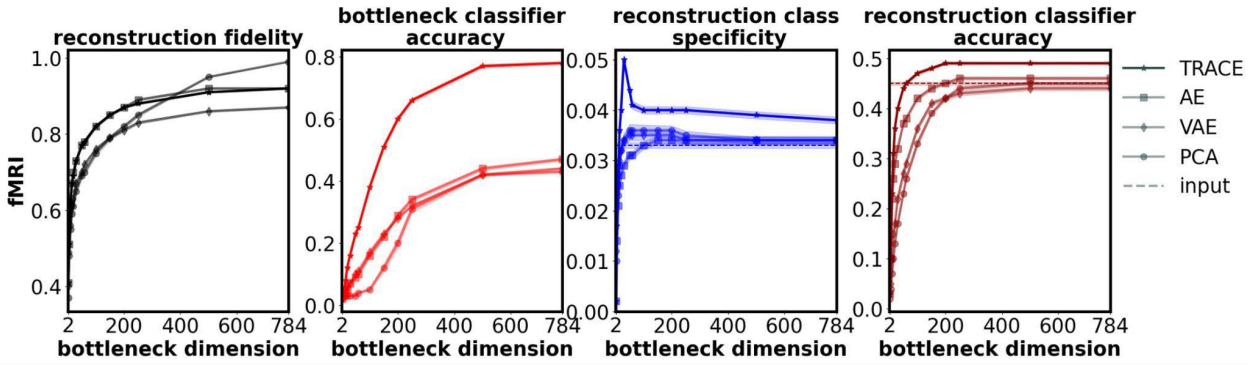

Figure 1.6. Comparison between quantitative metrics for TRACE and other models for fMRI dataset (n=59). TRACE shows superior performance in three out of four metrics (excluding reconstruction fidelity and only for d>250).

2.2.1 Exploration at optimal bottleneck dimensionality for fMRI data

As mentioned above, the maximal value for reconstruction class specificity was found at d=2 for the MNIST and Fashion MNIST datasets. For the fMRI dataset, we found that reconstruction class specificity peaked at d=30, so we proceeded with a parallel analysis to that done above at this dimensionality.

Crucially, at d=30, TRACE's performance on the fMRI dataset mimicked its exemplary performance on the MNIST and Fashion MNIST datasets with the exception of reconstruction fidelity, which was only slightly smaller for TRACE than for AE at dimensionalities of d>250 (and PCA at higher dimensions [i.e., d>500]) (**Figure 1.6**). To quantify this superiority, we performed a one-way repeated measures ANOVA for each outcome metric with factor model (4 levels), followed by planned pairwise contrasts comparing TRACE to every other model. Results revealed significant main effects of model for all four outcome metrics, and that TRACE outperformed the other models in 11 of these planned comparisons (with the exception of reconstruction fidelity between TRACE and AE; see **Table S4** for statistics).

Ultimately, as our goal is to learn about representations in human VTC, we also might want to visualize clusters for the 40 classes of the fMRI dataset. However, unlike for MNIST and Fashion MNIST where optimal bottleneck dimensionality was d=2, for the fMRI dataset we found the optimal bottleneck dimensionality at d=30. Therefore, we cannot easily visualize the class clusters in a scatterplot, and performing further dimensionality reduction for the sake of visualization would be inappropriate since assumptions of whichever dimensionality reduction technique we chose would impact the visualizations. Instead, we can use the same Cohen's d approach, described above, to characterize the tightness of the class clusters even in higher dimensionalities. The average effect size separating within- and between-class Euclidean distances across all 59 subjects was 0.38 (±0.09) for TRACE, 0.12 (±0.03) for AE, 0.11 (±0.02) for VAE, and 0.08 (±0.02) for PCA again showing TRACE's superiority.

As a final evaluation of TRACE's ability to filter out task-irrelevant information, we calculated the within- versus between-class Euclidean distance Cohen's d in the reconstructions. Pushing the distilled representations back into input space is particularly exciting for the use of TRACE with fMRI data if one wishes to discover a particular target pattern for further anatomical analysis, or for use with real-time neuroimaging (e.g., DecNef). However, visually examining fMRI reconstructions would not provide particularly useful information about the 'cleanliness' of the reconstruction, as the patterns are not visually meaningful to begin with, so we must again rely on a quantitative comparison. The average Cohen's d here again showed TRACE's superiority, with mean Cohen's d of 0.14 (±0.02) across subjects for TRACE, 0.09 (±0.02) for AE, 0.08 (±0.02) for VAE, and 0.08 (±0.02) for PCA. In other words, TRACE was able to reduce task-irrelevant information and thus extract a 'cleaner' representation, even in the reconstructions.

3. Discussion

3.1 Summary of findings

Most dimensionality-reduction approaches do not have a specific mechanism to ensure that the lower dimensional representations they reveal are particularly relevant to the question an experimenter wishes to answer. Further, many state-of-the-art deep learning models are of limited utility for discovering and characterizing meaningful representations in input-dimension-rich but exemplar-poor datasets, as they tend to overfit [30–32]. Together, these facts make discovering neural representations in within-subject fMRI datasets –

which also often contain a high degree of noise and task-irrelevant information – extremely challenging [17–19]. Further, to address these issues we proposed the Autoencoder with Classifier Enhancement (TRACE) model: a simple autoencoder with a classifier attached to the bottleneck. The classifier forces the model to learn not just lower dimensional representations of the data, but those that are also task-relevant. To quantify TRACE's superiority over a standard autoencoder (AE), a variational autoencoder (VAE), and principal components analysis (PCA), we used four metrics (see **Methods Section 4.4**): 1. reconstruction fidelity; 2. bottleneck classifier accuracy; 3. reconstruction class specificity; and 4. reconstruction classifier accuracy.

TRACE outperformed all other models in all metrics, with the exception of reconstruction fidelity (sometimes). Moreover, at the 'optimal' bottleneck dimensionality, TRACE's superior capacity for extracting task-relevant information is evident in both the bottleneck and reconstruction, and TRACE's reconstructions can even outperform the inputs on a measure of task-relevant behavior (reconstruction class specificity). TRACE's advantage over other models appears due to TRACE's capacity to minimize task-irrelevant, idiosyncratic information unique to a particular sample. This is evident in the one occasional exception to TRACE's sweeping superiority: reconstruction fidelity for the fMRI dataset. However, this seeming underperformance – especially in the fMRI dataset – is actually a strength: AE tried "too hard" to encode idiosyncratic details of a particular exemplar in the bottleneck, when some of these details are merely noise for the task that the observer is performing. Thus, precise reconstruction of noisy data may not be suitable.

Critically, all of these behaviors were maintained by TRACE even under extreme data truncation for the MNIST and Fashion MNIST datasets, and carried over into a real-world fMRI dataset. These results suggest that TRACE can extract lower-dimensional representations of data for both reconstruction and classification purposes and can do so even when there is a highly undesirable balance of input-dimensions versus samples. We speculate that the better performance under the scarcity of sample size is due to adding additional label information to the bottleneck which acts as an auxiliary function to help the network to learn the general pattern in the face of scarcity of sample size. Since this scarcity of sample size is typical in fMRI data, the superior performance of TRACE suggests the strong promise of this approach for both fMRI datasets and for other biological-scale data with many more input-dimensions than samples.

3.2 Relation to previous work

TRACE is not the only model which can accomplish dimensionality reduction, but one of many techniques. So is TRACE really necessary? Why would principal components analysis (PCA) [25–27] not suffice? PCA focuses on creating new features that can best explain the variance in data – including the noise and task-irrelevant information, which we know to be problematic especially in fMRI data [18,52–56] – and thus lacks explicit mechanisms to ensure the discovered lower representations contain task-relevant information. Additionally, we also note that PCA-based methods are not assumption-free (that is, they are parametric); these assumptions about the functional form of the dimensionality reduction limit the discovered features to adhering to those assumptions.

Our approach builds on previous successes with classifier-enhanced autoencoders [33–36] to extract task-relevant representations in non-biological datasets such as linguistic datasets, standard computer vision object datasets, and fault diagnosis applications. However, TRACE goes beyond these previous successes by explicitly demonstrating with otherwise matched architecture (TRACE vs AE) that the simple addition of a classifier can improve extraction of task-relevant latent representations *even under extreme data paucity*. This demonstration is especially important for the types of data used in cognitive neuroscience, which are often sample-poor. We also demonstrate that TRACE can improve reconstruction classifier accuracy and reconstruction class specificity such that it exceeds even input-level for these metrics, which could be a boon for real-time decoded neurofeedback (DecNef [39,40,49]). We discuss these implications in more detail in Implications and Future Directions, below.

Other techniques have been developed including nonparametric techniques beyond the fully-connected AE and VAE[57,58] used here [59,60]: adversarial autoencoders [61], generative adversarial networks (GANs) [62], deep convolutional GANs (DCGANs) [63], and so on. While comprehensive exploration of these is beyond the scope of this manuscript, we note that many of these models do still suffer from the fact that the discovered lower dimensional representations are not explicitly crafted to be task-relevant [64]. In fact, we can demonstrate that an implementation of a GAN modified to allow selection of specific categories of reconstruction (a conditional GAN, or cGAN [65]), fails quite miserably when trained only 2% of the MNIST or Fashion MNIST datasets (see **Supplementary Material S3** and **Figure S3**). These considerations led to the development of InfoGAN [24], an

unsupervised learning technique which modifies a generative adversarial network (GAN) in order to learn interpretable, low-dimensional representations. InfoGAN accomplishes this task by maximizing mutual information between noise in the GAN network and observations. Yet despite the tremendous success of InfoGAN [24], it is highly disadvantaged for the limited (sample-poor) data type targeted here. Specifically, InfoGAN's success has been demonstrated only on large-scale training datasets consisting of tens of thousands of training images. Further, exploring and characterizing latent spaces in GANs in general is highly nontrivial [7,66]; for these reasons, GANs generally do not accomplish the goal targeted by the TRACE network.

Attempts to mitigate the curse of dimensionality in fMRI datasets by pooling data across subjects to create larger training sets have of course been established to try to mitigate this significant challenge, including the shared response model [16], hyperalignment [67–69], and more recently decoder + autoencoder approaches [70]. However, while these can pool fMRI data to create more training exemplars, they do not explicitly seek subject-specific response patterns and instead presuppose that all subjects share a common response pattern.

In sum, although we do not benchmark TRACE against InfoGAN, hyperaligned data, or the expansive space of model variants, we argue that TRACE's utility is not only in its ability to distill task-relevant, low-dimensional representations, but also to do so in exemplar-limited, biological-scale datasets such as those collected in human neuroimaging experiments within a single subject.

3.3 Limitations

One limitation of the present approach is that we (deliberately) made TRACE and other models extremely simple (as in, few layers), which could have limited their performance. We did not investigate whether TRACE-like architecture (addition of a classifier on the bottleneck layer) would similarly improve performance for more complex networks, or whether multi-layer perceptrons or convolutional neural network (CNN) classifiers would surpass the simple logistic regression classifiers used here. We also could have opted to make the models deeper, with many hidden layers, which might have resulted in benefits in classification or reconstruction. However, we reiterate that we selected a simple architecture to be able to best evaluate TRACE's advantages over a "plain vanilla" fully-connected autoencoder, as more complex architectures could obscure TRACE's advantages. Future work may wish to explore other possible TRACE-like architectures.

It is also worth mentioning that for the sake of consistency we kept all hyperparameters for all networks and datasets the same. However, during training TRACE on a new dataset, it is always possible to tune the hyperparameters (learning rate, batchsize, regularization, etc) in order to achieve better performance (e.g. better bottleneck classification accuracy). Future studies may also more comprehensively explore the impact of specific hyperparameter tuning choices on TRACE's behavior.

3.4 Implications & future directions

Our findings have potentially exciting implications for the discovery of both low-dimensional representations and representations in the original (and anatomically- and/or functionally-relevant, in the case of fMRI) input space. For example, if a study's goal is to

induce canonical target patterns of neural activity for a particular object category with real-time decoded neurofeedback (DecNef [39,40,49]), one might wish to instead 'de-noise' the data by maximizing reconstruction classifier accuracy instead of reconstruction class specificity. In the fMRI dataset presented here, reconstruction classifier accuracy peaked at about d=200. It is possible that in other fMRI datasets, reconstruction classifier accuracy might peak at a non-maximal bottleneck dimensionality, in which case it could be used to select the best dimensionality for the task at hand. Alternatively, one could choose to select optimal bottleneck dimensionality based on when reconstruction class specificity or classifier accuracy exceeds the analogous metric calculated directly from the raw input data. Here we showed that TRACE either exceeds these benchmarks sooner than other models, or does so even when other models do not. Thus, the process for selecting the best bottleneck dimensionality can flexibly adapt to an experimenter's goals, and future research seeking to use TRACE to extract neural patterns for use with DecNef should explore how different bottleneck dimensionalities impact the success of the neurofeedback process.

Regardless of the method one uses to select bottleneck dimensionality, it seems likely that TRACE can remove task-irrelevant information in a way that is useful for DecNef. To demonstrate this possibility, we did one final exploratory test. Recall that the fMRI dataset used in this study is in part overlapping with the dataset used by Taschereau-Dumouchel and colleagues [49], and as such we can directly compare their binary ("cat" versus "everything that is not a cat") decoding accuracy with the decoding accuracy we achieved on TRACE's reconstructions. To translate the reconstruction classifier accuracy we

achieved to a binary scale, we counted a prediction to be correct if the correct class was in the top 20 (out of 40) of predicted classes from our one-versus-all classifier (with chance classification accuracy at 2.5%). Taschereau-Dumouchel and colleagues [49] observed binary logistic regression classification accuracies of 71.7% on average within-subject (~1 hour of fMRI data per person). (Relying on hyperalignment[67] to pool their 30 subjects and subsequently train such classifiers, they observed mean 82.4% using a 30-subject concatenated dataset). When we trained logistic regression classifiers on each individual subject (i.e., no hyperalignment) – some of whom are actually the original subjects in that former study – and translated the classification accuracies as described to be on the same scale as binary classification, we achieved the equivalent of 94.4% binary accuracy at bottleneck dimensionality d=30 (where reconstruction class specificity was maximized). Thus, TRACE facilitates distillation of class-specific representations in native space that are superior to the original representations themselves for this purpose.

Another interesting future possibility would be to investigate the extent to which TRACE excels over other methods as a function of neuroanatomical area – for the purposes of DecNef or simply to investigate neural representations themselves. Here, we focused on object representations in high level visual cortex (VTC), but in theory one could ask how early in the visual processing pipeline we might find evidence that task relevance plays a meaningful role. In the fMRI dataset used here, the task was for subjects to identify the object category of the image, and as a result the images were not standardized across lower level visual features such that object category did indeed covary with lower level visual properties such as color or spatial frequency (e.g. the background color of the

‘dolphin’ images is predominantly blue, whereas this is not the case for the ‘key’ images). Future studies may wish to use standardized images to investigate to what extent TRACE may assist in extraction of task-relevant representations versus low-level visual properties, depending on task and brain area; due to the limitations of the dataset used here for this first proof of concept, we leave these questions to future investigations.

Given TRACE’s success here, we hope that its capacity to discover task-relevant information *despite* undesirable ratios of samples to input-dimensions can help discover truths about other biological processes. Future studies should apply TRACE to other biological-scale datasets, with the goal of discovering representations relevant to those researchers and domains.

As discussed earlier, discovering lower dimensional representations that are in fact more task relevant can greatly help researchers to interrogate these lower dimensions. It is important to acknowledge that utilizing deep learning models such as TRACE comes with the caveat of a more difficult interpretation. Thus, full exploration of the latent, low-dimensional representations extracted by TRACE remains a subject for further investigations using available explainable artificial intelligence methods [71].

4. Methods

4.1 Methods overview

We proposed the “Task-Relevant Autoencoder via Classifier Enhancement” (TRACE) model and directly compared its behavior to that of a standard autoencoder (AE), a variational autoencoder (VAE), and principal component analysis (PCA) with equivalent internal architecture. Additional information about the methodology can be found in **Supplementary Material S1.1**.

4.2 Datasets

We employed MNIST [37] and Fashion MNIST [38] to benchmark the TRACE against other models. Additionally, we used a previously collected fMRI dataset, partially reported by Taschereau-Dumouchel and colleagues [49], in order to demonstrate the TRACE’s efficacy in small-scale fMRI datasets. The fMRI data used in this study was obtained from 59 healthy individuals who viewed 3600 images from 40 different categories of objects (30 animals and 10 man-made objects) while the whole-brain BOLD responses were acquired. Refer to **Supplementary Material S1.2** for more information.

4.3 Implementation of Task-Relevant Autoencoder via Classifier Enhancement (TRACE)

The Task-Relevant Autoencoder via Classifier Enhancement (TRACE) model is identical to a standard autoencoder with two hidden layers (one in the encoding section and one in the decoding section). The key distinction lies in the attachment of a logistic regression classifier to the bottleneck (**Figure 1.1**). For the hidden layers, we used the hyperbolic tangent as the activation function in order to discover more complex nonlinear patterns in the data, as this function has been reported previously to be more sensitive in capturing

detailed and local information to represent the data with lower dimensions [72]. The activation function for the "decoder branch" of the network was the softmax function (also known as Boltzman distribution) which outputs a probability distribution for each class (e.g., Classes of 10 digits for MNIST). The objective function of TRACE consists of two components, defined as follow:

The first component of the objective function, **Equation 1.1,** adopts the mean square error (MSE) as the criterion to reconstruct the input:

$$L_R = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{X}_{ij} - X_{ij})^2 \qquad (1.1)$$

Where X is the input with m samples and n input-dimensions, and $\hat{X}$ is the reconstruction of the input. The second component of the object function (attached to the bottleneck of the network), **Equation 1.2,** was chosen as the cross-entropy loss function to find lower-dimensional representations that are optimized to be task-relevant.

$$L_{CE} = \frac{-1}{m} \sum_{i=1}^{m} \sum_{c=1}^{k} y_{ci} \, log(\hat{y}_{ci}) \qquad (1.2)$$

where k denotes the number of the classes, y is the label of observation, and $\hat{y}$ is the predicted label.

In the TRACE network, the final objective function is the summation of reconstruction loss and the categorical cross-entropy loss function (**Equations 1.2** & **2.2**), i.e.:

$$L_{TRACE} = L_R + L_{CE}$$

$$= \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{X}_{ij} - X_{ij})^2 - \frac{\alpha}{m} \sum_{i=1}^{m} \sum_{c=1}^{k} y_{ci} \, log(\hat{y}_{ci}) \quad (1.3)$$

where α, sets the weight for the classifier part of the loss function in order to control for its participation in updating the parameters.

For more detailed information about the implementation of TRACE and other models, see **Supplementary Material S1.3**.

4.4 Outcome metrics

In order to explore what is the best low dimensional feature space that explains within class characteristics while preserving the ability of the network to reconstruct the input, we evaluated four metrics as a function of bottleneck dimensionality in all models and for all three datasets: (1) *reconstruction fidelity,* (2) *reconstruction classifier accuracy*, (3) *bottleneck classifier accuracy,* and (4) *reconstruction class specificity*, as described below and shown in cartoons in **Figure 1.7**. A more detailed explanation of all outcome metrics is provided in the **Supplementary Material S1.6.**

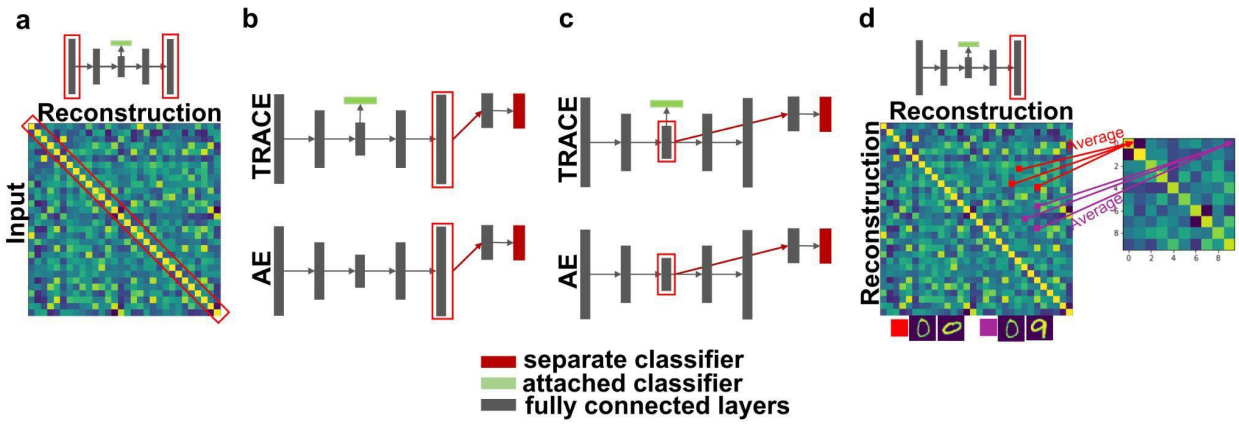

Figure 1.7. Graphical representation of the four quantitative outcome metrics. **(a)** Reconstruction fidelity, **(b)** reconstruction classifier accuracy, **(c)** bottleneck classifier accuracy, and **(d)** reconstruction class specificity. Small cartoons of the TRACE architecture use red rectangle overlays to indicate which sections of the model architecture are being utilized for each outcome metric. In **(b)** and **(c)**, red-filled boxes indicate separate classifiers, green-filled boxes indicate attached classifiers, and gray-filled boxes indicate fully-connected encoder and decoder layers.

4.4.1 Reconstruction fidelity

We quantified how well a model could reconstruct the input information with the average of all Pearson correlation coefficients between each input trial of the test set and the corresponding reconstruction of that sample (**Figure 1.7a**). We computed this correlation coefficient at all bottleneck dimensionalities tested. In the Results, below, this metric is referred to as *reconstruction fidelity* or $fidelity_R$.

$$fidelity_R = E(\rho_R) \tag{1.6}$$

Where ρ_R is the correlation between each input exemplar and its reconstruction, and E denotes the expected value.

4.4.2 Reconstruction classifier accuracy

To quantify how well the reconstructed input represents a certain class, we used a separate logistic classifier (**Equation 1.7**) and trained it using reconstructed inputs for all dimensions in the bottleneck (**Figure 1.7b**). Using the same train/test folds as for training all models, we trained the data for 30 epochs for MNIST and Fashion MNIST and 300 epochs for fMRI.

$$L_{RCA} = \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{k} y_{ci}\, log(\hat{y}_{ci}) + \lambda \sum_{r=1}^{p} w_r^2 \quad (1.7)$$

Where L_{RCA} is the cross-entropy loss for reconstructed input, and λ is the regularization parameter and w and p are the weight matrices and the number of parameters of the classifier respectively.

4.4.3 Bottleneck classifier accuracy

We quantified the task-relevance of the features in the bottleneck via the accuracy of the logistic regression classifier with such bottleneck node activity as inputs (**Figure 1.7c**). For all models, this classifier is trained separately, *after* the training of models is finished. We first extracted the bottleneck features after the training of all networks was complete, and then trained a completely separate logistic regression classifier to classify the category of the input image as it was now represented in the low-dimensional bottleneck of each model.

$$L_{BCA} = \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{k} y_{ci}\, log(\hat{y}_{ci}) + \lambda \sum_{b=1}^{q} w_b^2 \quad (1.8)$$

where w and q are the weight matrices and the number of parameters of the classifier respectively. The hyperparameter λ was set to 0.007 which was manually tuned to maximize the classification accuracy.

4.4.4 Reconstruction class specificity

Another measure of the task-relevancy of the reconstructed information is the degree of similarity of representations within a class versus between classes which we defined as the average of the diagonal (within class) of this similarity matrix minus the average of the off-diagonal (between class) of this matrix (**Figure 1.7d**), i.e.

$$RCS = E(\rho_{R,within}) - E(\rho_{R,between}) \quad (1.9)$$

where RCS is the class specificity in the reconstruction of the input, and $\rho_{R,within}$ and $\rho_{R,between}$ are the Pearson correlation matrices between trials within-class and between-class respectively. See **Supplementary Material S1.6.4** for a more detailed description.

4.4.5 Benchmarks against original inputs

To quantify the reduction in noise and the success of task-relevant feature extraction, we benchmark the reconstructions from all models in two ways.

First, we examined the classification accuracy of a simple logistic regression classifier applied to the input data in comparison to the accuracy of an identical classifier applied to the reconstructions (**Methods Section 4.4.2**). That is, if a representation has been

successfully de-noised through the compression (and task-relevant feature extraction, in the case of TRACE), then the reduction in task-irrelevant noise should be apparent in the superior classification accuracy of a logistic regression classifier. Thus, we trained logistic regression classifiers on the input space as well as the reconstruction (**Methods Section 4.4.2**) at each bottleneck dimensionality, and reported the accuracy (**Equation 1.7**).

Second, a final test of the ability of models to extract task-relevant representations can be quantified via comparing the reconstruction class specificity (**Methods Section 4.4.4**) against input class specificity, calculated equivalently to reconstruction class specificity (**Equation 1.9**).

Chapter 2. Domain adaptation in small-scale and heterogeneous biological datasets (under review)

Seyedmehdi Orouji[1], Martin C. Liu[2,3], Tal Korem[3,4,5,*,†], Megan A. K. Peters[1,5,6,*,†]

Author affiliations
1 Department of Cognitive Sciences, University of California Irvine, Irvine, CA
2 Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY
3 Program for Mathematical Genomics, Department of Systems Biology, Columbia University Irving Medical Center, New York, NY
4 Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY
5 CIFAR Azrieli Global Scholars Program, CIFAR, Toronto, Canada
6 CIFAR Fellow, Program in Brain, Mind, & Consciousness, CIFAR, Toronto, Canada

(* These authors contributed equally to this work.)
(† Corresponding authors. Email: tal.korem@columbia.edu, megan.peters@uci.edu)

Abstract

Machine learning techniques are steadily becoming more important in modern biology, and are used to build predictive models, discover patterns, and investigate biological problems. However, models trained on one dataset are often not generalizable to other datasets from different cohorts or laboratories, due to differences in the statistical properties of these datasets. These could stem from technical differences, such as the measurement technique used, or from relevant biological differences between the populations studied. Domain adaptation, a type of transfer learning, can alleviate this problem by aligning the statistical distributions of features and samples among different datasets so that similar models can be applied across them. However, a majority of state-of-the-art domain adaptation methods are designed to work with large-scale data, mostly text and images, while biological datasets often suffer from small sample sizes, and possess complexities such as heterogeneity of the feature space. This Review aims to synthetically discuss domain adaptation methods in the context of small-scale and highly heterogeneous biological data. We describe the benefits and challenges of domain adaptation in biological research and critically discuss some of its objectives, strengths, and weaknesses through key representative methodologies. We argue for the incorporation of domain adaptation techniques to the computational biologist's toolkit, with further development of customized approaches.

Keywords: Machine learning; biological-scale datasets; small datasets; neuroimaging; microbiome; domain adaptation; transfer learning.

1. Introduction

In the computational biological sciences, we are interested in learning informative "truths" about biological systems through machine learning or similar quantitative modeling techniques[3]. Contrary to "purely statistical" correlations, we expect such "truths" to generalize beyond a specific dataset or population, indicating that they offer a grounded biological meaning. However, collecting (and sometimes labeling) biological datasets is difficult, expensive, and time-consuming, leading to many small but related datasets which are collected from different sources and under different environmental and experimental conditions (e.g. different labs, equipment, settings, humidity, etc). For example, in the widely used Autism Brain Imaging Dataset (ABIDE), functional magnetic resonance imaging (fMRI) data was collected at multiple sites, which hindered the ability to directly aggregate data[1]. Beyond creating challenges in data curation and metadata standards[5,6], this variability in the sources of small biological datasets creates different *domains* of data that have different statistical distributions.

While this variety is a strength that can facilitate discovery of generalizable truths, it also presents a significant challenge to computational biology: Applying knowledge gained from one dataset (a *source*) to another (a *target*) will fail if the two datasets possess highly divergent distributions – a phenomenon known as *domain shift* or *data bias*[73,74]. In short, we cannot blindly apply a model (of any kind) trained on a source dataset collected under one set of conditions to new target data and expect it to perform effectively. In an age of open datasets and keen interest in adhering to FAIR principles (Findability, Accessibility,

Interoperability, and Reuse of digital assets) to accelerate scientific discovery, it is increasingly urgent that we acknowledge the strengths and challenges of combining datasets.

To best extract generalizable insights while making use of *all* collected data from varying sources – especially in biological disciplines where data are expensive – and to apply these insights to newly collected data, we must discover how to best leverage the use of all existing and continuously growing small biological datasets[4]. In the field of machine learning, *transfer learning* aims to use knowledge gained from learning a task on one dataset to performing a similar task on a different but related dataset, with the purpose of transferring knowledge across datasets[75–79]. *Domain adaptation* (DA), a subfield of transfer learning, has been developed to address this issue of different statistical distributions by aligning the distributions of the source and target domains. Of note, while there are some similarities to "batch correction" often applied in high-throughput molecular measurements[80,81], the objective is different: domain adaptation aims to learn generalizable models across domains, while batch correction is primarily aimed at removing technical variation. Importantly, DA is more than just "lining up the features" and training a model on both datasets; not only is this often impossible to do (especially if features are unlabeled), but statistical differences between the domains can often guarantee that such a brute force aggregation is doomed to failure. Instead, through DA, a model is forced to learn *domain invariant* features, i.e. features that are common across all domains, such that the learned model can be generalized and perform relatively well on a separate target domain. Another benefit of DA is that the integration of multiple

datasets effectively increases the sample size, allowing for improved inference of statistical signals. This allows better use of available data and resources, reducing the need to collect and annotate expensive data[82–84].

However, using DA methods to extract informative and generalizable insights from different datasets is difficult in general, and is particularly difficult in computational biology. Compared to datasets typically used to train machine learning models[85–88], many "biological-scale" datasets are smaller in sample size, have many more features than samples, and have a complicated feature space (e.g. different numbers of features in each dataset, missing values, etc.). Therefore, while developing effective DA techniques that can work well with these small "biological scale" datasets to find general truths about biological systems is highly desirable, it presents a specific set of challenges to machine learning research.

In this Review, we aim to critically discuss the benefits and challenges of applying current DA methodologies and frameworks to such biological datasets. To this end, we use the token examples of functional magnetic resonance imaging (fMRI) and microbiome datasets, two seemingly different disciplines in biology, to show the common considerations critical to developing effective DA techniques in such data. Our goal is to lay out the key components that require consideration in selecting an effective DA technique, and highlight important areas of future methodological research in DA methods that can be maximally effective in biological datasets – especially as data sharing and metadata curation continues to mature.

2. Domain adaptation: a powerful tool for biological data

In the biological sciences – and especially as re-analysis and meta-analysis is facilitated through open data sharing – researchers often work with multiple distinct datasets collected through various procedures and techniques. These datasets may contain unique idiosyncrasies that are specific to a dataset, and which may or may not necessarily offer any biological insights (for example, different MRI participants or scanners[1,29,67,68], or different patient populations for microbiome profiling[89,90]). Additionally, each dataset alone may have high feature dimensionality despite small sample size, and thus may be overfit by modern, state of the art models[30,31,91,92] – making it challenging to learn robust models that will generalize. These factors make it particularly attractive to apply DA to aggregate such biological datasets, reducing overfitting and facilitating the discovery of "generalizable truths". Before assessing the challenges of doing so, we would like to briefly examine three specific benefits of DA for biological research.

2.1 Mitigating small sample size and large feature space

Ideally, a successful approach in computational biology is to fit a model with few free parameters across many samples. However, complex biological systems often need to be modeled with many free parameters, while training samples remain quite few. This degree of model complexity in the face of insufficient training exemplars can reduce generalizability and increase the risk of overfitting, where a machine learning model fits the training data all too well but fails to generalize to new, unseen data (e.g., cross-validation fails). To address this issue, domain adaptation (DA) can be used to integrate

multiple individual datasets to increase the number of training samples available. This approach helps to achieve two essential goals: it provides access to a larger and more diverse set of training data, thereby reducing the potential negative impact of having a large number of parameters, and it also encourages models to be more properly regulated so they can better extract true signals rather than being overly sensitive to noise. Increasing generalizability in this way can also support other benefits, discussed next.

2.2 Transferring knowledge

Beyond simply increasing the number of training samples available, DA can also be used to transfer knowledge across different biological contexts (different cells, tissues, organisms, individuals, ecosystems, in-vitro, and in-vivo) – assuming that domains share some commonalities in between features and task or goals. This could help scientists and physicians to transfer knowledge from some existing rich datasets to a different (but related) dataset that is smaller in size. For instance, in many situations there exists a large amount of labeled data from adults' MRI scans but much less data for infants; therefore, DA might be especially helpful to transfer insights gained from adults to newborns[93]. DA could also help transfer drug response insights gained from richly annotated pre-clinical cell lines to more poorly annotated human settings[94], or to use DNA methylation data from multiple distinct tissues to predict donors' age[95]. In general, it is highly desirable to transfer knowledge gained from existing labeled datasets to other different, but related, datasets that are sparse in terms of sample size or annotation. It is easy to envision the benefit of applying models trained on publicly-available data to a locally-collected, small dataset – a process made potentially much more powerful through DA.

2.3 Discovering generalizable patterns

As introduced above, DA can also help drive at our primary scientific goal: to reveal true, meaningful, and generalizable biological insights rather than associations that are merely due to artifacts, confounds, idiosyncrasies unique to one dataset, or meaningful biological differences between domains which are separable from a particular question at hand. This is crucial since biological datasets are often composed of many different small cohorts collected from different laboratories and under different environmental and experimental conditions[87,91]. For example, many fMRI[96–98] datasets are small, consisting of 30 human subjects or fewer per scanning site, but different hardware components or settings across MRI machines may result in data with different statistical distributions – e.g., different noise characteristics, signal magnitudes, correlations between features, or stationarity of these components across time for each scan site. In the microbiome field, the vaginal microbiome has been studied in over a dozen cohorts in the context of preterm birth[99,100], and the gut microbiome has been similarly studied in the context of colorectal cancer[101,102], yet variability in microbiome profiling across laboratories has been repeatedly noted[89].

As these smaller, individual datasets are increasingly shared and curated into large databases, challenges of discovering domain-invariant patterns while using *as much data as possible* become immediately apparent. Because of the idiosyncratic nature of each individual dataset, machine learning models can learn non-relevant information in one single 'training' dataset that can lead to incorrect *general* conclusions about biological processes. For example, even sophisticated computer vision models can discover

'shortcuts' when detecting COVID-19 from chest radiographs: Instead of detecting clinically relevant factors, they rely on confounding factors such as laterality markers or patient positioning. This not only hinders the ability of the models to generalize to new data (i.e., when tested on a new patient from a new hospital)[103] but also might lead to misinterpretation of results within a single dataset. This issue is related also to *batch effects*[104] – essentially, the effect of non-biological artifacts that changes the distribution of the data for an experimental subset of a particular experiment. When experimental batches (e.g., plates for DNA extraction, or days for MRI appointments) are also associated with the outcome of interest, it may even lead to incorrect conclusions (batch *confounding*). DA can be used to correct for these domain-specific idiosyncrasies when combining batches or cohorts, facilitating discovery of domain-invariant signals which may be more meaningful biologically.

3. Challenges of domain adaptation in bio-scale data and a path forward

Despite the clear utility of DA in biological data, its successful application to small datasets with complex features comes with significant challenges – many of which stem from the very reasons we would want to use it in the first place. In service of laying out a path forward to effective deployment of DA methodologies in biological scale datasets across multiple fields, we next explore in more detail why existing DA techniques may not be able to perform effectively on biological datasets. The purpose of this discussion is to help researchers learn to evaluate DA approaches for appropriateness in their own research, as well as to highlight deficiencies in current DA applications to biological questions which

may be alleviated through improved collaboration between DA researchers and computational biologists.

3.1. Number of samples and features

Most DA methods have been designed in the fields of computer vision, text mining, or language processing[105–108] with reference to – and evaluation on – large-scale text and image data, where there can be tens of thousands (or even millions) of samples available for training (e.g. MNIST, CIFAR10; refs.[109–111]). In contrast, the number of samples in biological datasets is often small, but they simultaneously have many features, a problem known as curse of dimensionality[112]. For instance, in a typical fMRI or microbiome dataset we might only have a few dozens to hundreds of samples while the number of features could exceed thousands[90,113,114]. As introduced above, this imbalance between the number of samples and features can potentially lead to overfitting problems[30,31], which in turn hinders the effectiveness of DA techniques on biological datasets[112]. There do exist several datasets typically used to benchmark DA approaches that may be somewhat closer in size to biological-scale data, including Office31 (ref. [115]), which contains image data of objects collected from 3 source domains with different resolutions, for a total of 4,110 images from 31 object categories (132 images per category). However, while one might hope that DA methods that have shown success on Office31[116–118] could be useful for biological data with similar sample size per category, it must be acknowledged that many biological datasets have significantly different properties than imaging data[119–122], and are even smaller, with only several hundred training samples in total. There is a need for DA algorithm development to specifically target success in the face of fewer training samples.

3.2 Differences in feature complexity

However, simply checking that DA approaches can perform adequately on small datasets is unfortunately unlikely to be enough. Another barrier to applying DA approaches to biological data is that features in biological domains are inherently much more complex than those in image data. For example, in many machine learning datasets such as MNIST or Office-31, image data are essentially pixel luminance values in the RGB and alpha channels that can be relatively simple to aggregate with other source data, for example by resizing the image[73,123–126]. In the case of biological datasets however, the inherent complexity of features can significantly hinder our ability to aggregate different sources of data. For example, biological datasets often contain missing values[127–130], or have different numbers of features with unknown mapping orders between domains[67] (i.e., which features in a source are “the same” as which features in a target domain). They can also exhibit non-linear relationships or interactions between features[91,130–132], and unique data preprocessing requirements for each source can substantially increase the complexity of developing DA techniques for biological datasets. In other words, in addition to feature-to-sample ratio and number of categories, we need to take into account the complexity and heterogeneity of biological domains before using DA techniques on biological datasets. This increased complexity stems from several sources which we next discuss in more detail.

3.2.1 Missing values

Biological samples often contain many missing feature values. For example, microbiome data typically only consists of a few taxa that are shared by the majority of samples, and even less so across cohorts. Many taxa are rare and are only detected in very low

abundances, a phenomenon known as zero inflation in statistics[133]. In human neuroimaging, PET or MRI scans combined with patients' genetic information can help with early diagnosis of Alzheimer's disease. However, the very common problem of missing values (i.e. not every subject has completed multi-modality data) can impede the ability of these multimodal models to make reliable predictions[134–136]. Missing data is less problematic in many traditional datasets used to train DA approaches, meaning that these approaches may not deal with missing data well; to be successful with biological data, DA algorithms need to adequately handle small data and missing values.

3.2.2 Heterogeneity of features

Biological domains also often possess different numbers of features, and the features also often do not lie in the same rank order across domains. For example, fMRI data from a given brain region will have different numbers of voxels from one human subject to the next, and the information represented, for example, in voxel 1 in person A is unlikely to functionally align with the information encoded by voxel 1 in person B. While functional alignment approaches have been developed[67,69], they do not explicitly perform DA operations. In microbiome research, it can be unclear whether a particular taxa is the same across datasets, especially because sometimes the measurement techniques differ (e.g., taxa are characterized using different regions of a marker gene, such that the same taxa might be represented by different features in different datasets). These examples are in stark contrast to most image-based DA approaches, which can exploit physical proximity of features (pixels) through spatial convolution or learn feature importance maps based on spatial features alone (e.g., the center of an image may often be more informative than the edges).

Additionally, domains may have some overlapping features but also some non-shared (distinct) features – i.e., those that are specific to one domain but not the other[137]. Current DA techniques may not be very effective on such datasets since domains may lack supplementary information such as labels[138] or information about matching features or samples between datasets[78]. This limitation could force researchers to remove domain-specific features and hence lose the capacity of DA models to benefit from these unique features in the learning process. Ideally, DA for biology could benefit from a specific focus on both feature alignment (ideally unlabeled) and principled ways to deal with shared versus non-shared features.

3.2.3 Distribution of feature importance

In biological datasets, feature importance distributions can be more highly skewed than in many standard benchmarks used to test DA approaches. That is, in biology, a *few* features can be *very important* for the ultimate performance of a model; in contrast, in typical benchmark datasets, many features can have similar importances[119–122]. This difference in skewness of feature importance distributions can lead to extreme challenges for many DA approaches, such that DA models which succeed even on small 'typical' benchmark datasets may fail in biological applications.

3.3 Contributions of data collection and preprocessing procedures

Biological datasets often require extensive preprocessing after the data collection stage which can be inconsistent across datasets or laboratories (DADA2 or deblur for 16S rRNA amplicon data[139,140], fMRIPrep[141] versus AFNI[142,143], or FSL[144–146] for fMRI images[147]). As

a result, machine learning methods used in biology typically are limited to being highly context- and preprocessing-specific, requiring careful design and tailoring to test the desired hypothesis appropriately[148]. This often occurs despite targeted efforts in bridging this gap by the means of setting up standards in generating and preprocessing the data[149], since some lab- and individual-specific idiosyncrasies are wholly unavoidable. For example, in fMRI data correction for subject's head movement, using different scanning sequences or scanners can introduce data shifts that makes applying DA techniques even more difficult[1,150–154]. Such preprocessing idiosyncrasies can exacerbate or interact with other batch effects, including introducing or altering interdependencies among features[91].

3.4 Interpretability of features and feature spaces

Interpretability is an important aspect of biological research, in contrast to at least some other ML applications. However, alignment steps in DA – which often require finding a latent representation of data by projecting the domains into a shared feature space[155–157] – are frequently carried out by machine learning and deep learning methods. This means that DA in biological data inherits the same problem that plagues machine learning more broadly: failures in interpretability due to the black-box nature of these machine learning and deep learning methods. In fact, the shared feature space is particularly challenging to interpret[158] because the shared feature space is defined as a latent space that bridges two or more domains, rather than the latent space defined by one domain alone. Therefore, DA research can and should aim particularly at understanding how input features are related to the common feature space when utilizing these methods[159,160].

3.5 Theoretical limitations of domain adaptation

It is also important that we discuss a critical theoretical limitation of DA, especially as it might impact biological data. The primary driver of DA's potential success is the *adaptability* between the source and target domains[161,162] – essentially, the theoretically maximal ability of an ideal model to jointly model them[163,164]. Failure of adaptability is thus a potentially fatal concern. While considering additional source domains provides the benefits of a larger and more diverse sample set (or additional labels), these domains might have inherently different distributions of features or different joint distribution with the labels, which would make a model considering them less accurate. Thus, applying DA might ultimately bring more cost than benefit[162]: in the worst-case scenario, *negative transfer* (i.e., applying knowledge from a source domain negatively affects the performance of the model in a target domain) can happen[161,165]. Crucially, the potential for negative transfer can be further amplified when working with biological data, due to its already-heterogeneous nature and the smaller sample size of each dataset. Therefore, it is imperative that the adaptability of the particular biological datasets in question be explicitly quantified or estimated before applying DA methods. Unfortunately, while there exist a few methods to quantify adaptability between domains[161,163], analysis in the context of different biological sub-fields is exceedingly rare. The development of adaptability analysis methods thus may be a fruitful and critical area of future research into DA application to biological datasets.

4. Considerations for domain adaptation

Despite the challenges noted above, even in their current state, DA approaches can still provide benefit in biological data at this critical expansion of data sharing and open science practices in biology. But there are a great many methods to choose from. How should a scientist select the best DA approaches for their own datasets or scientific questions? In this section, we outline specific considerations for biologists in selecting and applying DA approaches in their own research.

We begin this section by presenting a formal definition of *domain* and *domain adaptation*. We then present a taxonomy which can be useful in gaining a better understanding of what to search for in the literature. In this Review we focus on the primary subcategory of DA which addresses *data bias* or *covariate shift*; this DA subcategory tries to align shifts in the feature spaces between domains (or the change in the marginal distribution of data samples across domains). Other specialized subcategories of domain shift include *label shift*[166], which indicates that different domains contain different number of labels for each class, and *concept shift*[167], in which the data distribution remains the same but the conditional distribution changes (i.e. $P_s(y|X) \neq P_t(y|X)$). Interested readers should refer to these surveys[7,168] for a comprehensive overview of the different types of shifts in the DA field.

4.1 What is a domain?

A domain can be defined as $D = \{\chi, P(X)\}$, where χ is a feature space, $X = \{x_1, x_2, \ldots, x_n\}$ is an instance set with x_i denoting a given feature, n denotes the number of features or dimensions in the data (e.g., in fMRI data voxel activities or taxa in microbiome data), and $P(X)$ denotes the marginal probability distribution of all samples in that dataset. This

formal definition is typically used in discussions of DA across a wide variety of disciplines[169,170].

4.2 The terminology of domain adaptation

For a specific domain, we define the task (e.g., predicting what image a subject is looking at from neuroimaging data, or predicting a disease state from microbiome composition) as $T = \{y, f(\cdot)\}$, where y denotes the labels to be predicted and $f(\cdot)$ denotes a decision function (i.e., the posterior probability distribution of $P(y|X)$ of the joint distribution $P(X, y)$) that needs to be learned in order to map input features to the corresponding labels.

Given these definitions, domain adaptation is faced with the following problem, in which distributions or relative alignment of features across domains are different but the task remains approximately the same. Thus, a DA problem with covariate shift can be formally defined as follows:

$$P(X_{s_1}) \neq P(X_{s_2}) \neq \ldots \neq P(X_{s_k}) \neq P(X_t)\,,$$

$$T_{s_1} \approx T_{s_2} \approx \ldots \approx T_{s_k} \approx T_t$$

where s denotes the source domain, t denotes the target domain, k is the number of source domains, $P(X)$ is the marginal distribution of a specific instance set in a given domain, and T is the task performed in each domain. Here, the goal of DA is to improve the performance of target decision function $f(\cdot)_t$ in target domain D_t by leveraging the information from source domain D_s and decision function $f(\cdot)_s$ (which is learned on the

source domain after the source and target domains are aligned). In other words, DA intends to adapt the model(s) trained from a source (or sources) to a different, but related, target dataset. It does this by aligning the distributions of features and samples belonging to different domains so that the models emphasize learning *domain invariant* features that are not dependent on a specific dataset (**Figure 2.1**).

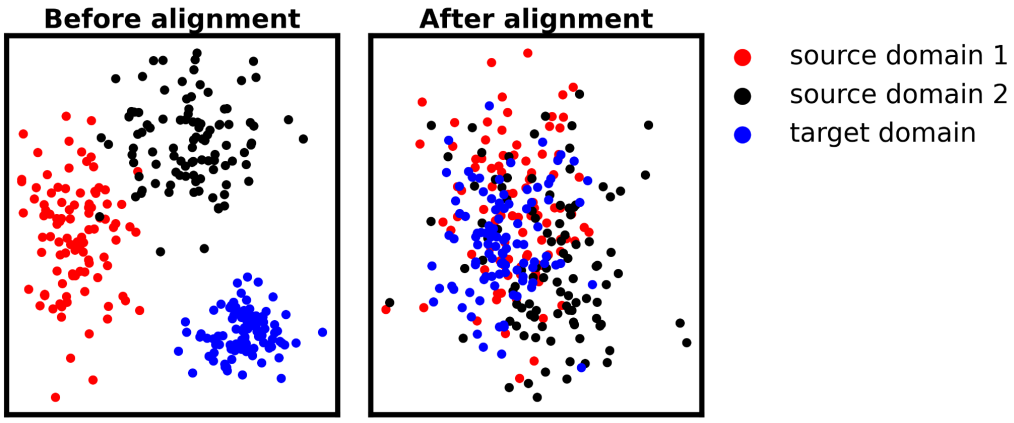

Figure 2.1. A cartoon representation of source and target domains before and after alignment. In this cartoon, features vary in their values along two dimensions, and each domain's features take on a different mean and covariance. Unless the domains are aligned, these differences could both obscure other meaningful variation in the data that is shared across domains, and prevent models trained on one domain from generalizing to another.

4.3 A taxonomy of domain adaptation

Generally, when undertaking a DA analysis, we should consider three main factors:

1. The data used to train a model may be collected from **multiple sources** or just from a **single source**.
2. Depending on the availability of labels in the target domain, we might choose **supervised**, **semi-supervised**, or **unsupervised** models.

3. The feature spaces in the source(s) and target domains can be **homogenous**, meaning that they have the same dimensionality and "meaning", e.g., feat4–0w5ure A in source 1 represents the same "type" of information as feature A in source 2; or **heterogenous**, meaning that the feature spaces may differ in terms of dimensionality and/or meaning.

In the following, we discuss these three factors in more detail. **Table 1** also shows a summary of these categories accompanied by mathematical annotations.

4.3.1 Single- vs. Multi-source

In selecting a DA method, one question you will want to ask is how many domains are present. As mentioned above, DA techniques can be divided into two categories of "single-source" and "multi-source"[12]. In single-source DA, the source domain is usually labeled, while the target domain belongs to another domain that possesses a different distribution[79,157]. Single-source DA is simpler than multi-source DA since there are only two distributions of data – source and target. Therefore, single-source DA is a good technique when there is enough data available in both the source and target domains to effectively train a model that can perform well on the target domain[171,172173,174].

However, in modern real-world data sharing initiatives, most biological data come from many sources[9,175], and using this data to its full extent can facilitate novel insights. Therefore it is advantageous to find models that leverage all available sources. This problem can be addressed through multi-source DA, which aims to combine multiple sources of labeled data to make predictions about a similar task on a target

dataset[9,12,176,177]. A naive way to solve this problem is to combine multiple sources into one big source domain and then approach the problem as a single-source DA[12,178]. However, these methods can show very limited improvement in performance – and sometimes even worse performance – in comparison to using only one source[179], specifically stemming from challenges of aligning the sources to begin with. Another way to tackle this problem could be to train a model on each source independently, apply each trained model to the target domain, and then vote for the 'correct' label in the target domain based on the prediction across sources[180]. One could also attempt to first discover domain-invariant features among all source and target domains[181], or use a two-stage alignment technique that first tries to find domain-invariant feature spaces for each source-target pairing and then align model outputs across these spaces[179]. In all cases, though, Multi-source DA is significantly more challenging than single-source DA – a problem made worse by the particular characteristics of biological data, as discussed above.

4.3.2 Supervised vs. semi-supervised vs. unsupervised

It is also important to assess what kinds of labels are available for your data, across all the domains you need to align; this will dictate whether you should select a *supervised*, *semi-supervised*, or *unsupervised* DA method. These labels have been applied in varying ways[12,79,182–184]. Here we have chosen a categorization based strictly on the usage of target labels: in unsupervised DA, no label is available in the target domain[117,157,185,186]; in semi-supervised DA[187–189], some labels are available to use; and in supervised DA, labels in the target domain are available for most samples[168]. Although the majority of DA techniques in existing literature focus on unsupervised DA (since it is often utilized for the

purpose of annotating unlabeled data in the target domain), in the case of biological data, any of the supervised, semi-supervised, or unsupervised scenarios is possible. This is because the primary goal of domain adaptation in biological settings is to uncover insights about biological systems that generalize across domains. Thus, even when labeled data are available in the target domain, one can still benefit from utilizing DA techniques on different datasets to find generalizable patterns across domains.

4.3.3 Homogeneous vs. heterogeneous

Finally, it is important to understand how the features are related across your different domains. DA can be divided into two categories based on the relationships between these features: homogeneous or heterogeneous[12,168,169]. In homogeneous DA, the source and target domains have the same feature space, $\chi_s = \chi_t$, but the data distributions of instances of these feature spaces are different, $P(X_s) \neq P(X_t)$. That is, feature 1 in domain 1 represents the same "meaning" as feature 1 in domain 2 – for example, they both represent a specific voxel at a specific coordinate in the brain, or represent the same microbe (Note: $\chi_s = \chi_t$ means that the feature space in both domains is homogenous, but if $X_s = X_t$ then it means that X_s and X_t are identical datasets such that there is no difference between the source and target datasets at all). In heterogeneous DA, conversely, the feature space is related but different between the domains. Many DA techniques that have been developed so far tend to focus on homogeneous DA[156,190–198]. For instance, the source data could be the fMRI data obtained from a subject with one scanner and the target domain is the fMRI data obtained from the same subject with the same protocol but a different scanner. Alternatively, different domains could contain gut metagenomic sequencing data from different studies aligned against the same reference database.

Addressing the domain shift in a homogeneous DA problem is relatively simpler since it is possible to simply perform the feature alignment directly on the original instances of the domains without the need to project them into a common feature space.

Unfortunately, however, most biological datasets are heterogeneous in nature[91,130] since these data are collected in different laboratories, under different environmental and experimental conditions, and sometimes even for answering different but related questions. In other words, neither the feature spaces nor the marginal distributions are the same (i.e. $\chi_s \neq \chi_t$, $P(X_s) \neq P(X_t)$). As a result, biological datasets very often have different feature dimensionalities, and sometimes these features even have different labels or come from different modalities of data collection (e.g., fMRI versus another neuroimaging modality like electroencephalography). For instance, the fMRI data from the brains of two individuals have different numbers of voxels (features) which also are not meaningfully aligned across individuals regarding their functional properties (e.g., voxel 1 in person A is unlikely to encode the same information as voxel 1 in person B) – even when the scanner, protocol and performed task are exactly the same.

	Categories | Definitions	**Domains, $\boldsymbol{D} = \{\chi, \boldsymbol{P}(\boldsymbol{X})\}$ & Tasks, $\boldsymbol{T} = \{\boldsymbol{Y}, \boldsymbol{f}(\cdot)\}$**	**Verbal description**
	Traditional ML	$D_s = D_t$ & $T_s = T_t$	When the source (i.e. training set) and target (i.e. test set) have the same distribution and the task is exactly the same.

Traditional ML vs transfer learning	**Transfer Learning (TL)**	$D_s \neq D_t$ or $T_s \neq T_t$ or both	When the source and target domains have different distributions or the performed task on source and target are different, or both.
Single- vs multi-source DA	**Single-source DA**	$P(X_s) \neq P(X_t)$ & $T_s \approx T_t$	When there is only one source domain and the marginal distribution of the feature space between source and target domain is different. The task in the target domain is similar to that in the source domain.
	Multi-source DA	$P(X_{s_1}) \neq P(X_{s_2}) \neq \ldots \neq P(X_{s_k}) \neq P(X_t)$, & $T_{s_1} \approx T_{s_2} \approx \ldots \approx T_{s_k} \approx T_t$	When there are multiple sources available which can have different distributions, and when these distributions differ from that of the target domain. The task is similar across all domains.
	Supervised	$P(X_s) \neq P(X_t)$, *with all target labels*	When source and target domains are both labeled.
	Semi-supervised	$P(X_s) \neq P(X_t)$, *with some target labels*	When source is labeled but target is partially

Supervised, semi - , or unsupervised			labeled.
	Unsupervised	$P(X_s) \neq P(X_t)$, *with no target labels*	When source is labeled but target is not labeled.
Homogeneous vs heterogeneous	**Homogeneous DA**	$P(X_s) \neq P(X_t)$ & $\chi_s = \chi_t$ & $T_s \approx T_t$	When the feature spaces have the same dimensionality and same meaning.
	Heterogeneous DA	$P(X_s) \neq P(X_t)$ & $\chi_s \neq \chi_t$ & $T_s \approx T_t$	When the feature spaces have different dimensionality or different meanings.

Table 2.1. Difference among traditional machine learning, transfer learning, and various kinds of domain adaptation. ML, machine learning; DA, domain adaptation. χ represents feature space, and $P(X)$ is the marginal distribution of instance set X, T denotes the performed task, and $f(\cdot)$ is the decision function to map each sample to the corresponding label. s denotes the source domain, t denotes the target domain, and k is the number of source domains.

4.4 Case studies and practical examples

Given the nature of most biological datasets, which often contain limited samples and originate from many different sources, the most common DA setting in this field is multi-source heterogenous DA settings. For instance, aggregating fMRI data from multiple subjects or even multiple sites[96–98] can be considered a multi-source heterogeneous domain adaptation. It is multi-source because the data is coming from multiple subjects or multiple sites with different MRI scanners, and it is heterogeneous because the number of voxels (i.e. features) from each subject and the information they represent is different. (Note: number of voxels can be equated through spatial normalization to a standardized template, but this does not address that each voxel will still represent different information across individuals.) In the microbiome field, integration of data from multiple microbiome datasets in order to predict a phenotype on a held-out study[99,100,199] is once again multi-source and heterogeneous, as data are often amplicons of different regions of the 16S rRNA gene. To illustrate the utility of existing DA approaches and explore their categorization with the taxonomy discussed above, here we select several methods to discuss in slightly more detail.

One DA method, the PRECISE method[94], has been used to predict patients' drug response based on available pre-clinical datasets such as cell lines, and patient driven xenografts (PDXs). To achieve this, the authors first extracted factors from cell lines, PDXs and human tumors using principal component analysis (PCA). Then they aligned these subspaces from human tumor data with pre-clinical data using geometric transformations, and extracted common features associated with biological processes

followed by training a regression model using consensus genes and validated with known biomarker-drug associations to accurately predict drug response in patients. In this study, DA was homogenous, as the features (genes) in the source and target domains were the same; multi-source, as various source domains were used (i.e. cell lines); and supervised, as the labels of all samples were used.

Another method, Adversarial Inductive Transfer Learning (AITL)[200], similarly aims to utilize largely available source domains such as cell lines and clinical trials to predict drug responses on small and hard-to-obtain gene expression data from patients. To this end, researchers first used a feature extractor network to map the source and target into a common feature space. This mapping aimed to alleviate the domain shift by using a global discriminator to learn domain-invariant features. Then, these domain-invariant features were used to build a regression model for the source task (i.e. predicting IC50) and a classification network to make predictions on the target task (i.e. predicting whether there is reduction in the size of the tumor). This study aimed to address both prior and covariate shifts in the source and target domains. The data used in this study came from multiple heterogeneous sources including thousands of cell lines from different cancer types. Finally, the target samples were labeled. This study can thus be characterized as a multi-source and supervised heterogeneous (i.e. drug response is categorized differently between preclinical and clinical settings) DA scenario.

Other methods such as WENDA[95] (Weighted Elastic Net for unsupervised Domain Adaptation) aim to predict a human's age using DNA methylation data, which are known

to be different across different tissues. WENDA aims to use the available DNA methylation data from some tissues (source domains) to predict the age of the human subject using DNA methylation from a different tissue (target domain) by giving more importance to features that are more robust and behave in a similar fashion across source and target domains. In this study, data from 19 different tissues with chronological age ranging from 0 to 103 years old were used as the source domain. The target domain came from 13 different tissues, with chronological age ranging from 0 to 70 years old. In the application of WENDA, the source domain remained unchanged, while each tissue type was viewed as a distinct target domain. This thus represents a multi-source, unsupervised, homogenous DA scenario.

In another study, Li and colleagues[1] propose a multi-source domain adaptation approach by using resting-state fMRI "Autism Brain Imaging Data Exchange" (ABIDE) datasets[201] from multiple academic sites (UMI, NYU, USM, UCLA). Their goal was to improve the classification accuracy of autism diagnosis by detecting biomarkers. In this study, the feature space, denoted as χ, was extracted features from fMRI sites such that $\chi_i = \chi_j$, with i and j representing different institutions (the data can be spatially normalized across participants by warping to MNI space). From this perspective, this problem is a homogeneous domain adaptation scenario. Subsequently, the authors utilized a Mixture of Experts (MoE)[202,203], combining multiple neural networks – each of which is specialized in solving a specific task – in order to improve the overall performance of the model, and adversarial domain alignment methods to minimize the discrepancies between the domains, and successfully demonstrated the advantage of using federated domain

adaptation techniques in using multi-site fMRI dataset to classify autism. Additionally, they were able to reveal possible biomarkers in the brain for autism classification. Therefore, in this framing this can be considered as a multi-source and supervised homogeneous DA problem.

Finally, Gao and colleagues proposed the deep cross-subject adaptation decoding (DCAD)[204] method: a single source, unsupervised, heterogeneous domain adaptation technique. DCAD uses a 3D feature extraction framework using 3D convolution and pooling operations based on volume fMRI data to learn common spatiotemporal patterns within a source domain to generate labels[204]. Subsequently, an unsupervised domain adaptation method minimizes the discrepancy between source and target distributions. This process considers different subjects as different sources and aids in the precise decoding of cognitive states (in working memory tasks) across subjects. To validate the approach, they applied task-fMRI (tfMRI) data from the HCP[205] dataset. The experimental outcomes revealed exceptional decoding performance, achieving state-of-the-art accuracy rates of 81.9% and 84.9% under two conditions (4 brain states and 9 brain states, respectively) during working memory tasks. Additionally, this study demonstrated that unsupervised domain adaptation effectively mitigates data distribution shifts, offering an excellent solution to enhance cross-subject decoding performance without relying on annotations.

5. Future directions

5.1 What is missing from DA approaches in biological applications?

Despite these exciting successes, continued development of DA approaches tailored to the challenges of biological data is critically needed. This is especially important in light of the increasing availability of curated open datasets, complemented by increasing standardization of metadata standards[5,6]. We thus hope the machine learning community will continue to develop techniques that can address relevant limitations of biological datasets, including:

1. Models must be able to capture the non-linear and complex patterns in biological systems, ideally with minimal or no assumptions. Therefore, many linear-based domain adaptation techniques (usually focused on some sort of transformation from source to target domain) might not be adequate.
2. Ideally we want to utilize domain adaptation to discover the underlying mechanisms of biological phenomena, rather than simply aggregating data for automatic annotation. Unfortunately, many existing techniques are primarily developed for addressing automatic annotation of unlabeled data. Therefore, to fully unleash the power of DA in biological systems, we must focus on methods that seek to discover domain-invariant features that are common across datasets. This usually happens by mapping all domains into a common feature space.
3. This domain-invariant mapping should be done using methods that work with limited data in individual cohorts. Although deep learning models are great tools to uncover highly nonlinear and complex relations in data with no specific assumptions, they often require many samples. Recently, simpler neural network architectures such as TRACE[2] and Fader networks[151] have shown promise with

small fMRI datasets. However, many of the powerful neural network architectures such as GANs might not be suitable for biological datasets as they usually require vast amounts of data[206,207].

4. Methods should be developed to address domains' *adaptability* with specific focus on biological datasets. As mentioned earlier, methods do exist to quantify adaptability between domains[161,163], but limited attention has been paid to how such methods may fare in biological contexts.

In sum, it is incumbent upon us in the biological disciplines to challenge machine learning research to design more flexible and broadly applicable DA methods that can perform under the constraints of real-world biological datasets. An important step towards this goal will be to test and evaluate existing approaches on our own data, and on data available through increasingly comprehensive and consistently annotated shared data repositories, to comprehensively explore and categorize their current shortcomings. Thus, we hope that, with the help of the topics discussed in this Review, researchers in biological disciplines will feel empowered to try out existing DA approaches and to help catalog their successes and shortcomings.

If you would like to use DA techniques to augment your own data processing pipeline, we urge you to begin by gaining a comprehensive perspective on your data using the definitions and taxonomy described above. For example, How many sources do you have available? What is the sample size in each source? Do these sources contain equal amounts of features? If not, what are the nature of features in each source? Are these

features in each source known and have a label? What task are you trying to achieve? Depending on the answers, you can choose the appropriate DA approaches, and set about examining their successes or failures. We hope that the tools and information provided in this Review will encourage you to do so, and to report your findings so that iterative improvements in DA approaches can be made to best serve our fields.

5.2 Promises for the future

In this piece we have focused on human neuroimaging (specifically fMRI) and microbiome sciences as token examples to speculate the potential promises of DA in computational biology as a whole. We hope that these selected case studies have helped to show off the potential of DA in numerous and varied biological disciplines, from electrophysiology, multi-omics, DNA sequencing, and scRNA sequencing to and protein localization – all of which face similar challenges in data collection and labeling to the case study fields discussed here. Differences in equipment, experimental setup, or even individuals can lead to a shift in the distribution of data, even when the task is identical. In all cases, however, our goal as researchers and clinicians is to go beyond domain-specific or dataset-specific models in order to discover domain-general and informative “truths” about biological systems.

Thus, DA could be extremely useful to aggregate diverse biological datasets available across the Open Science Framework, OpenNeuro, Neurosynth, Dryad, CEDAR, and more in search of meaningful and even clinically relevant outcomes[208–211]. But much work is needed to address the existing challenges. It is the intention of this paper to help and facilitate these processes by bringing more awareness of DA, and the need to develop

new techniques that are compatible with the limitations of biological datasets in order to make it accessible to biologists. If we are successful in identifying the challenges of performing DA on biological data, we are optimistic that DA and transfer learning methodologies can greatly benefit biologists.

Chapter 3. Exploring Interactions Between VTC and PFC Using Domain Adaptive Task-Relevant Autoencoding

Abstract

Understanding the interactions between brain regions during visual processing is crucial for decoding the neural basis of cognition. This study focuses on the ventral temporal cortex (VTC) and the prefrontal cortex (PFC), key regions involved in object recognition and conscious awareness, respectively. We propose a novel machine learning approach, Domain Adaptive Task Relevant to Autoencoding via Classifier Enhancement (DATRACE), to explore the shared neural representations between these regions. By leveraging domain adaptation techniques, DATRACE aims to align and decode the voxel activities of the VTC and PFC to find the shared feature space between these two regions. Our approach involves an encoder-decoder architecture that predicts voxel activities in both the VTC and PFC from a shared latent space, while a logistic classifier ensures the relevance of these representations for object recognition tasks. Through rigorous evaluation, we determined the optimal dimensionality of the shared representation and employed representational similarity analysis to examine the clustering of object categories in this shared feature space and within VTC itself. Preliminary results demonstrate that the shared representations between VTC and PFC capture similar categories of objects (e.g. insects, animals, objects). To interrogate the features in this shared space, we conducted a feature perturbation analysis in the shared feature space by perturbing one individual feature at a time while keeping the rest of the feature intact. However, this perturbation analysis indicated that single feature disruptions do not significantly affect classifier performance, implying the need for further studies where we

perturb combinations of these features to reveal meaningful semantic interpretations of the shared dimensions encoding visual objects in both PFC and VTC.

1. Introduction

The human brain functions as a network of interconnected regions that process sensory inputs[212,213]. Visual processing is particularly complex due to the involvement of multiple specialized areas[214,215]. In this study we examine the interactions between two key regions involved in conscious visual perception: the ventral temporal cortex (VTC)[216] and the prefrontal cortex (PFC). The VTC is central to object recognition and categorization, while the PFC is instrumental in bringing conscious awareness to visual stimuli[217]. The exchange of information between these regions is essential for transforming static visual inputs into enriched, contextually relevant interpretations that guide behavior[218,219].

Communication between the VTC and PFC involves both bottom-up and top-down processes[220]. Bottom-up processes typically involve the direct flow of sensory data from lower sensory areas to higher cognitive areas, facilitating immediate perceptual experiences[221]. Conversely, top-down processes involve the modulation of sensory interpretation by higher cognitive functions such as attention, memory, and expectation, which are critical for integrating sensory input with previous experiences and knowledge[222]. Understanding the interplay of these processes is crucial for exploring how the brain interprets complex visual scenes[223].

Recent advances in neuroimaging and computational neuroscience have shed light on the functional connectivity and interactive dynamics of the VTC and PFC[224]. However, many details about the specific features (e.g. animacy, size, color) of visual stimuli that

are exchanged between these regions and how those shared features influence cognitive processes remain poorly understood. A key question in this domain is how information transfer occurs at a neural level and which visual features are most significant in these inter-regional interactions.

In this study, we incorporate a machine learning approach known as domain adaptation (DA) to address this challenging problem. Originally developed within the field of computer science, DA is designed to enhance the performance of models to recognize and utilize information across different datasets. This is achieved by transferring knowledge from a well-labeled "source domain" to an unlabeled or poorly labeled "target domain"[7,9–11]. DA accomplishes this by aligning the distributions of the source and target domains, allowing a model trained on the source domain to effectively predict labels in the target domain[12]. In the context of brain imaging, we can take advantage of this “byproduct” of DA in order to find a shared representation between different brain regions (e.g. VTC and PFC). This alignment helps in understanding what these interconnected activity patterns mean between these brain regions.

Our approach, the Domain Adaptive Task Relevant to Autoencoding via Classifier Enhancement (DATRACE), is designed to capitalize on this DA methodology. DATRACE aims to uncover shared neural representations that facilitate object recognition across both brain regions. The initial phase of our model development focuses on determining the optimal dimensionality within the network's bottleneck—a critical layer where the most important features for both regions are hypothesized to converge. By analyzing how well

the predicted voxel activity in the PFC aligns with actual observed activity, we refine our model to better capture the essence of shared cognitive processing between these regions.

We also utilize a logistic multi-class classifier connected to the bottleneck of the network. This classifier's role is to categorize objects based on the shared representations, thereby testing the functional relevance of these neural codes and putting the constraints on finding features that solve for the same task in both regions.

Finally, we apply individual feature perturbation to this shared space and conduct representational similarity analysis to probe deeper into the nature of the features encoded within this shared representation. These analyses aim to disentangle the complexities of the encoded features to better understand what categories and attributes of objects they represent. Through these analyses, we hope to elucidate the functional integration and communication between the VTC and PFC, as well as to explore the interpretability and applicability of deep learning models in the realm of neuroscientific research.

2. Reconstruction-based domain adaptation

Reconstruction-based domain adaptation (DA) approaches have been gaining attention due to their capability to improve the transfer of knowledge between domains by focusing

on the reconstruction of data from both the source and target samples[225]. This methodology assumes that successful data reconstruction will help to enhance the adaptability and performance of DA algorithms. The main strategies in reconstruction-based DA can be broadly classified into two categories: encoder-decoder reconstruction and adversarial reconstruction. In this study we mainly focus on encoder-decoder reconstruction techniques. This approach primarily utilizes stacked autoencoders (SAEs)[226,227] to implement the encoder-decoder reconstruction mechanism[228]. An encoder in this setup aims to capture a latent low-dimensional representation of the input data, whereas the decoder focuses on reconstructing the input data from this abstract representation which results in the minimization of reconstruction error. This process not only helps in preserving the essential characteristics of the data but also ensures that the features learned are robust and significant for DA tasks.

Additionally, autoencoders can be progressively stacked to form a deeper architecture, where each layer is trained to refine the feature representation obtained from the previous layer. This stacking enhances the model's ability to encode more complex patterns and relationships within the data. For instance, the stacked denoising autoencoder introduced by Glorot et al[228] extends this idea by incorporating a noise reduction mechanism, which further improves the robustness of the feature encoding against variations in the input data. However, one challenge in using these stacked autoencoders is that these networks usually contain many parameters which in turn can lead to overfitting problems – especially when we have a dataset with poor sample-to-feature ratio as is common in human neuroimaging datasets. For this reason we propose a Domain Adaptive Task

Relevant to Autoencoding via Classifier Enhancement (DATRACE), which follows a similar architecture and logic to that of our previous TRACE model[2].

3. Methods

In this section, we outline the dataset used in this study. We then detail the network architecture, objective function, and computational analysis employed to identify the optimal bottleneck and to measure the distances between different object categories. To explore the shared representation between VTC and PFC, we examined representational dissimilarity matrices in the shared feature space as well as the low-dimensional representation of VTC. Lastly, we perturbed features at the bottleneck layer and observed how these perturbations affected the classifier's predictions.

3.1 Dataset

We utilized a previously collected and partially reported fMRI dataset[49] from 60 healthy individuals who each viewed 3600 images spanning 40 categories of objects, including 30 animal types and 10 man-made objects. While these images were being viewed, BOLD signals from the ventral temporal cortex (VTC) and prefrontal cortex (PFC) were recorded. The number of voxels in the VTC and PFC varied for each individual, but on average, there were 2382 ± 303 and 1452 ± 195 voxels for VTC and PFC respectively across the 60 subjects.

3.2 DATRACE architecture

To identify the shared representation between voxel activities in the ventral temporal cortex (VTC) and the prefrontal cortex (PFC), we developed an architecture capable of uncovering a low-dimensional representation that is shared between VTC and PFC. As shown in **Figure 3.1**, the architecture of DATRACE consists of four main components: an encoder, a classifier, and two decoders.

Encoder: This component receives voxel data from the VTC as input features. The data passes through a hidden layer with tangent hyperbolic (Tanh) activation function, projecting the input into a 500-dimensional space. To prevent overfitting, a dropout layer follows with p=0.1 (i.e. 10 percent of neurons are dropped out in the feed forward computation). The bottleneck layer then holds the shared representation between the VTC and PFC.

Classifier: The second component, a logistic classifier similar to that used in TRACE^2, ensures that the shared representation is optimized for the specific task (i.e., object recognition). This layer uses a softmax activation function to provide a probability distribution across all classes, and applies ridge regularization to prevent overfitting.

VTC Decoder: The third component is a decoder that reconstructs VTC voxels from the encoded information in the bottleneck. This decoder has a hidden layer of 500 units with Tanh activation function, followed by a dropout layer for regularization purposes. The final layer matches the original VTC dimensions.

PFC Decoder: The fourth component predicts voxel activities in the PFC based on the encoded information in the bottleneck. It also contains a hidden layer of 500 units with Tanh activation and is regularized through a dropout layer. The final layer matches the original PFC dimensions.

Since this architecture can predict both VTC and PFC activities from the latent representation in the bottleneck through W_{BV} and W_{BP} respectively (**Figure 3.1**), we hypothesize that these features must hold a shared representation between the two regions. This shared representation could enhance our understanding of the information shared between these regions in object recognition tasks.

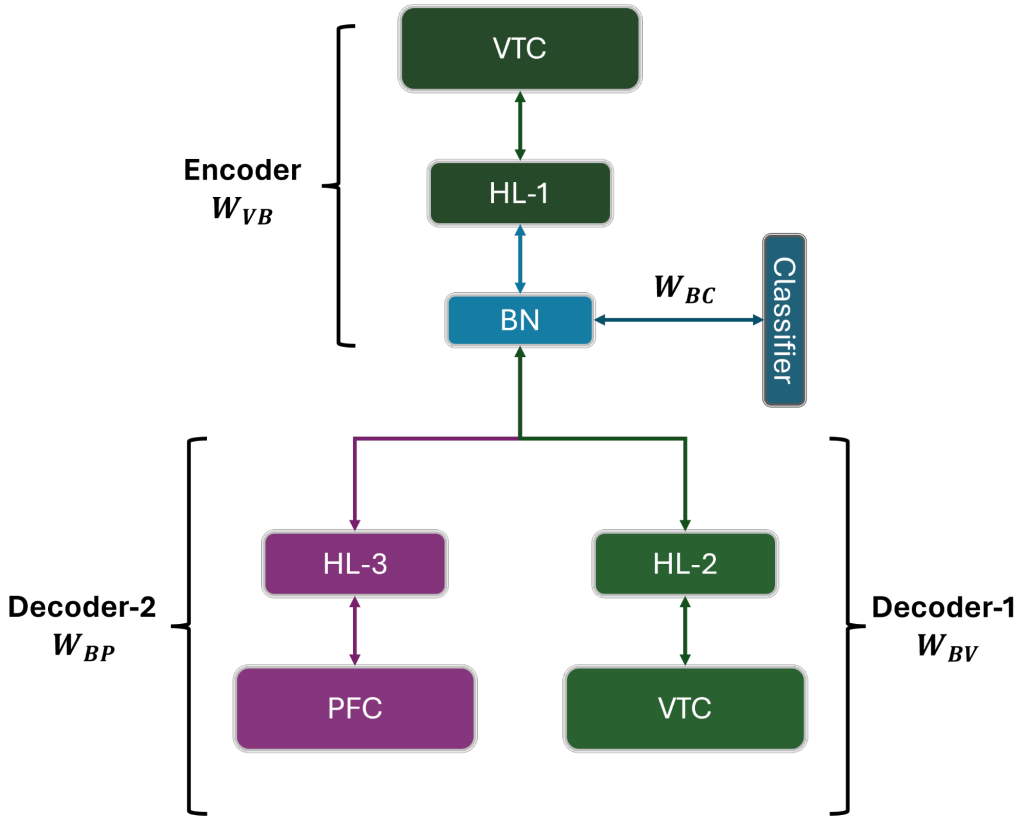

Figure 3.1. The DATRACE architecture functions as follows: the encoder (W_{VB}) maps VTC voxels into a bottleneck layer, while the logistic classifier (W_{BC}) ensures that this shared space is optimized for the task at hand. Decoder-1 (W_{BV}) utilizes the shared features to reconstruct voxel activities in VTC, and Decoder-2 (W_{BV}) projects the shared feature space into PFC.

3.3 DATRACE objective function

Given the architecture described above, the loss function in DATRACE has three components. We used mean absolute error (MAE) for reconstruction and prediction of VTC and PFC respectively, and as for the classifier we used multiclass cross-entropy loss function. In order to control the contribution of the classifier to network we used a parameter α and we set it to 0.1 as was reported by Orouji et al.[2] **Equation 3.1** describes the final objective function of the network.

$$Loss_{final} = \frac{1}{m \times n_V} \sum_{i=1}^{m} \sum_{j=1}^{n} abs(\hat{V}_{ij} - V_{ij}) + \frac{1}{m \times n_P} \sum_{i=1}^{m} \sum_{j=1}^{n} abs(\hat{P}_{ij} - P_{ij}) - \frac{\alpha}{m} [\sum_{i=1}^{m} \sum_{c=1}^{k} y_{ci} \, log(\hat{y}_{ci}) + \lambda \sum w_{BC}^2] \quad (3.1)$$

Where V and P denote the actual voxel activities, and $\hat{V}$ and $\hat{P}$ represent the predicted voxel activities in the VTC and PFC, respectively. Here, n_V and n_P refer to the dimensionality of the VTC and PFC, while m indicates the number of samples, and c represents the number of classes. y and $\hat{y}$ denote the actual and predicted class labels, respectively, while λ is the regularization factor which was set to 0.001.

3.4 Training DATRACE

To train DATRACE, we divided the dataset into training and testing sets. The training set consisted of 2700 trials, while the test set included 900 trials. The training data was scaled between -1 and 1, and this scaling transformation was also applied to the test set. For optimization, we utilized the Adam algorithm[229] to minimize the loss function. After experimenting with various epoch counts, we settled on 50 epochs, as this number achieves the lowest loss on the test set.

3.4 DATRACE Optimal Bottleneck Dimension

To evaluate the performance of the network we looked at two metrics. The first metric was reconstruction/prediction fidelity which measures the precision of the predictions of the voxel activities in VTC and PFC respectively, similar to what was reported previously[2]. The second metric was the classification accuracy of predicted VTC and PFC. For that we trained a separate logistic classifier on the predicted VTC and PFC and reported the prediction accuracy on the test set. We evaluated these metrics across varying dimensions of the bottleneck (i.e. d = 20, 30, 50, 100, 150, 200, 500). This process was repeated for all 60 subjects, and we report the mean and standard error.

3.5. Representational Dissimilarity Matrix

To build the representational dissimilarity matrix (RDM) for each subject, we measured the Euclidean distances between the bottleneck activity patterns elicited by each object in the test dataset, across all pairs of objects in all the 40 classes. We computed the mean Euclidean distance between each pair at the bottleneck layer with an optimal dimension

of d = 100 (see Results for justification of this selected dimensionality). First, we calculated the pairwise Euclidean distances across all possible pairs of trials, then grouped all instances belonging to the two classes that are being compared. The mean distance between these groups was considered as the measure of dissimilarity between the two classes. The resulting RDM is a 40x40 matrix which shows the distances between each pair of classes. Finally, we normalized the distances in the matrix to lie between 0 and 1. **Equation 3.2**. describes how the RDM was calculated.

$$D(i,j) = \frac{1}{S \times M} \sum_{s=1}^{S} \sum_{m=1}^{M} d_{s,m}(i,j)) \quad (3.2)$$

where S is the number of subjects, M is the number of pairs between categories i and j, and d is the Euclidean distance between two pairs of exemplars.

To contrast the low-dimensional representation of VTC alone with the shared representation, we computed the RDM for the VTC's low-dimensional representation with 100 dimensions. For that, we first utilized the TRACE model[2] as previously reported to derive this low-dimensional representation for VTC only, and subsequently calculated its RDM following the procedures described above.

3.6. Bottleneck Feature Perturbation

In the final analysis, we extracted features from the DATRACE bottleneck layer at d=100 for the test set and trained a separate logistic classifier on these shared representations. To understand the role of features in the bottleneck of the network, we perturbed each feature individually by setting it to zero while keeping the rest of the feature unchanged,

following by feeding the modified features into the pre-trained classifier, and examined the resulting confusion matrix across the trials. As a result, each trial in the confusion matrix contained 100 predictions.

4. Results

In this section, we discuss how we determined the optimal dimensionality of the DATRACE shared feature space (i.e., bottleneck). We also evaluate the resulting representational dissimilarity matrix (RDM) from the shared representations, and examine the confusion matrix resulting from feature perturbation to investigate these shared spaces in individuals.

4.1. Optimal bottleneck dimensionality

To determine the optimal dimensionality for the shared representation in the bottleneck, we trained the network with varying bottleneck dimensions (i.e., d = 20, 30, 50, 100, 200, 250, 500) and evaluated the prediction fidelity and classifier accuracy for both VTC and PFC regions, as discussed in the methods section. **Figure 3.2** depicts the mean prediction fidelity and prediction accuracy for 60 subjects for both VTC and PFC. We found that at d = 100, fidelity for both brain regions started to plateau, and the classifier accuracy reached its maximum level. Therefore, we selected d = 100 as the optimal dimension for these shared representations and conducted the rest of the analysis using this dimension. Consistent with the observations reported by Orouji et al.[2], we found that for both VTC and PFC, it was possible to obtain a more task-relevant (‘purer’)

representation through DATRACE, indicated by the fact that VTC reconstruction and PFC prediction classifier accuracy surpassed classifier accuracy for a separate model trained on these inputs (for both regions), all without significantly compromising the fidelity score. This suggests that our model can effectively capture the essential features needed for object recognition tasks while maintaining a high fidelity in its predictions.

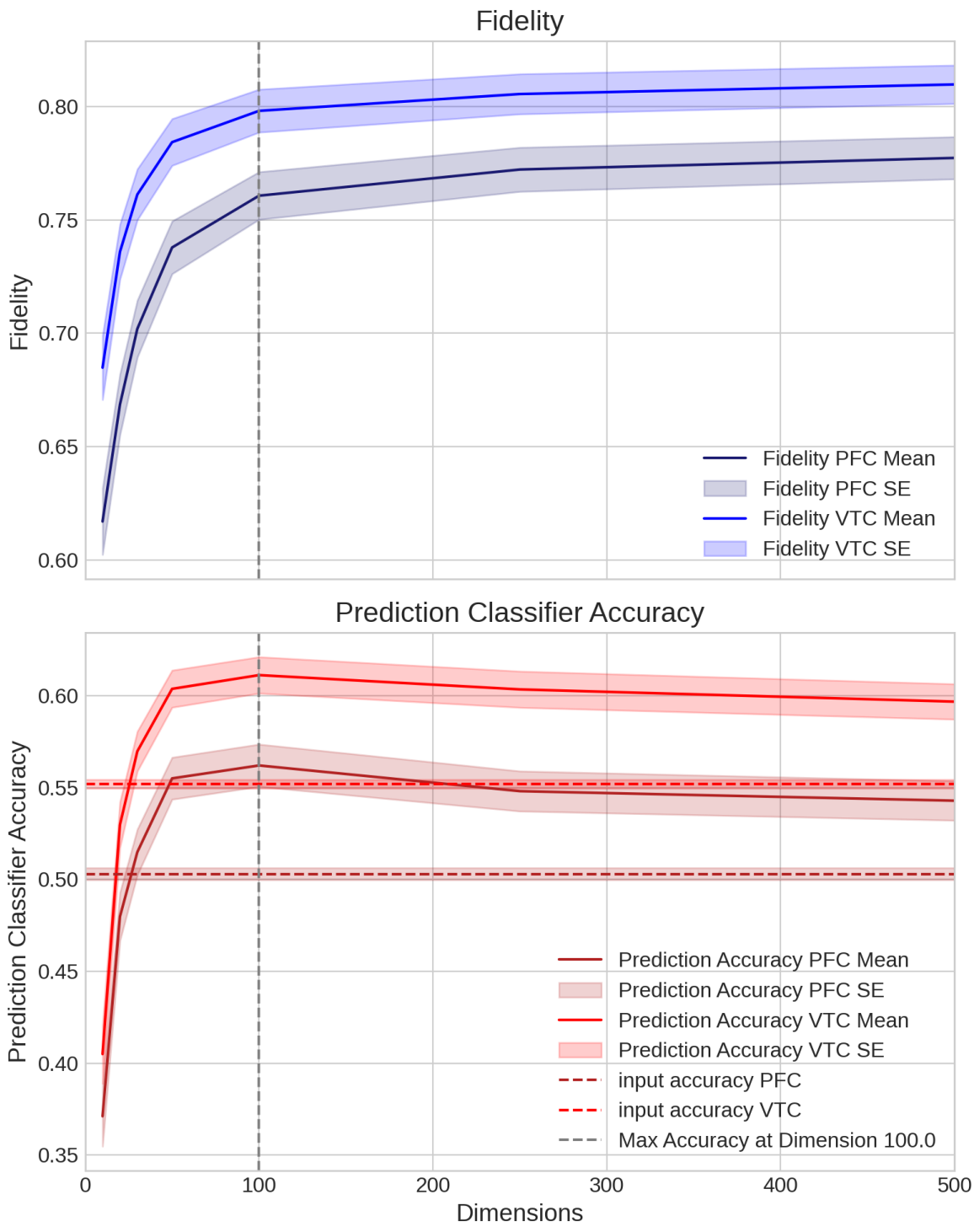

Figure 3.2. Mean values for Fidelity and Prediction Class Accuracy for 60 subjects across various bottleneck dimensions (d = 20, 30, 50, 100, 200, 250, 500) are presented. In the top panel, Fidelity plateaus at d=100 for both predicted PFC (dark blue) and reconstructed VTC (light blue). Similarly, in the bottom panel, Prediction Class Accuracy for VTC (light red) and PFC (dark red) reaches its maximum at d=100. The shaded areas indicate the

standard error. Dashed horizontal lines indicate classification accuracy achieved for a separate classifier trained on the VTC and PFC empirical data rather than predicted PFC or reconstructed VTC.

4.2. Representational Dissimilarity Matrix (RDM)

We generated a Representational Dissimilarity Matrix (RDM) for each subject as described in the methods section, for both (a) the shared representations between VTC and PFC and (b) a low-dimensional representation of VTC alone trained via the TRACE model. With the RDM we can ask two questions. Firstly, what categories of objects do these regions represent in the form of clusters? And secondly, how do these clusters differ between VTC-only representations versus shared VTC-PFC representations, if at all? To facilitate this approach, we organized the object categories in the RDM such that subgroups of insects, reptiles, birds, mammals, aquatic animals, and man-made objects are positioned adjacent to each other. Upon analyzing the RDM profile for the shared representations between VTC and PFC (**Figure 3.3a**), we observed some distinct clusters. Notably, there exists a cluster for insects and reptiles with similar characteristics. Additionally, distinct clusters were evident for birds, mammals, and aquatic animals as well as a distinct cluster for man-made objects. This clustering in the shared representation illustrates how these brain regions process different categories of visual objects.

The RDM profile of the low-dimensional representations for VTC was very similar to the shared representation between VTC and PFC. This similarity suggests that the categorical object representations in VTC are very similar to those shared with PFC. However, to confidently affirm these observations, further detailed investigation is

needed. Such studies would help validate the consistency of these findings across different data sets and experimental conditions.

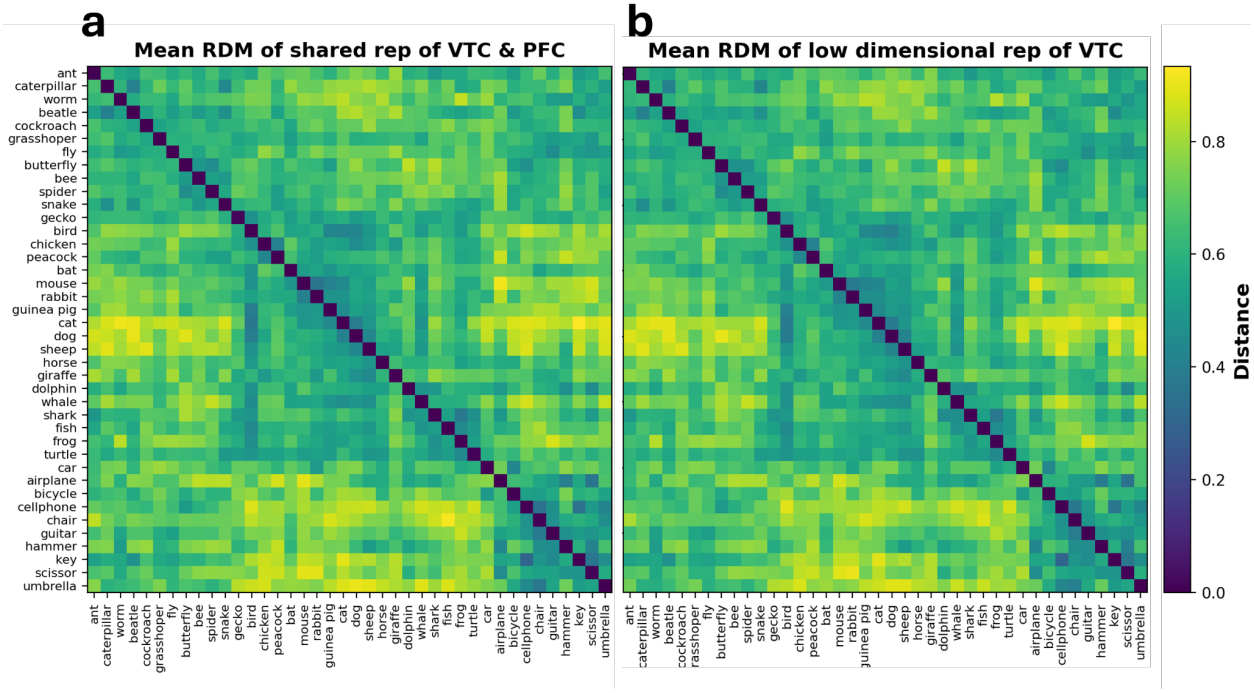

Figure 3.3 RDMs for the shared representation of VTC and PFC (left) and for the low dimensional representation of VTC (right). Both RDM profiles are very similar (but they are not identical), suggesting the information encoded in VTC alone is similar to the information being shared between VTC and PFC.

4.3. Feature perturbation

To analyze the role of each individual feature in the shared representation between VTC and PFC within the test set in order to better understand their semantic meaning, we implemented a method of perturbation where each feature was set to zero one at a time, while keeping the rest of the features' activities unchanged. We then input these altered features into a logistic regression classifier that had been previously trained on the training set. This was done to observe how the classifier's predictions varied with each feature's perturbations. The results of these tests were then gathered into a confusion

matrix, with the rows representing instances from the test set and the columns representing the features when they were set to zero. This analysis was carried out on three subjects to explore whether any consistent patterns emerged that might reveal what each individual feature encodes in an object. **Table 3.1** presents the preliminary results for these three subjects. We observed that perturbing individual features does not cause the classifier to make mistakes. This is likely because the shared representation consists of many features (around 100), making the impact of a single feature negligible. Unfortunately, without confusing the classifier, we cannot speculate on what specific features in the latent representation are encoding in an object. Therefore, further investigation is necessary, in which we might perturb various combinations of features to effectively confuse the classifier.

	Feat0	Feat1	Feat2	Feat3	Feat4	Feat5	Feat6	Feat7	Feat8	Feat9	Feat10
cat	cat	cat	cat	cat	cat	cat	cat	cat	cat	cat	cat
cat	cat	cat	cat	cat	cat	cat	cat	cat	cat	cat	cat
cat	cat	cat	cat	cat	cat	cat	cat	cat	cat	cat	cat
cat	cat	cat	cat	cat	cat	sheep	cat	cat	cat	cat	cat

Table 3.1. Confusion matrix for the first 10 features of the shared representation of the instance “cat" for one subject. Our preliminary results show that perturbing individual features will not cause significant confusion in the classifier prediction.

5. Discussion and future directions

In this study, we explored the information transfer between the ventral temporal cortex (VTC) and the prefrontal cortex (PFC) during visual processing. The VTC, essential for object recognition and categorization, is known for operating at the interface of perception and cognition, while the PFC contributes to the conscious awareness of visual stimuli[217].

Understanding the interaction between these regions is crucial for unraveling the cognitive processes underlying visual object recognition.

Our approach, Domain Adaptive Task Relevant to Autoencoding via Classifier Enhancement (DATRACE), leverages domain adaptation to uncover shared neural representations between VTC and PFC. The encoder-decoder architecture effectively predicts voxel activities in both VTC and PFC from the shared representation, while the classifier validates the relevance of these representations for object recognition tasks. This methodology provides a robust framework for analyzing neural interactions for any other regions in the brain. We also arranged classes of objects in a hierarchical manner (e.g. insects, birds, mammals, etc), similar to those reported in the THINGS dataset[230] and explored the resulting RDMs between these hierarchical categorization for all 60 individuals. The preliminary results indicate that VTC encodes object classes by representing animacy, which is in agreement with previous findings[231]. Additionally, we found that there are apparent categorizations encoded in VTC based on insects versus animals (i.e. birds, mammals, aquatic animals). Our model indicates that the nature of the information passed to PFC (i.e. shared with PFC) is very similar to those encoded in VTC. Further investigation is needed to examine whether the same pattern exists in other regions of the visual stream.

We also performed individual feature perturbation to investigate the nature of these features. Preliminary results indicate that individual features do not significantly confuse the classifier probably due to the high dimension of the shared space (i.e. d=100).

Therefore, further investigation is necessary to perturb combinations of features to determine what combinations of features encode for specific attributes in objects.

DISCUSSION OF THE DISSERTATION

In this dissertation we first introduced TRACE (a task-relevant autoencoding approach) to achieve two goals. First to find a lower dimensional representation of high dimensional fMRI data and secondly to distill information related to the specific task. We then presented a perspective on domain adaptation (DA) techniques and their potential utility in the context of biological dataset. Finally, we integrated DA and TRACE (DATRACE) in order to find a shared representation between brain regions (i.e. VTC and PFC). Our findings indicate that TRACE and DATRACE are able to find a lower dimensional representation that is optimized to represent both the input data and the task at hand. Additionally, we demonstrated that DATRACE is indeed capable of finding a shared representation between these brain regions. Further investigation is necessary to interrogate and explain the nature of this shared representation.

Application of TRACE on Small fMRI Datasets

In the first chapter, we proposed the Autoencoder with Classifier Enhancement (TRACE) model designed to distill task-relevant information small-scale datasets fMRI datasets. Many state-of-the-art deep learning models, despite being powerful, tend to overfit when used with small datasets due to the poor sample-to-feature ratio in fMRI datasets. TRACE mitigates these issues by adopting a simple autoencoder architecture with relatively fewer parameters. Central to the architecture of TRACE is the integration of a logistic classifier into its bottleneck layer, compelling the model to learn not just low-dimensional representations but task-relevant ones.

Our results showed that TRACE outperformed principal components analysis (PCA), standard autoencoders (AEs), and variational autoencoders (VAEs) across several metrics (i.e. reconstruction fidelity, bottleneck classifier accuracy, reconstruction class specificity, and reconstruction classifier accuracy). The attachment of a classifier to the bottleneck of TRACE not only optimized it to extract task-relevant low-dimensional representations but also helped to reduce noise and task-irrelevant information in the original input space. This was particularly evident in the fMRI dataset, where TRACE's reconstructions achieved higher classification accuracy compared to the raw input. Moreover, TRACE maintained its superior performance even under extreme data truncation in MNIST and Fashion MNIST datasets which highlighted its robustness in dealing with datasets with a poor sampel-to-feature ratio.

Utility of Domain Adaptation in Biological Data

Despite advances in transfer learning and domain adaptation (DA), methods that are suitable and effective on complex biological data still require further development. The growing availability of open datasets highlights the need for models that can capture the common knowledge between them .

Many existing DA techniques are primarily designed for automatic annotation of unlabeled data by transferring the knowledge from available labeled datasets. In biological sciences however, researchers are more interested in uncovering the underlying biological ‘truths’. Fortunately, DA can be a potentially great tool for this purpose. This is because DA models align the distributions of different datasets which consequently results in revealing domain-invariant features across those data.

Exploring and evaluating existing DA approaches on biological data is crucial for identifying their current potentials and limitations. Before adopting any DA approach, it is important to consider the properties of the data, including the number of sources, sample sizes, and feature nature, to choose appropriate DA techniques and examine their efficacy.

DATRACE: Integrating TRACE and Domain Adaptation

Building on the principles of TRACE and domain adaptation, we proposed Domain Adaptive Task Relevant to Autoencoding via Classifier Enhancement (DATRACE) to explore neural representations that are transferred between brain regions. Specifically, we investigated the interaction between the ventral temporal cortex (VTC) and the prefrontal cortex (PFC) during visual tasks.

DATRACE leverages an autoencoder-based DA to uncover shared neural representations between VTC and PFC. Using an encoder-decoder architecture, DATRACE predicts voxel activities in both regions from a shared representation, with a classifier attached to the bottleneck that ensures the relevance of these representations to the visual task. Preliminary results indicated that the encoded object classes in VTC are very similar to those shared and transferred to PFC.

Additionally, we performed feature perturbation to investigate the individual features' role. However, we found that single features did not significantly confuse the classifier which is probably because many features in this shared space are contributing to the prediction of the classifier. Further investigation is needed to understand what combinations of features encode for specific object attributes.

DATRACE's framework provides a promising method for analyzing the interactions between VTC and PFC, and potentially other brain regions. Future research should explore DATRACE's applicability and its potential to enhance our understanding of visual perception.

References

1. Li, X. *et al.* Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med. Image Anal.* **65**, 101765 (2020).
2. Orouji, S., Taschereau-Dumouchel, V. & Cortese, A. ‘ Task-relevant autoencoding’ enhances machine learning for human neuroscience. *arXiv preprint arXiv* (2022).
3. Ross, L. N. & Bassett, D. S. Causation in neuroscience: keeping mechanism meaningful. *Nat. Rev. Neurosci.* **25**, 81–90 (2024).
4. Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S. & Bhattacharyya, D. K. Big Data Analytics in Bioinformatics: A Machine Learning Perspective. *arXiv [cs.CE]* (2015).
5. Zizienová, M. New OSF Metadata to Support Data Sharing Policy Compliance. (2023).
6. Musen, M. A. *et al.* The center for expanded data annotation and retrieval. *J. Am. Med. Inform. Assoc.* **22**, 1148–1152 (2015).
7. Wilson, G. & Cook, D. J. A Survey of Unsupervised Deep Domain Adaptation. *ACM Trans Intell Syst Technol* **11**, 1–46 (2020).
8. Jiang, J. J. A literature survey on domain adaptation of statistical classifiers. (2007).
9. Sun, S., Shi, H. & Wu, Y. A survey of multi-source domain adaptation. *Inf. Fusion* (2015).
10. Farahani, A., Voghoei, S. & Rasheed, K. A brief review of domain adaptation. *Advances in Data Science* (2021).
11. Zhao, S. *et al.* A review of single-source deep unsupervised visual domain adaptation. *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 473–493 (2022).
12. Zhao, S., Li, B., Reed, C., Xu, P. & Keutzer, K. Multi-source Domain Adaptation in the Deep Learning Era: A Systematic Survey. *arXiv [cs.LG]* (2020).
13. Sidhu, G. S., Asgarian, N., Greiner, R. & Brown, M. R. G. Kernel Principal Component Analysis for dimensionality reduction in fMRI-based diagnosis of ADHD. *Front. Syst. Neurosci.* **6**, 74 (2012).
14. Mannfolk, P., Wirestam, R., Nilsson, M., Ståhlberg, F. & Olsrud, J. Dimensionality reduction of fMRI time series data using locally linear embedding. *MAGMA* **23**, 327–338 (2010).
15. Yang, Z., LaConte, S., Weng, X. & Hu, X. Ranking and averaging independent component analysis by reproducibility (RAICAR). *Hum. Brain Mapp.* **29**, 711–725 (2008).
16. Chen, P.-H. (cameron) *et al.* A Reduced-Dimension fMRI Shared Response Model. in *Advances in Neural Information Processing Systems* (eds. Cortes, C., Lawrence, N., Lee, D., Sugiyama, M. & Garnett, R.) vol. 28 460–468 (Curran Associates, Inc., 2015).
17. Bejjanki, V. R., da Silveira, R. A., Cohen, J. D. & Turk-Browne, N. B. Noise correlations in the human brain and their impact on pattern classification. *PLoS Comput. Biol.* **13**, e1005674 (2017).
18. Liu, T. T. Noise contributions to the fMRI signal: An overview. *Neuroimage* **143**, 141–151 (2016).
19. Peltier, S. J. *Characterization and Compensation of Systematic Noise in Functional Magnetic Resonance Imaging*. (University of Michigan., 2003).
20. Haxby, J. V. *et al.* Distributed and overlapping representations of faces and objects

in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
21. Haynes, J.-D. & Rees, G. Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* **7**, 523–534 (2006).
22. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10**, 424–430 (2006).
23. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8619–8624 (2014).
24. Chen, X. *et al.* InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. in *Advances in Neural Information Processing Systems* (eds. Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R.) vol. 29 (Curran Associates, Inc., 2016).
25. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **10**, 1299–1319 (1998).
26. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572 (1901).
27. Eckart, C. & Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1**, 211–218 (1936).
28. Braun, M. L., Buhmann, J. M. & Müller, K.-R. On relevant dimensions in kernel feature spaces. *J. Mach. Learn. Res.* **9**, 1875–1908 (2008).
29. Cox, D. D. & Savoy, R. L. Functional magnetic resonance imaging (fMRI) 'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* vol. 19 261–270 Preprint at https://doi.org/10.1016/s1053-8119(03)00049-1 (2003).
30. Wasikowski, M. & Chen, X.-W. Combating the Small Sample Class Imbalance Problem Using Feature Selection. *IEEE Trans. Knowl. Data Eng.* **22**, 1388–1400 (2010).
31. He, H. & Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
32. Nie, D., Zhang, H., Adeli, E., Liu, L. & Shen, D. 3D Deep Learning for Multi-modal Imaging-Guided Survival Time Prediction of Brain Tumor Patients. *Med. Image Comput. Comput. Assist. Interv.* **9901**, 212–220 (2016).
33. Socher, R., Pennington, J., Huang, E. H., Ng, A. Y. & Manning, C. D. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* 151–161 (Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011).
34. Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D. & Li, W. Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation. in *Computer Vision – ECCV 2016* 597–613 (Springer International Publishing, 2016).
35. Li, X. *et al.* Intelligent cross-machine fault diagnosis approach with deep auto-encoder and domain adaptation. *Neurocomputing* **383**, 235–247 (2020).
36. Hosoya, H. CIGMO: Learning categorical invariant deep generative models from grouped data. (2020).
37. LeCun, Y., Cortes, C. & Burges, C. MNIST handwritten digit database, 1998. *URL*

http://www. research. att. com/\ yann/ocr/mnist (1998).
38. Xiao, H., Rasul, K. & Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv [cs.LG]* (2017).
39. Shibata, K., Sasaki, Y., Kawato, M. & Watanabe, T. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Journal of Vision* vol. 12 282–282 Preprint at https://doi.org/10.1167/12.9.282 (2012).
40. Watanabe, T., Sasaki, Y., Shibata, K. & Kawato, M. Advances in fMRI Real-Time Neurofeedback. *Trends Cogn. Sci.* **21**, 997–1010 (2017).
41. Amano, K., Shibata, K., Kawato, M., Sasaki, Y. & Watanabe, T. Learning to Associate Orientation with Color in Early Visual Areas by Associative Decoded fMRI Neurofeedback. *Curr. Biol.* **26**, 1861–1866 (2016).
42. Koizumi, A. *et al.* Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nature Human Behaviour* **1**, 1–7 (2016).
43. Shibata, K. *et al.* Toward a comprehensive understanding of the neural mechanisms of decoded neurofeedback. *Neuroimage* **188**, 539–556 (2019).
44. LaConte, S. M., Peltier, S. J. & Hu, X. P. Real-time fMRI using brain-state classification. *Hum. Brain Mapp.* **28**, 1033–1044 (2007).
45. deCharms, R. C. *et al.* Learned regulation of spatially localized brain activation using real-time fMRI. *Neuroimage* **21**, 436–443 (2004).
46. Scharnowski, F., Hutton, C., Josephs, O., Weiskopf, N. & Rees, G. Improving visual perception through neurofeedback. *J. Neurosci.* **32**, 17830–17841 (2012).
47. Cortese, A., Amano, K., Koizumi, A., Kawato, M. & Lau, H. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat. Commun.* **7**, 13669 (2016).
48. Scheinost, D. *et al.* Orbitofrontal cortex neurofeedback produces lasting changes in contamination anxiety and resting-state connectivity. *Transl. Psychiatry* **3**, e250 (2013).
49. Taschereau-Dumouchel, V. *et al.* Towards an unconscious neural reinforcement intervention for common fears. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 3470–3475 (2018).
50. Shibata, K., Watanabe, T., Kawato, M. & Sasaki, Y. Differential Activation Patterns in the Same Brain Region Led to Opposite Emotional States. *PLoS Biol.* **14**, e1002546 (2016).
51. Cortese, A., Amano, K., Koizumi, A., Lau, H. & Kawato, M. Decoded fMRI neurofeedback can induce bidirectional confidence changes within single participants. *NeuroImage* vol. 149 323–337 Preprint at https://doi.org/10.1016/j.neuroimage.2017.01.069 (2017).
52. Caballero-Gaudes, C. & Reynolds, R. C. Methods for cleaning the BOLD fMRI signal. *Neuroimage* **154**, 128–149 (2017).
53. Shmueli, K. *et al.* Low-frequency fluctuations in the cardiac rate as a source of variance in the resting-state fMRI BOLD signal. *Neuroimage* **38**, 306–320 (2007).
54. Birn, R. M., Diamond, J. B., Smith, M. A. & Bandettini, P. A. Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *Neuroimage* **31**, 1536–1548 (2006).
55. Fullana, M. A. *et al.* Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry* vol. 21 500–508

Preprint at https://doi.org/10.1038/mp.2015.88 (2016).
56. Hofmann, S. G., Ellard, K. K. & Siegle, G. J. Neurobiological correlates of cognitions in fear and anxiety: A cognitive–neurobiological information-processing model. *Cognition and Emotion* **26**, 282–299 (2012).
57. Kingma, D. P., Mohamed, S., Rezende, D. J. & Welling, M. Semi-supervised learning with deep generative models. in *Advances in neural information processing systems* 3581–3589 (2014).
58. Maaløe, L., Sønderby, C. K., Sønderby, S. K. & Winther, O. Improving semi-supervised learning with auxiliary deep generative models. in *NIPS Workshop on Advances in Approximate Bayesian Inference* (2015).
59. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
60. Wang, S., Ding, Z. & Fu, Y. Coupled Marginalized Auto-Encoders for Cross-Domain Multi-View Learning. in *IJCAI* 2125–2131 (2016).
61. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. & Frey, B. Adversarial Autoencoders. *arXiv [cs.LG]* (2015).
62. Goodfellow, I. *et al.* Generative Adversarial Nets. in *Advances in Neural Information Processing Systems* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K. Q.) vol. 27 (Curran Associates, Inc., 2014).
63. Radford, A., Metz, L. & Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv [cs.LG]* (2015).
64. Wang, S., Ding, Z. & Fu, Y. Feature Selection Guided Auto-Encoder. *AAAI* **31**, (2017).
65. Mirza, M. & Osindero, S. Conditional Generative Adversarial Nets. *arXiv [cs.LG]* (2014).
66. Santurkar, S., Schmidt, L. & Mądry, A. A Classification-Based Study of Covariate Shift in GAN Distributions. *arXiv [cs.LG]* (2017).
67. Haxby, J. V. *et al.* A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).
68. Guntupalli, J. S. *et al.* A Model of Representational Spaces in Human Cortex. *Cereb. Cortex* bhw068–bhw068 (2016).
69. Busch, E. L. *et al.* Hybrid Hyperalignment: A single high-dimensional model of shared information embedded in cortical patterns of response and functional connectivity. *Cold Spring Harbor Laboratory* 2020.11.25.398883 (2020) doi:10.1101/2020.11.25.398883.
70. Huang, J. *et al.* Learning shared neural manifolds from multi-subject FMRI data. *arXiv [q-bio.NC]* (2021).
71. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **23**, (2020).
72. Ng, W. W. Y., Zeng, G., Zhang, J., Yeung, D. S. & Pedrycz, W. Dual autoencoders features for imbalance classification problem. *Pattern Recognit.* **60**, 875–889 (2016).
73. Zhao, S., Zhao, X., Ding, G. & Keutzer, K. EmotionGAN: Unsupervised Domain Adaptation for Learning Discrete Probability Distributions of Image Emotions. in *Proceedings of the 26th ACM international conference on Multimedia* 1319–1327 (Association for Computing Machinery, New York, NY, USA, 2018).

74. Torralba, A. & Efros, A. A. Unbiased look at dataset bias. in *CVPR 2011* 1521–1528 (2011).
75. Duan, L., Xu, D. & Tsang, I. Learning with Augmented Features for Heterogeneous Domain Adaptation. *arXiv [cs.LG]* (2012).
76. Harel, M. & Mannor, S. Learning from Multiple Outlooks. *arXiv [cs.LG]* (2010).
77. Prettenhofer, P. & Stein, B. Cross-Language Text Classification Using Structural Correspondence Learning. in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* 1118–1127 (Association for Computational Linguistics, Uppsala, Sweden, 2010).
78. Zhou, J., Pan, S., Tsang, I. & Yan, Y. Hybrid Heterogeneous Transfer Learning through Deep Learning. *AAAI* **28**, (2014).
79. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
80. Ling, W. *et al.* Batch effects removal for microbiome data via conditional quantile regression. *Nat. Commun.* **13**, 5418 (2022).
81. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
82. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
83. Zhang, Y. *et al.* Collaborative Unsupervised Domain Adaptation for Medical Image Diagnosis. *IEEE Trans. Image Process.* **29**, 7834–7844 (2020).
84. Chen, C., Dou, Q., Chen, H., Qin, J. & Heng, P.-A. Synergistic Image and Feature Adaptation: Towards Cross-Modality Domain Adaptation for Medical Image Segmentation. *AAAI* **33**, 865–872 (2019).
85. Lan, K. *et al.* A Survey of Data Mining and Deep Learning in Bioinformatics. *J. Med. Syst.* **42**, 139 (2018).
86. Shastry, K. A. & Sanjay, H. A. Machine Learning for Bioinformatics. in *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications* (eds. Srinivasa, K. G., Siddesh, G. M. & Manisekhar, S. R.) 25–39 (Springer Singapore, Singapore, 2020).
87. Berger, B., Daniels, N. M. & Yu, Y. W. Computational Biology in the 21st Century: Scaling with Compressive Algorithms. *Commun. ACM* **59**, 72–80 (2016).
88. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
89. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
90. Schloss, P. D. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *MBio* **9**, (2018).
91. Xu, C. & Jackson, S. A. Machine learning and complex biological data. *Genome Biol.* **20**, 76 (2019).
92. Bzdok, D., Altman, N. & Krzywinski, M. Statistics versus machine learning. *Nat. Methods* **15**, 233–234 (2018).
93. Herndon, N. & Caragea, D. Naıve bayes domain adaptation for biological sequences. in *Proceedings of the 4th International Conference on Bioinformatics Models, Methods and Algorithms, BIOINFORMATICS* 62–70 (2013).
94. Mourragui, S., Loog, M., van de Wiel, M. A., Reinders, M. J. T. & Wessels, L. F. A.

PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics* **35**, i510–i519 (2019).
95. Handl, L., Jalali, A., Scherer, M., Eggeling, R. & Pfeifer, N. Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data. *Bioinformatics* **35**, i154–i163 (2019).
96. Wang, M. *et al.* Identifying Autism Spectrum Disorder With Multi-Site fMRI via Low-Rank Domain Adaptation. *IEEE Transactions on Medical Imaging* vol. 39 644–655 Preprint at https://doi.org/10.1109/tmi.2019.2933160 (2020).
97. Mensch, A., Mairal, J., Bzdok, D., Thirion, B. & Varoquaux, G. Learning Neural Representations of Human Cognition across Many fMRI Studies. *arXiv [stat.ML]* (2017).
98. Zhang, H., Chen, P.-H. & Ramadge, P. Transfer Learning on fMRI Datasets. in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (eds. Storkey, A. & Perez-Cruz, F.) vol. 84 595–603 (PMLR, 2018).
99. Huang, C. *et al.* Meta-analysis reveals the vaginal microbiome is a better predictor of earlier than later preterm birth. *BMC Biol.* **21**, 199 (2023).
100. Golob, J. L. *et al.* Microbiome preterm birth DREAM challenge: Crowdsourcing machine learning approaches to advance preterm birth research. *Cell Rep Med* **5**, 101350 (2024).
101. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
102. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
103. DeGrave, A. J., Janizek, J. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **3**, 610–619 (2021).
104. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
105. Liu, P., Qiu, X. & Huang, X. Adversarial Multi-task Learning for Text Classification. *arXiv [cs.CL]* (2017).
106. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
107. Venkateswara, H., Eusebio, J., Chakraborty, S. & Panchanathan, S. Deep hashing network for unsupervised domain adaptation. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 5018–5027 (2017).
108. Peng, X. *et al.* Moment matching for multi-source domain adaptation. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 1406–1415 (2019).
109. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
110. Torralba, A., Fergus, R. & Freeman, W. T. 80 million tiny images: a large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 1958–1970 (2008).
111. Krizhevsky, A., Hinton, G. & Others. Learning multiple layers of features from tiny images. (2009).
112. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nat. Methods* **15**, 399–

400 (2018).
113. Turner, B. O., Paul, E. J., Miller, M. B. & Barbey, A. K. Small sample sizes reduce the replicability of task-based fMRI studies. *Commun Biol* **1**, 62 (2018).
114. Zhou, S., Cox, C. R. & Lu, H. Improving Whole-Brain Neural Decoding of fMRI with Domain Adaptation. in *Machine Learning in Medical Imaging* 265–273 (Springer International Publishing, 2019).
115. Saenko, K., Kulis, B., Fritz, M. & Darrell, T. Adapting Visual Category Models to New Domains. in *Computer Vision – ECCV 2010* 213–226 (Springer Berlin Heidelberg, 2010).
116. Long, M., Zhu, H., Wang, J. & Jordan, M. I. Deep Transfer Learning with Joint Adaptation Networks. *arXiv [cs.LG]* (2016).
117. Long, M., Cao, Y., Wang, J. & Jordan, M. Learning Transferable Features with Deep Adaptation Networks. in *Proceedings of the 32nd International Conference on Machine Learning* (eds. Bach, F. & Blei, D.) vol. 37 97–105 (PMLR, Lille, France, 2015).
118. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K. & Darrell, T. Deep Domain Confusion: Maximizing for Domain Invariance. *arXiv [cs.CV]* (2014).
119. Shavitt, I. & Segal, E. Regularization Learning Networks: Deep Learning for Tabular Datasets. *arXiv [stat.ML]* (2018).
120. Arik, S. Ö. & Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *AAAI* **35**, 6679–6687 (2021).
121. McElfresh, D. *et al.* When Do Neural Nets Outperform Boosted Trees on Tabular Data? *Adv. Neural Inf. Process. Syst.* **36**, 76336–76369 (2023).
122. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? in *Advances in Neural Information Processing Systems* (eds. Koyejo, S. et al.) vol. 35 507–520 (Curran Associates, Inc., 2022).
123. Yang, Y. & Soatto, S. FDA: Fourier domain adaptation for semantic segmentation. in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 4085–4095 (IEEE, 2020).
124. Hong, W., Wang, Z., Yang, M. & Yuan, J. Conditional generative adversarial network for structured domain adaptation. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1335–1344 (IEEE, 2018).
125. Motiian, S., Jones, Q., Iranmanesh, S. & Doretto, G. Few-shot adversarial domain adaptation. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
126. Sohn, K. *et al.* Unsupervised Domain Adaptation for Face Recognition in Unlabeled Videos. *arXiv [cs.CV]* (2017).
127. Ghosh-Dastidar, B. & Schafer, J. L. Multiple Edit/Multiple Imputation for Multivariate Continuous Data. *J. Am. Stat. Assoc.* **98**, 807–817 (2003).
128. Eisemann, N., Waldmann, A. & Katalinic, A. Imputation of missing values of tumour stage in population-based cancer registration. *BMC Med. Res. Methodol.* **11**, 129 (2011).
129. van Dijk, D. *et al.* MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv* 111591 (2017) doi:10.1101/111591.
130. Zitnik, M. *et al.* Machine Learning for Integrating Data in Biology and Medicine:

Principles, Practice, and Opportunities. *Inf. Fusion* **50**, 71–91 (2019).
131. Marmarelis, V. Z. Identification of nonlinear biological systems using Laguerre expansions of kernels. *Ann. Biomed. Eng.* **21**, 573–589 (1993).
132. Singh, D., Climente-Gonzalez, H., Petrovich, M., Kawakami, E. & Yamada, M. FsNet: Feature Selection Network on High-dimensional Biological Data. in *2023 International Joint Conference on Neural Networks (IJCNN)* 1–9 (IEEE, 2023).
133. Pan, A. Y. Statistical analysis of microbiome data: The challenge of sparsity. *Current Opinion in Endocrine and Metabolic Research* **19**, 35–40 (2021).
134. Zhou, T., Liu, M., Thung, K.-H. & Shen, D. Latent Representation Learning for Alzheimer's Disease Diagnosis With Incomplete Multi-Modality Neuroimaging and Genetic Data. *IEEE Trans. Med. Imaging* **38**, 2411–2422 (2019).
135. Samartsidis, P. *et al.* Estimating the prevalence of missing experiments in a neuroimaging meta-analysis. *Res. Synth. Methods* **11**, 866–883 (2020).
136. Zhou, T. *et al.* Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data. *Med. Image Anal.* **60**, 101630 (2020).
137. Wei, P., Ke, Y. & Goh, C. K. A General Domain Specific Feature Transfer Framework for Hybrid Domain Adaptation. *IEEE Trans. Knowl. Data Eng.* **31**, 1440–1451 (2019).
138. Wang, C. & Mahadevan, S. Heterogeneous Domain Adaptation using Manifold Alignment. in *Twenty-Second International Joint Conference on Artificial Intelligence* (2011).
139. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
140. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, e00191–16 (2017).
141. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
142. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
143. Cox, R. W. & Hyde, J. S. Software tools for analysis and visualization of fMRI data. *NMR Biomed.* **10**, 171–178 (1997).
144. Woolrich, M. W. *et al.* Bayesian analysis of neuroimaging data in FSL. *Neuroimage* **45**, S173–86 (2009).
145. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23 Suppl 1**, S208–19 (2004).
146. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *Neuroimage* **62**, 782–790 (2012).
147. Andronache, A. *et al.* Impact of functional MRI data preprocessing pipeline on default-mode network detectability in patients with disorders of consciousness. *Front. Neuroinform.* **7**, 16 (2013).
148. Cammarota, G. *et al.* Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 635–648 (2020).
149. Kim, Y.-M., Poline, J.-B. & Dumas, G. Experimenting with reproducibility: a case study of robustness in bioinformatics. *Gigascience* **7**, (2018).
150. Liu, X. *et al.* Domain Adaptation via Low Rank and Class Discriminative

Representation for Autism Spectrum Disorder identification: A Multi-site fMRI Study. *IEEE Trans. Neural Syst. Rehabil. Eng.* **PP**, (2023).
151. Pominova, M., Kondrateva, E., Sharaev, M., Bernstein, A. & Burnaev, E. Fader networks for domain adaptation on fMRI: ABIDE-II study. in *Thirteenth International Conference on Machine Vision* vol. 11605 570–577 (SPIE, 2021).
152. Liu, X. & Huang, H. Alterations of functional connectivities associated with autism spectrum disorder symptom severity: a multi-site study using multivariate pattern analysis. *Sci. Rep.* **10**, 4330 (2020).
153. van Opbroek, A., Ikram, M. A., Vernooij, M. W. & de Bruijne, M. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Trans. Med. Imaging* **34**, 1018–1030 (2015).
154. Kouw, W. M., Loog, M., Bartels, L. W. & Mendrik, A. M. MR Acquisition-Invariant Representation Learning. *arXiv [cs.CV]* (2017).
155. Yang, X., Deng, C., Liu, T. & Tao, D. Heterogeneous Graph Attention Network for Unsupervised Multiple-Target Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1992–2003 (2022).
156. Liu, F., Zhang, G. & Lu, J. Heterogeneous Domain Adaptation: An Unsupervised Approach. *IEEE Trans Neural Netw Learn Syst* **31**, 5588–5602 (2020).
157. Ganin, Y. *et al.* Domain-adversarial training of neural networks. in *Domain Adaptation in Computer Vision Applications* 189–209 (Springer International Publishing, Cham, 2017).
158. Vinuesa, R. & Sirmacek, B. Interpretable deep-learning models to help achieve the Sustainable Development Goals. *Nature Machine Intelligence* **3**, 926–926 (2021).
159. Arpit, D. *et al.* A Closer Look at Memorization in Deep Networks. in *Proceedings of the 34th International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) vol. 70 233–242 (PMLR, 06--11 Aug 2017).
160. Koh, P. W. & Liang, P. Understanding Black-box Predictions via Influence Functions. in *Proceedings of the 34th International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) vol. 70 1885–1894 (PMLR, 06--11 Aug 2017).
161. Mehra, A., Kailkhura, B., Chen, P.-Y. & Hamm, J. Understanding the limits of unsupervised domain adaptation via data poisoning. *Adv. Neural Inf. Process. Syst.* 17347–17359 (2021).
162. Ben-David, S., Lu, T., Luu, T. & Pál, D. Impossibility theorems for domain adaptation. *AISTATS* 129–136 (2010).
163. Redko, I., Habrard, A. & Sebban, M. On the analysis of adaptability in multi-source domain adaptation. *Mach. Learn.* **108**, 1635–1652 (2019).
164. Liu, H., Long, M., Wang, J. & Jordan, M. Transferable Adversarial Training: A General Approach to Adapting Deep Classifiers. in *Proceedings of the 36th International Conference on Machine Learning* (eds. Chaudhuri, K. & Salakhutdinov, R.) vol. 97 4013–4022 (PMLR, 09--15 Jun 2019).
165. Wang, Z., Dai, Z., Póczos, B. & Carbonell, J. Characterizing and Avoiding Negative Transfer. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11285–11294 (IEEE, 2019).
166. Chan, Y. S. & Ng, H. T. Word sense disambiguation with distribution estimation. https://www.ijcai.org/Proceedings/05/Papers/1543.pdf.

167. Kouw, W. M. & Loog, M. An introduction to domain adaptation and transfer learning. *arXiv [cs.LG]* (2018).
168. Wang, M. & Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* (2018).
169. Csurka, G. Domain Adaptation for Visual Applications: A Comprehensive Survey. *arXiv [cs.CV]* (2017).
170. Liu, X. *et al.* Deep Unsupervised Domain Adaptation: A Review of Recent Advances and Perspectives. *APSIPA Transactions on Signal and Information Processing* **11**, (2022).
171. Fernando, B., Habrard, A., Sebban, M. & Tuytelaars, T. Unsupervised visual domain adaptation using subspace alignment. in *Proceedings of the IEEE international conference on computer vision* 2960–2967 (2013).
172. Gong, B., Shi, Y., Sha, F. & Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. in *2012 IEEE Conference on Computer Vision and Pattern Recognition* 2066–2073 (ieeexplore.ieee.org, 2012).
173. Long, M., Cao, Y., Wang, J. & Jordan, M. Learning Transferable Features with Deep Adaptation Networks. in *Proceedings of the 32nd International Conference on Machine Learning* (eds. Bach, F. & Blei, D.) vol. 37 97–105 (PMLR, Lille, France, 2015).
174. Sun, B. & Saenko, K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. in *Computer Vision – ECCV 2016 Workshops* 443–450 (Springer International Publishing, 2016).
175. Bhatt, H. S., Rajkumar, A. & Roy, S. Multi-Source Iterative Adaptation for Cross-Domain Classification. https://www.ijcai.org/Proceedings/16/Papers/519.pdf.
176. Matsuura, T. & Harada, T. Domain Generalization Using a Mixture of Multiple Latent Domains. *AAAI* **34**, 11749–11756 (2020).
177. Montesuma, E. F. & Mboula, F. M. N. Wasserstein Barycenter for Multi-Source Domain Adaptation. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 16785–16793 (IEEE, 2021).
178. Mansour, Y., Mohri, M. & Rostamizadeh, A. Domain Adaptation with Multiple Sources. *Adv. Neural Inf. Process. Syst.* **21**, (2008).
179. Zhu, Y., Zhuang, F. & Wang, D. Aligning Domain-Specific Distribution and Classifier for Cross-Domain Classification from Multiple Sources. *AAAI* **33**, 5989–5996 (2019).
180. Xu, R., Chen, Z., Zuo, W., Yan, J. & Lin, L. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3964–3973 (2018).
181. Zhao, H. *et al.* Adversarial multiple source domain adaptation. *Adv. Neural Inf. Process. Syst.* **31**, 8559–8570 (2018).
182. Guan, H. & Liu, M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans. Biomed. Eng.* **69**, 1173–1185 (2022).
183. Daumé, H., III. Frustratingly Easy Domain Adaptation. *arXiv [cs.LG]* (2009).
184. Saito, K., Ushiku, Y. & Harada, T. Asymmetric Tri-training for Unsupervised Domain Adaptation. in *Proceedings of the 34th International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) vol. 70 2988–2997 (PMLR, 06--11 Aug 2017).

185. Shrivastava, A. *et al.* Learning from Simulated and Unsupervised images through adversarial training. *arXiv [cs.CV]* 2107–2116 (2016).
186. Zhuo, J., Wang, S., Zhang, W. & Huang, Q. Deep Unsupervised Convolutional Domain Adaptation. in *Proceedings of the 25th ACM international conference on Multimedia* 261–269 (Association for Computing Machinery, New York, NY, USA, 2017).
187. Wen Li, Lixin Duan, Dong Xu & Tsang, I. W. Learning With Augmented Features for Supervised and Semi-Supervised Heterogeneous Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 1134–1148 (2014).
188. Tzeng, E., Hoffman, J., Darrell, T. & Saenko, K. Simultaneous Deep Transfer Across Domains and Tasks. *2015 IEEE International Conference on Computer Vision (ICCV)* Preprint at https://doi.org/10.1109/iccv.2015.463 (2015).
189. Saito, K., Kim, D., Sclaroff, S., Darrell, T. & Saenko, K. Semi-supervised domain adaptation via minimax entropy. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 8050–8058 (IEEE, 2019).
190. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* vol. 90 227–244 Preprint at https://doi.org/10.1016/s0378-3758(00)00115-4 (2000).
191. Yang, J., Yan, R. & Hauptmann, A. G. Cross-domain video concept detection using adaptive svms. in *Proceedings of the 15th ACM international conference on Multimedia* 188–197 (Association for Computing Machinery, New York, NY, USA, 2007).
192. Duan, L., Tsang, I. W., Xu, D. & Chua, T.-S. Domain adaptation from multiple sources via auxiliary classifiers. in *Proceedings of the 26th Annual International Conference on Machine Learning* 289–296 (Association for Computing Machinery, New York, NY, USA, 2009).
193. Duan, L., Xu, D. & Chang, S.-F. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. in *2012 IEEE Conference on Computer Vision and Pattern Recognition* 1338–1345 (2012).
194. Zhou, J. T., W.Tsang, I., Pan, S. J. & Tan, M. Heterogeneous Domain Adaptation for Multiple Classes. in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (eds. Kaski, S. & Corander, J.) vol. 33 1095–1103 (PMLR, Reykjavik, Iceland, 2014).
195. Sun, B., Feng, J. & Saenko, K. Return of Frustratingly Easy Domain Adaptation. *AAAI* **30**, (2016).
196. Guo, J., Shah, D. & Barzilay, R. Multi-Source Domain Adaptation with Mixture of Experts. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* Preprint at https://doi.org/10.18653/v1/d18-1498 (2018).
197. Mancini, M., Porzi, L., Bulo, S. R., Caputo, B. & Ricci, E. Boosting domain adaptation by discovering latent domains. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3771–3780 (IEEE, 2018).
198. Zhao, S. *et al.* Multi-Source Distilling Domain Adaptation. *AAAI* **34**, 12975–12983 (2020).
199. Austin, G. I. *et al.* Processing-bias correction with DEBIAS-M improves cross-study generalization of microbiome-based prediction models. *bioRxiv* (2024) doi:10.1101/2024.02.09.579716.

200. Sharifi-Noghabi, H., Peng, S., Zolotareva, O., Collins, C. C. & Ester, M. AITL: Adversarial Inductive Transfer Learning with input and output space adaptation for pharmacogenomics. *Bioinformatics* **36**, i380–i388 (2020).
201. Di Martino, A. *et al.* The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).
202. Masoudnia, S. & Ebrahimpour, R. Mixture of experts: a literature survey. *Artificial Intelligence Review* **42**, 275–293 (2014).
203. Shazeer, N. *et al.* Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv [cs.LG]* (2017).
204. Gao, Y., Zhang, Y., Cao, Z., Guo, X. & Zhang, J. Decoding Brain States From fMRI Signals by Using Unsupervised Domain Adaptation. *IEEE J Biomed Health Inform* **24**, 1677–1685 (2020).
205. Van Essen, D. C. *et al.* The Human Connectome Project: a data acquisition perspective. *Neuroimage* **62**, 2222–2231 (2012).
206. Hou, L. Regularizing label-augmented generative adversarial networks under limited data. *IEEE Access* **11**, 28966–28976 (2023).
207. Webster, R., Rabin, J., Simon, L. & Jurie, F. Detecting overfitting of deep generative networks via latent recovery. *arXiv [cs.LG]* 11273–11282 (2019).
208. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798 (2018).
209. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
210. Amir, A., Ozel, E., Haberman, Y. & Shental, N. Achieving pan-microbiome biological insights via the dbBact knowledge base. *Nucleic Acids Res.* **51**, 6593–6608 (2023).
211. Abdill, R. J. *et al.* Integration of 168,000 samples reveals global patterns of the human gut microbiome. *bioRxiv* (2023) doi:10.1101/2023.10.11.560955.
212. van den Heuvel, M. P., Mandl, R. C. W., Kahn, R. S. & Hulshoff Pol, H. E. Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Hum. Brain Mapp.* **30**, 3127–3141 (2009).
213. Fallani, F. & Babiloni, F. *The Graph Theoretical Approach in Brain Functional Networks: Theory and Applications*. (Springer Nature, 2022).
214. Gershenson, C., Aerts, D. & Edmonds, B. *Worldviews, Science and Us: Philosophy and Complexity : University of Liverpool, UK, 11-14 September 2005*. (World Scientific, 2007).
215. McClurkin, J. W., Optican, L. M., Richmond, B. J. & Gawne, T. J. Concurrent processing and complexity of temporally encoded neuronal messages in visual perception. *Science* **253**, 675–677 (1991).
216. *A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex*. (Massachusetts inst of tech cambridge ma center for biological and computational learning, 2005).
217. Libedinsky, C. & Livingstone, M. Role of prefrontal cortex in conscious visual perception. *J. Neurosci.* **31**, 64–69 (2011).
218. Eriksson, J., Larsson, A., Ahlström, K. R. & Nyberg, L. Similar frontal and distinct

posterior cortical regions mediate visual and auditory perceptual awareness. *Cereb. Cortex* **17**, 760–765 (2007).
219. Kar, K. & DiCarlo, J. J. Fast recurrent processing via ventral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *bioRxiv* 2020.05.10.086959 (2020) doi:10.1101/2020.05.10.086959.
220. Intaitė, M., Noreika, V., Šoliūnas, A. & Falter, C. M. Interaction of bottom-up and top-down processes in the perception of ambiguous figures. *Vision Res.* **89**, 24–31 (2013).
221. Magosso, E. & Ursino, M. The Sensory-Cognitive Interplay: Insights into Neural Mechanisms and Circuits. *J. Integr. Neurosci.* **22**, 3 (2022).
222. Gazzaley, A. Top-down modulation: the crossroads of perception, attention and memory. in *Human Vision and Electronic Imaging XV* vol. 7527 78–88 (SPIE, 2010).
223. Yan, Y., Zhan, J., Ince, R. A. A. & Schyns, P. G. Top-down predictions of visual features dynamically reverse their bottom-up processing in the occipito-ventral pathway to facilitate stimulus disambiguation and behavior. *bioRxiv* 2021.10.12.464078 (2021) doi:10.1101/2021.10.12.464078.
224. Alkhasli, I., Mottaghy, F. M., Binkofski, F. & Sakreida, K. Preconditioning prefrontal connectivity using transcranial direct current stimulation and transcranial magnetic stimulation. *Front. Hum. Neurosci.* **16**, 929917 (2022).
225. Ghifary, M., Kleijn, W. B., Zhang, M. & Balduzzi, D. Domain generalization for object recognition with Multi-task autoencoders. *arXiv [cs.CV]* 2551–2559 (2015).
226. Zhuang, F., Cheng, X., Luo, P., Pan, S. J. & He, Q. Supervised representation learning: Transfer learning with deep autoencoders. *Int Jt Conf Artif Intell* 4119–4125 (2015).
227. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D. & Erhan, D. Domain Separation Networks. in *Advances in Neural Information Processing Systems* (eds. Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R.) vol. 29 (Curran Associates, Inc., 2016).
228. Glorot, X., Bordes, A. & Bengio, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. *ICML* 513–520 (2011).
229. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).
230. Hebart, M. N. *et al.* THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife* **12**, (2023).
231. Thorat, S., Proklova, D. & Peelen, M. V. The nature of the animacy organization in human ventral temporal cortex. *Elife* **8**, (2019).