Equivariance Allows Handling Multiple Nuisance Variables When Analyzing Pooled Neuroimaging Datasets

Vishnu Suresh Lokhande lokhande@cs.wisc.edu
Rudrasis Chakraborty rudrasischa@gmail.com
Sathya N. Ravi sathya@uic.edu
Vikas Singh vsingh@biostat.wisc.edu

Abstract

Pooling multiple neuroimaging datasets across institutions often enables improvements in statistical power when evaluating associations (e.g., between risk factors and disease outcomes) that may otherwise be too weak to detect. When there is only a single *source of variability (e.g., different scanners), domain adaptation and matching the distributions of representations may suffice in many scenarios. But in the presence of* more than one *nuisance variable which concurrently influence the measurements, pooling datasets poses unique challenges, e.g., variations in the data can come from both the acquisition method as well as the demographics of participants (gender, age). Invariant representation learning, by itself, is ill-suited to fully model the data generation process. In this paper, we show how bringing recent results on equivariant representation learning (for studying symmetries in neural networks) instantiated on structured spaces together with simple use of classical results on causal inference provides an effective practical solution. In particular, we demonstrate how our model allows dealing with more than one nuisance variable under some assumptions and can enable analysis of pooled scientific datasets in scenarios that would otherwise entail removing a large portion of the samples. Our code is available on* https://github.com/vsingh-group/DatasetPooling

1. Introduction

Observational studies in many disciplines acquire cross-sectional/longitudinal clinical and imaging data to understand diseases such as neurodegeneration and dementia [44]. Typically, these studies are sufficiently powered for the primary scientific hypotheses of interest. However, *secondary* analyses to investigate weaker but potentially interesting associations between risk factors (such as genetics) and disease outcomes are often difficult when using common statistical significance thresholds, due to the small-/medium sample sizes.

Over the last decade, there are coordinated large scale multi-institutional imaging studies (e.g., ADNI [26], NIH All of Us and HCP [19]) but the types of data collected or the project's scope (e.g., demographic pool of participants) may not be suited for studying specific *secondary* scientific questions. A "pooled" imaging dataset obtained from combining roughly similar studies across different institutions/sites, when possible, is an attractive alternative. The pooled datasets provide much larger sample sizes and improved statistical power to identify early disease biomarkers – analyses which would not otherwise be possible [16,30]. But even when study participants are consistent across sites, pooling poses challenges. This is true even for linear regression [56] – improvement in statistical power is not always guaranteed. Partly due to these as well as other reasons, high visibility projects such as ENIGMA [47] have reported findings using meta-analysis methods.

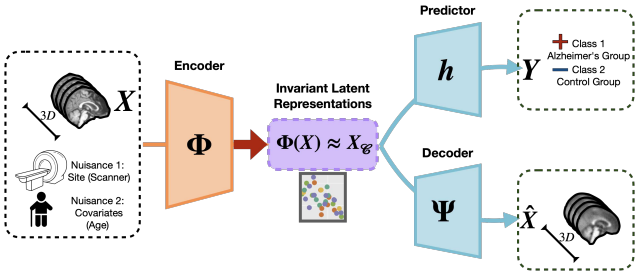
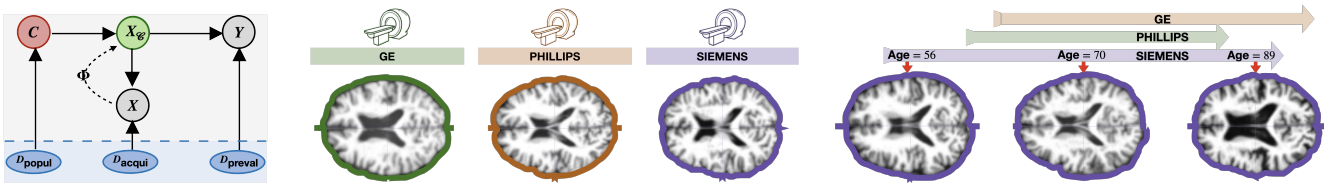


Figure 1. **Learning Invariant Representations.** In our framework, input images X are pooled together from multiple sites. An encoder Φ maps X to the latent representations $\Phi(X)$ that corresponds to high-level causal features $X_{\mathscr{C}}$ that influences the label prediction. Unlike the input images X, $\Phi(X)$ is robust to nuisance attributes like site (scanner) and covariates (age). Φ is trained alongside predictor h and decoder Ψ.

Data pooling and fairness. Even under ideal conditions, pooling imaging datasets across sites requires care. Assume that the participants across two sites, say site_1 and site_2, are perfectly gender matched with the same proportion of male/female and the age distribution (as well as the proportion of diseased/health controls) is also identical. In this idealized setting, the only difference between sites may come from variations in scanners or acquisition (e.g., pulse sequences). When training modern neural networks for a regression/classification task with imaging data obtained in this scenario, we may ask that the representations learned by



(a) Causal Diagram (b) Variation due to site (scanner) for particular age group. (c) Variation due to covariates (age) in Siemens scanner.

Figure 2. **(a)** A **Causal Diagram** listing variable of interest and their relationship for multi-site pooling problem. Nodes D_{popul}, D_{acqui} and D_{preval} denote the population, acquisition and prevalence biases that vary across sites. C's are covariates (like age or gender). $X_{\mathscr{C}}$ denotes the high-level causal features of an image X that influences the labels Y. Nodes in red d-separate the nodes in blue and green. **(b)** MRI images on control subjects from the ADNI [26] dataset for different **scanners** in the age group 70-80. **(c)** Images obtained from the Siemens scanner (i.e., fixing site) on control subjects for three extreme **age groups**. The gantt chart on top of the image indicates the respective age range in the Phillips and GE scanners. As observed, different scanner groups do not share a common support on "age" covariates, resulting in samples outside of the common support to be discarded in naïve pooling approaches.

the model be *invariant* to the categorical variable denoting "site". While this is not a "solved" problem, this strategy has been successfully deployed based on results in invariant representation learning [3, 5, 34] (see Fig. 1). One may alternatively view this task via the lens of fairness – we want the model's performance to be fair with respect to the site variable. This approach is effective, via constraints [52] or using adversarial modules [17, 53]. This setting also permits re-purposing tools from domain adaptation [35, 50, 55] or transfer learning [12] as a pre-processing step, before analysis of the pooled data proceeds.

Nuisance variables/confounds. Data pooling problems one often encounters in scientific research typically violates many of the conditions in the aforementioned example. The measured data X at each site is influenced not only by the scanner properties but also by a number of other covariates / nuisance variables. For instance, if the age distribution of participants is not identical across sites, comparison of the site-wise distributions is challenging because it is influenced **both** by age and the scanner. An example of the differences introduced due to age and scanner biases is shown in Figures 2b, 2c. With multiple nuisance variables, even effective tools for invariant representation learning, when used directly, can provide limited help. The data generation process, and the role of covariates/nuisance variables, available via a causal diagram (Figure 2a), can inform how the formulation is designed [6, 45]. Indeed, concepts from causality have benefited various deep learning models [37, 41]. Specially, recent work [31] has shown the value of integrating structural causal models for domain generalization, which is related to dataset pooling.

Causal Diagram. Dataset pooling under completely arbitrary settings is challenging to study systematically. So, we assume that the site-specific imaging datasets are *not* significantly different to begin with, although the distributions for covariates such as age/disease prevalence may not be perfectly matched and each of these factors will influence the data. We assume access to a causal diagram describing how these variables influence the measurements. We show how the distribution matching criteria provided by a causal diagram can be nicely handled for some ordinal covariates that are not perfectly matched across sites by adapting ideas from equivariant representation learning.

Contributions. We propose a method to pool multiple neuroimaging datasets by learning representations that are robust to site (scanner) and covariate (age) values (see Fig. 1 for visualization). We show that continuous nuisance covariates which do not have the same support and are not identically distributed across sites, can be effectively handled when learning invariant representations. We do not require finding "closest match" participants across sites – a strategy loosely based on covariate matching [39] from statistics which is less feasible if the distributions for a covariate (e.g., age) do not closely overlap. Our model is based on adapting recent results on equivariance together with known concepts from group theory. When tied with common invariant representation learners, our formulation allows far improved analysis of pooled imaging datasets. We first perform evaluations on common fairness datasets and then show its applicability on two separate neuroimaging tasks with multiple nuisance variables.

2. Reducing Multi-site Pooling to Infinite Dimensional Optimization

Let X denote an image of a participant and let Y be the corresponding (continuous or discrete) response variable or target label (such as cognitive score or disease status). For simplicity, consider only two sites – site_1 and site_2. Let D represent the site-specific shifts, biases or covariates that we want to take into account. One possible data generation process relating these variables is shown in Figure 2a.

Site-specific biases/confounds. Observe that Y is, in fact, influenced by high-level (or latent) features $X_{\mathscr{C}}$ specific to the participant. The images (or image-based disease biomarkers) X are simply our (lossy) measurement of the participant's brain $X_{\mathscr{C}}$ [14]. Further, X also includes an (unknown) confound: contribution from the scanner (or acquisition protocol). Figure 2a also lists covariates C, such as age and other factors which impact $X_{\mathscr{C}}$ (and therefore, X). A few common site-specific biases D are shown in

Fig. 2a. These include **(i)** *population bias* D_{popul} that leads to differences in age or gender distributions of the cohort [9]; **(ii)** we must also account for *acquisition shift* D_{acqui} resulting from different scanners or imaging protocols – this affects X but not $X_{\mathscr{C}}$; **(iii)** data are also influenced by a *class prevalence bias* D_{preval}, e.g., healthier individuals over-represented in site$_2$ will impact the distribution of cognitive scores across sites.

For imaging data, in principle, site-invariance can be achieved via an encoder-decoder style architecture to map the images X into a "site invariant" latent space $\Phi(X)$. Here, $\Phi(X)$ in the idealized setting, corresponds to the true "causal" features $X_{\mathscr{C}}$ that is comparable across sites. In practice, we know that images cannot fully capture the disease – so, $\Phi(X)$ is simply a surrogate, limited by the measurements we have on hand. Given these caveats, an architecture is shown in Fig. 1. Ideally, the encoder will minimize Maximum Mean Discrepancy (MMD) [20] or another discrepancy between the distributions of latent representations $\Phi(\cdot)$ of the data from site$_1$ and site$_2$.

The site-specific attributes D are often **unobserved** or otherwise unavailable. For instance, we may not have full access to D_{popul} from which our participants are drawn. To tackle these issues, we use a causal diagram, see Fig. 2a, similar to existing works [31,55] with minimal changes. For dealing with unobserved D's, some standard approaches are known [22]. Let us see how it can help here. Applying d-separation (see [22, 36]) on Fig. 2a, we see that the nodes $(D_{\text{popul}}, C, X_{\mathscr{C}})$ form a so-called "head-to-tail" branch and the nodes $(D_{\text{acqui}}, X, X_{\mathscr{C}})$, $(D_{\text{preval}}, Y, X_{\mathscr{C}})$ form a "head-to-head" branch. This implies that $X_{\mathscr{C}} \perp\!\!\!\perp D \mid C$. This is exactly an invariance condition: $X_{\mathscr{C}}$ should not change across different sites for samples with the **same value** of C. To enforce this using $\Phi(\cdot)$, we must optimize a discrepancy between site-wise $\Phi(X)$'s at a given value of C,

$$\min_{\Phi} \mathcal{MMD}\Big(P_{\text{site}_1}\Big(\Phi(X) \mid C\Big), P_{\text{site}_2}\Big(\Phi(X) \mid C\Big)\Big) \quad (1)$$

Provably solving (1)? A brief comment on the difficulty of the distributional optimization in (1) is useful. Generic tools for (worst case) convergence rates for such problems are actively being developed [51]. For the average case, [38] presents an online method for a specific class of (finite dimensional) distributionally robust optimization problems that can be defined using standard divergence measures. Observe that even these convergence guarantees are local in nature, i.e., they output a point that satisfies necessary conditions and may not be sufficient.

In practice, the outlook is a little better. Intuitively, an optimal *matching* of the conditional distributions $P(\Phi(X) \mid C)$ across the two sites corresponds to a (globally) optimal solution to the probabilistic optimization task in (1). Existing works show that it is indeed possible to approach this *computationally* via sub-sampling methods [55] or by learning elaborate matching functions to identify image or object pairs across sites that are "similar" [31] or have the same value for C. Sub-sampling, by definition, reduces the number of samples from the two sites by discarding samples outside of the common support. This impacts the quality of the estimator – for instance, [55] must restrict the analysis only to that age range of C which overlaps or is shared across sites. Such discarding of samples is clearly undesirable when each image acquisition is expensive. Matching functions also do not work if the support of C is not identical across the sites, as briefly described next.

Example 2.1. *Let C denote an observed covariate, e.g., age. Consider X_i at site$_1$ with $C = c_1$ and X_j at site$_2$ with $C = c_2$. If $c_1 \approx c_2$, a matching will seek $\Phi(X_i) \approx \Phi(X_j)$ in $X_{\mathscr{C}}$ space. If c_1 falls outside the support of c's acquired at site$_2$, one must not only estimate $\Phi(\cdot)$ but also a transport expression $\Gamma_{c_2 \to c_1}(\cdot)$ on $X_{\mathscr{C}}$ such that $\Phi(X_i) \approx \Gamma_{c_2 \to c_1}(\Phi(X_j))$. The "transportation" involves estimating what a latent image acquired at age c_2 would look like at age c_1. This means that matching would need a solution to the key difficulty, obtained further upstream.*

2.1. Improved Distribution Matching via Equivariant Mappings may be possible

Ignoring Y for the moment, recall that matching here corresponds to a bijection between unlabeled (finite) conditional distributions. Indeed, if the conditional distributions take specific forms such as a Poisson process, it is indeed possible to use simple matching algorithms that only require access to pairwise ranking information on the corresponding empirical distributions [43], for example, the well-known Gale-Shapley algorithm [46]. Unfortunately, in applications that we consider here, such distributional assumptions may not be fully faithful with respect to site specific covariates C. In essence, we want representations Φ (when viewed as a function of C) that vary in a predictable (or say deterministic) manner – if so, we can avoid matching altogether and instead match a suitable property of the site-wise distributions of the representation $\Phi(X)$. We can make this criterion more specific. We want the site-wise distributions to vary in a manner where the "trend" is consistent across the sites. Assume that this were not true, say $P(\Phi(X) \mid C)$ is continuous and monotonically increases with C for site$_1$ but monotonically decreases for site$_2$. A match of $P(\Phi(X) \mid C)$ across the sites at a particular value of $C = c$ implies at least one $C = c'$ where $P(\Phi(X) \mid C)$ do not match. The monotonicity argument is weak for high dimensional Φ. Plus, we have multiple nuisance variables. It turns out that our requirement for $P(\Phi(X) \mid C)$ to vary in a predictable manner across sites can be handled using the idea of equivariant mappings, i.e., $P(\Phi(X) \mid C)$ must

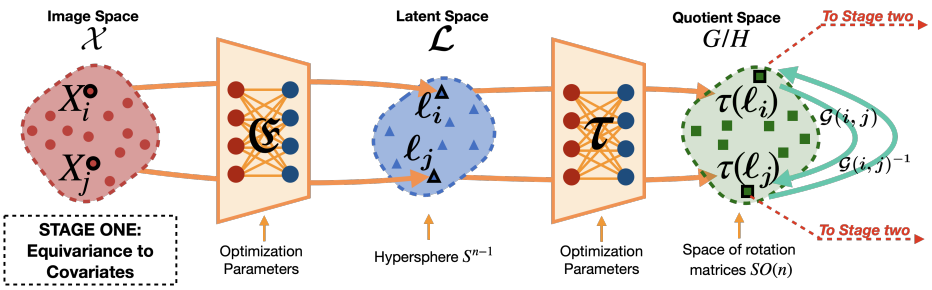


Figure 3. **Visualization of Stage one.** First, an image pair X_i, X_j are mapped onto a hypersphere using an encoder $\mathfrak{E}$. The resulting pair $\boldsymbol{\ell}_i, \boldsymbol{\ell}_j$ are passed through τ network to map them into the space of rotation matrices (which is the quotient group denoted by G/H). Fact 3 ensure that τ is a $G = SO(n)-$equivariant map. $\mathcal{G}(i,j)/\mathcal{G}(i,j)^{-1}$ is the group action of transforming $\tau(\boldsymbol{\ell}_i)$ to $\tau(\boldsymbol{\ell}_j)/\tau(\boldsymbol{\ell}_j)$ to $\tau(\boldsymbol{\ell}_i)$ respectively.

be **equivariant** with respect to C for both sites. In addition, we will also seek **invariance** to scanner attributes.

While we are familiar with the well-studied notion of invariance through the MMD criterion [29], we will briefly formalize our idea behind an equivariant mapping which is less common in this setting.

Definition 1. *A mapping $f : \mathcal{X} \to \mathcal{Y}$ defined over measurable Borel spaces $\mathcal{X}$ and $\mathcal{Y}$ is said to be $G-$equivariant under the action of group G iff*

$$f(g \cdot x) = g \cdot f(x), \quad g \in G$$

We refer the reader to two recent surveys, Section 2.1 of [7] and Section 3.1 of [8], which provide a detailed review.

Equivariance is often understood in the context of a group action (say, a matrix group) [24, 28]. While the covariates C is a vector (and every vector space is an abelian group), since this group will eventually act on the latent space of our images, imposing additional structure will be beneficial. To do so, we will utilize a mapping between C's and a group suitable for our setting. Once this is accomplished, we will derive an equivariant encoder. We discuss these steps next.

3. Methods

The dual goals of **(i)** equivariance to covariates (such as age) and **(ii)** invariance to site, involves learning multiple mappings. For simplicity, and to keep the computational effort manageable, we divide our method into two stages. Briefly, our stages are **(a) Stage one: Equivariance to Covariates.** We learn a mapping to a space that provides the essential flexibility to characterize changes in covariate C as a group action. This enables us to construct a space satisfying the equivariance condition as per Def. 1 **(b) Stage two: Invariance to Site.** We learn a second encoding to a generic vector space by apriori ensuring that the equivariance properties from Stage one are preserved. Such an encoding is then tuned to optimize the MMD criterion, thus generating a latent space that is invariant to site while remaining equivariant to covariates. We describe these stages one by one in the following sections.

3.1. Stage one: Equivariance to Covariates

Given the space of images, $\mathcal{X}$, with the covariates C, first, we want to characterize the effect of C on $\mathcal{X}$ as a group action for some group G. Here, an element $g \in G$ characterizes the change from covariate $c_i \in C$ to $c_j \in C$ (for short, we will use i and j). The change in C corresponds to a translation action which is difficult to instantiate in $\mathcal{X}$ without invoking expensive conditional generative models. Instead, we propose to learn a mapping to a latent space $\mathcal{L}$ such that the change in C can be characterized by a group action pertaining to G in the space $\mathcal{L}$ (the latent space of $\mathcal{X}$). As an example, let us say X_i goes to X_j in $\mathcal{X}$ as $(X_i \to X_j)$. This means that $(X_i \to X_j)$ is caused due to the covariate change $(c_i \to c_j)$ in C. Let $\mathfrak{E}$ be a mapping between the image space $\mathcal{X}$ and the latent space $\mathcal{L}$. In the latent space $\mathcal{L}$, for the moment, we want that $(\mathfrak{E}X_i \to \mathfrak{E}X_j)$ should correspond to the change in covariate $(c_i \to c_j)$.

Remark 2. *We are mostly interested in normalized covariates for example, in ℓ_p norm, while other volume based deterministic normalization functions may also be applicable. In the simplest case of $p = 2$ norm, the corresponding group action is naturally induced by the matrix group of rotations.*

Based on this choice of group, we will learn an autoencoder $(\mathfrak{E}, \mathfrak{D})$ with an encoder $\mathfrak{E} : \mathcal{X} \to \mathcal{L}$ and a decoder $\mathfrak{D} : \mathcal{L} \to \mathcal{X}$, here $\mathcal{L}$ is the encoding space. Due to Remark 2, we can choose $\mathcal{L}$ to be a hypersphere, $\mathbf{S}^{n-1}$, and $(\mathfrak{E}, \mathfrak{D})$ as a hyperspherical autoencoder [54]. Then, we can characterize the "action of C on $\mathcal{X}$" as the action of G on $\mathbf{S}^{n-1}$. That is to say that a covariate change (translation in C) is a change in angles on $\mathcal{L}$. This corresponds to a rotation due to the choice of our group G. Note that for $\mathcal{L} = \mathbf{S}^{n-1}$, G is the space of $n \times n$ rotation matrices, denoted by $\mathsf{SO}(n)$, and the action of G is well-defined. What remains is to encourage the latent space $\mathcal{L}$ to be G-equivariant. We start with some group theoretic properties that will be useful.

3.1.1 Review: Group theoretic properties of $\mathsf{SO}(n)$

Let $\mathsf{SO}(n) = \left\{X \in \mathbf{R}^{n \times n} | X^T X = I_n, \det(X) = 1\right\}$ be the group of $n \times n$ special orthogonal matrices. The group

$\mathsf{SO}(n)$ acts on $\mathbf{S}^{n-1}$ with the group action "$\cdot$" given by $g \cdot \boldsymbol{\ell} \mapsto g\boldsymbol{\ell}$, for $g \in \mathsf{SO}(n)$ and $\boldsymbol{\ell} \in \mathbf{S}^{n-1}$. Here we use $g\boldsymbol{\ell}$ to denote the multiplication of matrix g with $\boldsymbol{\ell}$. Under this group action, we can identify $\mathbf{S}^{n-1}$ with the quotient space G/H with $G = \mathsf{SO}(n)$ and $H = \mathsf{SO}(n-1)$ (see Ch. 3 of [13] for more details). Let $\tau : \mathbf{S}^{n-1} \to G/H$ be such an identification, i.e., $\tau(\boldsymbol{\ell}) = gH$ for some $g \in G$. The identification τ is equivariant to G in the following sense.

Fact 3. *Given $\tau : \mathbf{S}^{n-1} \to G/H$ as defined above, τ is equivariant with the action of G, i.e., $\tau(g \cdot \boldsymbol{\ell}) = g\tau(\boldsymbol{\ell})$.*

Next, we see that given two points $\boldsymbol{\ell}_i, \boldsymbol{\ell}_j$ on $\mathbf{S}^{n-1}$ there is a unique group element in G to move from $\tau(\boldsymbol{\ell}_i)$ to $\tau(\boldsymbol{\ell}_j)$.

Lemma 4. *Given two latent space representations $\boldsymbol{\ell}_i, \boldsymbol{\ell}_j \in \mathbf{S}^{n-1}$, and the corresponding cosets $g_i H = \tau(\boldsymbol{\ell}_i)$ and $g_j H = \tau(\boldsymbol{\ell}_j)$, $\exists! g_{ij} = g_j g_i^{-1} \in G$ such that $\boldsymbol{\ell}_j = g_{ij} \cdot \boldsymbol{\ell}_i$.*

Thanks to Fact 3 and Lemma A.1, simply identifying a suitable τ will provide us the necessary equivariance property. To do so, next, we parameterize τ by a neural network and describe a loss function to learn such a τ and $(\mathfrak{E}, \mathfrak{D})$.

3.1.2 Learning a G-equivariant τ with DNNs

Now that we established the key components: **(a)** an autoencoder $(\mathfrak{E}, \mathfrak{D})$ to map from $\mathcal{X}$ to the latent space $\mathbf{S}^{n-1}$ **(b)** a mapping $\tau : \mathbf{S}^{n-1} \to \mathsf{SO}(n)$ which is $G = \mathsf{SO}(n)$-equivariant, see Figure 3, we discuss how to learn such a $(\mathfrak{E}, \mathfrak{D})$ and a G-equivariant τ.

Let $X_i, X_j \in \mathcal{X}$ be two images with the corresponding covariates $i, j \in C$ with $i \neq j$. Let $\boldsymbol{\ell}_i = \mathfrak{E}(X_i), \boldsymbol{\ell}_j = \mathfrak{E}(X_j)$. Using Lemma A.1, we can see that a $g_{ij} \in G$ to move from $\boldsymbol{\ell}_i$ to $\boldsymbol{\ell}_j$ does exist and is unique. Now, to learn a τ that satisfies the equivariance property (Fact 3), we will need τ to satisfy two conditions, $\tau(g_{ij} \cdot \boldsymbol{\ell}_i) = g_{ij}\tau(\boldsymbol{\ell}_i)$ and $\tau(g_{ji} \cdot \boldsymbol{\ell}_j) = g_{ji}\tau(\boldsymbol{\ell}_j)\ \forall g \in G$. The two conditions are captured in the following loss function,

$$\boldsymbol{\ell}_i = \mathfrak{E}(X_i) \quad \boldsymbol{\ell}_j = \mathfrak{E}(X_j) \tag{2}$$

$$L_{\text{stage1}} = \sum_{\substack{\{(X_i,i),(X_j,j)\} \\ \subset \mathcal{X} \times C}} \begin{matrix} \|\mathcal{G}(i,j) \cdot \tau(\boldsymbol{\ell}_i) - \tau(\boldsymbol{\ell}_j)\|^2 \ + \\ \|\mathcal{G}^{-1}(i,j) \cdot \tau(\boldsymbol{\ell}_j) - \tau(\boldsymbol{\ell}_i)\|^2 \end{matrix} \tag{3}$$

Here, $\mathcal{G} : C \times C \to G$ will be a table lookup given by $(i,j) \mapsto g_{ij}$ is the function that takes two values for the covariate c, say, i, j corresponding to $X_i, X_j \in \mathcal{X}$ and simply returns the group element (rotation) g_{ij} needed to move from $\mathfrak{E}(X_i)$ to $\mathfrak{E}(X_j)$. **Choice of $\mathcal{G}$:** In general, learning $\mathcal{G}$ is difficult since C may not be continuous. In this work, we fix $\mathcal{G}$ and learn τ by minimizing (3). We will simplify the choice of $\mathcal{G}$ as follows: assuming that C is a numerical/ordinal random variable, we define $\mathcal{G}$ by $(i,j) \mapsto \mathsf{expm}((i-j)\mathbf{1}_m)$. Here $m = \binom{n}{2}$ is the dimension of G and expm is the matrix exponential, i.e., $\mathsf{expm} : \mathfrak{so}(n) \to \mathsf{SO}(n)$, where $\mathfrak{so}(n)$ is the Lie algebra [21] of $\mathsf{SO}(n)$. Since $\mathfrak{so}(n)$ is a vector space, hence $(i-j)\mathbf{1}_m \in \mathfrak{so}(n)$. To reduce the runtime of expm, we replace expm by a Cayley map [32, 42] defined by: $\mathfrak{so}(n) \ni A \mapsto (I-A)(I+A)^{-1} \in \mathsf{SO}(n)$. Here we used expm for parameterization (other choices also suitable).

Finally, we learn the encoder-decoder $(\mathfrak{E}, \mathfrak{D})$ by using a reconstruction loss constraint with L_{stage1} in (3). This can also be thought of as a combined loss for this stage as $L_{\text{stage1}} + \sum_i \|X_i - \mathfrak{D}(\mathfrak{E}(X_i))\|^2$ where the second term is the reconstruction loss. The loss balances two terms and requires a scaling factor (see appendix § A.7). A flowchart of all steps in this stage can be seen in Fig 3.

Algorithm 1 Learning representations that are *Equivariant* to *Covariates* and *Invariant* to *Site*

Input: Training Sets from multiple sites $(X,Y)_{\text{site1}}$, $(X,Y)_{\text{site2}}$. Nuisance covariates C.

Stage one: Equivariance to Covariates

1 : Parameterize Encoder-Decoder pairs $(\mathfrak{E}, \mathfrak{D})$ and τ mapping with neural networks

2 : Optimize over $(\mathfrak{E}, \mathfrak{D})$ and τ to minimize,
$L_{\text{stage1}} + \sum_i \|X_i - \mathfrak{D}(\mathfrak{E}(X_i))\|^2$

Output: First latent space mapping $\mathfrak{E}$ and a supporting mapping function τ. Here, τ is G-equivariant to the covariates C (see Lemma (A.1) and (3)).

Stage two: Invariance to Site

1 : Parameterize encoder b, predictor h and decoder Ψ with neural networks

2 : Preserve equivariance from stage one with an equivariant mapping Φ, (see Lemma (A.1))

3 : Optimize Φ, b, h and Ψ to minimize $L_{\text{stage2}} + \mathcal{MMD}$

Output: Second latent space mapping Φ. Here, Φ is equivariant to the covariates and invariant to site.

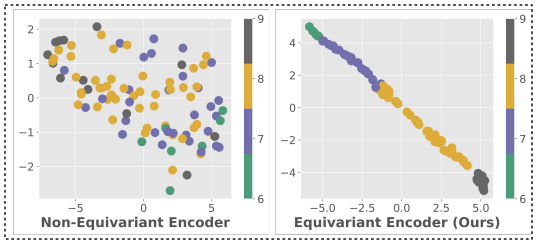
3.2. Stage two: Invariance to Site

Having constructed a latent space $\mathcal{L}$ that is equivariant to changes in the covariates C, we must now handle the site attribute, i.e., invariance with respect to site. Here, it will be convenient to project $\mathcal{L}$ onto a space that simultaneously preserves the equivariant structure from $\mathcal{L}$ *and* offers the flexibility to enforce site-invariance. The following lemma, inspired from the functional representations of probabilistic symmetries (§4.2 of [7]), provides us strategies to achieve this goal. Here, consider $\Phi : \mathcal{L} \to \mathcal{Z}$ to be the projection.

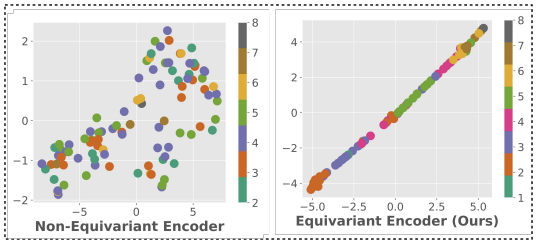
Lemma 5. *For a $\tau : \mathcal{L} \to G/H$ as defined above, and for any arbitrary mapping $b : \mathcal{L} \to \mathcal{Z}$, the function $\Phi : \mathcal{L} \to \mathcal{Z}$ defined by*

$$\Phi(\boldsymbol{\ell}) = \tau(\boldsymbol{\ell}) \cdot b\left(\tau(\boldsymbol{\ell})^{-1} \cdot \boldsymbol{\ell}\right) \tag{4}$$

is G-equivariant, i.e., $\Phi(g \cdot \boldsymbol{\ell}) = g\Phi(\boldsymbol{\ell})$.



(a) ADNI Dataset



(b) Adult Dataset

Figure 4. **t-SNE plots of latent representations $\tau(\boldsymbol{\ell})$.** For ADNI **(left)** and Adult **(right)**, an equivariant encoder ensures that the latent features are evenly distributed and bear a monotonic trend with respect to the changes in the age covariate value. The non-equivariant space is generated from the Naïve pooling baseline. Each color denotes a discretized **age** group. Age was discretized only for the figure to highlight the density of samples in each age group.

Proof is available in the appendix § A.1. Note that Φ remains equivariant for **any mapping** b. This provides us the option to parameterize b as a neural network and train the entirety of Φ for the desired site invariance where equivariance will be preserved due to (9). In this work, we learn such a $\Phi : \mathcal{L} \to \mathcal{Z}$ with the help of a decoder $\Psi : \mathcal{Z} \to \mathcal{L}$ by minimizing the following loss,

$$L_{\text{stage2}} = \sum_{\substack{\boldsymbol{\ell}=\mathfrak{E}(X)\in\mathcal{L} \\ X\in\mathcal{X},Y\in\mathcal{Y}}} \overbrace{\|\boldsymbol{\ell} - \Psi(\Phi(\boldsymbol{\ell}))\|^2}^{\text{Reconstruction loss}} + \overbrace{\|Y - h(\Phi(\boldsymbol{\ell}))\|^2}^{\text{Prediction loss}} \quad (5)$$

$$\text{subject to} \quad \underbrace{\Phi(\boldsymbol{\ell}) = \tau(\boldsymbol{\ell}) \cdot b\left(\tau(\boldsymbol{\ell})^{-1} \cdot \boldsymbol{\ell}\right)}_{G\text{-equivariant map}} \quad (6)$$

Minimizing the loss (5) with the constraint (6) allows learning the network $b : \mathcal{L} \to \mathcal{Z}$ and the decoder $\Psi : \mathcal{Z} \to \mathcal{L}$. We are now left with asking that $Z \in \mathcal{Z}$ be such that the representations are invariant across the sites. We simply use the following MMD criterion although other statistical distance measures can also be utilized.

$$\mathcal{MMD} = \| \mathop{\mathbb{E}}_{\substack{Z_1\sim \\ P(\Phi(\boldsymbol{\ell}))_{\text{site1}}}} \mathcal{K}(Z_1,\cdot) - \mathop{\mathbb{E}}_{\substack{Z_2\sim \\ P(\Phi(\boldsymbol{\ell}))_{\text{site2}}}} \mathcal{K}(Z_2,\cdot)\|_{\mathcal{H}} \quad (7)$$

The criterion is defined using a Reproducing Kernel Hilbert Space with norm $\|\cdot\|_{\mathcal{H}}$ and kernel $\mathcal{K}$. We combine (5), (6) and (7) as the objective function to ensure site invariance. Thus, the combined loss function $L_{\text{stage2}} + \mathcal{MMD}$ is minimized to learn (Φ, Ψ). Scaling factor details are available in the appendix § A.7.

Summary of the two stages. Our overall method comprises of two stages. The first stage, Section 3.1, involves learning the τ function. The function learned in this stage is G-equivariant by the choice of the loss L_{stage1}, see (3). Our next stage, Section 3.2, employs the learned τ function and a trainable mapping b to generate invariant representations. This stage preserves G-equivariance due to the Φ mapping in (9). The loss for the second step is $L_{\text{stage2}} + \mathcal{MMD}$, see (5). Our method is summarized in Algorithm 1. Convergence behavior of the proposed optimization (of τ, Φ) still seems challenging to characterize exactly, but recent papers provide some hope, and opportunities. For example, if the networks are linear, then results from [18] maybe applicable which explain the our superior empirical performance.

4. Experiments

We evaluate our proposed encoder for site-invariance and robustness to changes in the covariate values C. Evaluations are performed on two multi-site neuroimaging datasets, where algorithmic developments are likely to be most impactful. Prior to neuroimaging datasets, we also conduct experiments on two standard fairness datasets, German and Adult. The inclusion of fairness datasets in our analysis, provides us a means for sanity tests and optimization feasibility on an established problem. Here, the goal of achieving fair representations is treated as pooling multiple subsets of data indexed by separate sensitive attributes. We begin our analysis by first describing our measures of evaluation and then reporting baselines for comparisons.

Measures of Evaluation. Recall that our method involves learning τ as in (3) to satisfy the equivariance property. Moreover, we need to learn Φ as in (9)–(5) to achieve site invariance. Our measures assess the structure of the latent space $\tau(\boldsymbol{\ell})$ and $\Phi(\boldsymbol{\ell})$. The measures are: **(a)** $\boldsymbol{\Delta_{Eq}}$: This metric evaluates the ℓ_2 distance between $\tau(\boldsymbol{\ell}_i)$ and $\tau(\boldsymbol{\ell}_j)$ for all pairs i, j. Formally, it is computed as

$$\boldsymbol{\Delta_{Eq}} = \sum_{\substack{\{(X_i,i),(X_j,j)\}\subset\mathcal{X}\times C \\ \boldsymbol{\ell}_i=\mathfrak{E}_e(X_i),\boldsymbol{\ell}_j=\mathfrak{E}_e(X_j)}} |i-j| \|\tau(\boldsymbol{\ell}_i) - \tau(\boldsymbol{\ell}_j)\|^2 \quad (8)$$

A higher value of this metric indicates that $\tau(\boldsymbol{\ell}_i)$ and $\tau(\boldsymbol{\ell}_j)$ are related by the group action g_{ij}. Additionally, we use t-SNE [48] to qualitatively visualize the effect of τ. **(b)** $\boldsymbol{Adv}$: This metric quantifies the site-invariance achieved by the encoder Φ. We evaluate if $\Phi(\boldsymbol{\ell})$ for a learned $\boldsymbol{\ell} \in \mathcal{L}$ has any information about the site. A three layered fully network (see appendix § A.6) is trained as an adversary to predict site from $\Phi(\boldsymbol{\ell})$, similar to [49]. A lower value of Adv, that is close to random chance, is desirable. **(c)** $\boldsymbol{\mathcal{M}}$: Here, we compute the $\mathcal{MMD}$ measure, as in (7), on the test set. A smaller value of $\mathcal{M}$ indicates better invariance to site. Lastly, **(d)** $\boldsymbol{\mathcal{ACC}}$: This metric notes the test set

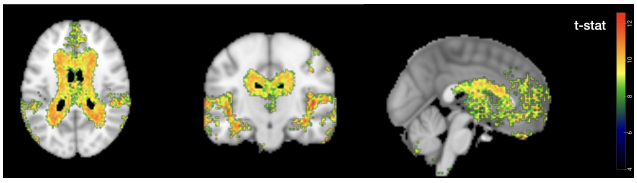


Figure 5. **Statistical Analysis on the reconstructed outputs.** The voxels that are significantly associated with Alzheimer's disease ($p < 0.001$) are shown. Adjustments for multiple comparisons were made using Bonferroni correction. A high density of significant voxels indicates that our method preserves disease related signal after pooling across scanners.

accuracy in predicting the target variable Y.

Baselines for Comparison. We contrast our method's performance with respect to a few well-known baselines. (i) **Naïve:** This method indicates a naïve approach of pooling data from multiple sites without any scheme to handle nuisance variables. (ii) **MMD [29]:** This method minimizes the distribution differences across the sites without any requirements for equivariance to the covariates. The latent representations being devoid of the equivariance property result in lower accuracy values as we will see shortly. (iii) **CAI [49]:** This method introduces a discriminator to train the encoder in a minimax adversarial fashion. The training routine directly optimizes the Adv measure above. While being a powerful implicit data model, adversarial methods are known to have unstable training and lack convergence guarantees [40]. (iv) **SS [55]:** This method adopts a Sub-sampling (SS) framework to divide the images across the sites by the covariate values C. An MMD criterion is minimized individually for each of the sub-sampled groups and an average estimate is computed. Lastly, (v) **RM [33]:** Also used in [31], RandMatch (RM) learns invariant representations on samples across sites that "match" in terms of the class label (we match based on both Y and C values) . Below, we summarize each method and nuisance attribute correction adopted by them.

Correction	Naïve	MMD [29]	CAI [49]	SS [55]	RM [33]	Ours
Site	✗	✓	✓	✓	✓	✓
Covariates	✗	✗	✗	✓	✓	✓

Table 1. Baselines in the paper and their nuisance attribute correction.

We evaluate methods on the test partition provided with the datasets. The mean of the metrics over three random seeds is reported. The hyper-parameter selection is done on a validation split from the training set, such that the prediction accuracy falls within 5% window relative to the best performing model [10] (more details in appendix § A.2).

4.1. Obtaining Fair Representations

We approach the problem of learning fair representations through our multi-site pooling formulation. Specifically, we consider each sensitive attribute value as a separate site. Results on two benchmark datasets, German and Adult [11], are described below.

German Dataset. This dataset is a classification problem used to predict defaults on the consumer loans in the German market. Among the several features in the dataset, the attribute *foreigner* is chosen as a sensitive attribute. We train our encoder while maintaining equivariance with respect to the continuous valued *age* feature. Table 2 provides a summary of the results in comparison to the baselines. Our equivariant encoder maximizes the Δ_{Eq} metric indicating the the latent space $\tau(\ell)$ is well separated for different values of *age*. Further, the invariance constraint improves the Adv metric signifying a better elimination of sensitive attribute information from the representations. The $\mathcal{M}$ metric is higher relative to the other baselines. The $\mathcal{ACC}$ for all the methods are within a 2% range.

Adult Dataset. In the Adult dataset, the task is to predict if a person has an income higher (or lower) than $\$50K$ per year. The dataset is biased with respect to gender, roughly, 1-in-5 women (in contrast to 1-in-3 men) are reported to make over $\$50K$. Thus, the female/male genders are considered as two separate sites with *age* as a nuisance covariate feature. As shown in Table 2, our equivariant encoder improves on metrics Δ_{Eq} and Adv relative to all the baselines similar to the German dataset. In addition to the quantitative metrics, we visualize the t-SNE plots of the representations $\tau(\ell)$ in Fig. 4 (right). It is clear from the figure that an equivariant encoder imposes a certain monotonic trend as the *Age* values as varied.

4.2. Pooling Brain Images across Scanners

For our motivating application, we focus on pooling tasks for two different brain imaging datasets where the problem is to classify individuals diagnosed with Alzheimer's disease (AD) and healthy control (CN).

Setup. Images are pre-processed by first normalizing and then skull-stripping using Freesurfer [15]. A linear (affine) registration is performed to register each image to MNI template space. Images are trained using 3D convolutions with ResNet [23] backbone (details in the appendix § A.6). Since the datasets are small, we report results over five random training-validation splits.

ADNI Dataset. The data for this experiment has been downloaded from the Alzheimers Disease Neuroimaging

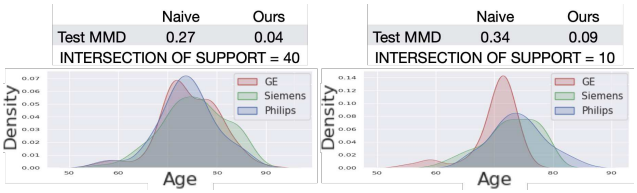


Figure 6. **Distribution of age covariate in the ADNI dataset.** Two settings are considered – **(left)** the intersection of the support is large, and **(right)** with a smaller common support. Despite the mismatch of support across scanner attributes, our approach minimizes the MMD measure (desirable) on the test set relative to the naïve pooling method.

$\boldsymbol{\Delta_{Eq}}$: Equivariance Gap, $\boldsymbol{Adv}$: Adversarial Test Accuracy, $\boldsymbol{\mathcal{M}}$: Test $\mathcal{MMD}$ measure, $\boldsymbol{\mathcal{ACC}}$: Test prediction accuracy
↑: Higher Value is preferred, ↓: Lower Value is preferred

	German				Adult				ADNI				ADCP			
	$\Delta_{Eq}\uparrow$	$Adv\downarrow$	$\mathcal{M}\downarrow$	$\mathcal{ACC}\uparrow$	$\Delta_{Eq}\uparrow$	$Adv\downarrow$	$\mathcal{M}\downarrow$	$\mathcal{ACC}\uparrow$	$\Delta_{Eq}\uparrow$	$Adv\downarrow$	$\mathcal{M}\downarrow$	$\mathcal{ACC}\uparrow$	$\Delta_{Eq}\uparrow$	$Adv\downarrow$	$\mathcal{M}\downarrow$	$\mathcal{ACC}\uparrow$
Naïve	4.6(0.7)	0.62(0.03)	7.7(0.8)	74(0.9)	3.4(0.7)	83(0.1)	9.8(0.3)	84(0.1)	3.1(1.0)	59(2.9)	27(1.6)	80(2.6)	4.1(0.9)	49(8.4)	90(8.7)	83(4.4)
MMD [29]	4.5(1.0)	0.66(0.04)	1.5(0.3)	73(1.5)	3.4(0.9)	83(0.1)	3.1(0.3)	84(0.1)	3.1(1.0)	59(3.3)	27(1.7)	80(2.6)	3.6(1.0)	49(11.9)	86(11.0)	84(6.5)
CAI [49]	1.9(0.6)	0.65(0.01)	1.2(0.2)	76(1.3)	0.1(0.0)	81(0.7)	4.2(2.4)	84(0.04)	2.4(0.7)	61(2.1)	27(1.5)	74(3.6)	2.8(1.6)	56(6.9)	85(12.3)	82(5.1)
SS [55]	3.8(0.5)	0.70(0.07)	1.5(0.6)	76(0.9)	2.8(0.5)	83(0.2)	1.5(0.2)	84(0.1)	3.7(0.5)	57(2.1)	26(1.6)	81(3.7)	3.4(1.3)	51(6.7)	88(14.6)	82(3.5)
RM [33]	3.4(0.4)	0.66(0.04)	7.5(0.9)	74(2.1)	0.8(0.1)	82(0.4)	4.8(0.7)	84(0.3)	0.8(0.9)	52(5.4)	22(0.6)	78(3.8)	0.4(0.5)	40(4.7)	77(13.8)	84(5.3)
Ours	**6.4(0.6)**	**0.54(0.01)**	2.7(0.6)	75(3.3)	**5.3(0.9)**	**75(1.4)**	7.1(0.6)	83(0.1)	**5.1(1.2)**	**50(4.2)**	**16(7.2)**	77(4.8)	**7.5(1.2)**	49(7.3)	**70(22.3)**	81(1.8)

Table 2. **Quantitative Results.** We show Mean(Std) results over multiple run. For our baselines, we consider a **Naïve** encoder-decoder model, learning representations via minimizing the **MMD** criterion [29] and Adversarial training [49], termed as **CAI**. We also compare against Sub-sampling (**SS**) [55] that minimizes the MMD criterion separately for every age group, and the RandMatch (**RM**) [33] baseline that generates matching input pairs based on the Age and target label values. The SS and RM baselines discard subset of samples if a match across sites is not available. The measure Adv represents the adversarial test accuracy except for the German dataset where ROC-AUC is used due to high degree of skew in the data.

Initiative (ADNI) database (adni.loni.usc.edu). We have three scanner types in the dataset, namely, GE, Siemens and Phillips. Similar to the fairness experiments, equivariance is sought relative to the covariate *Age*. The values of *Age* are in the range 50-95 as indicated in density plot of Fig. 6 (left). The *Age* distribution is observed to vary across different scanners, albeit minimally, in the full dataset. In the t-SNE plot, Fig. 4 (left), we see that the latent space has an equivariant structure. Closer inspection of the plot shows that the representations vary in the same order as that of *Age*. Different colors indicate different **Age** sub-groups. Next, in Fig. 5, we present the t-statistics in the template space on the reconstructed images after pooling. Here, the t-statistics measure the association with AD/CN target labels. As seen in the figure, the voxels significantly associated with the Alzheimer's disease ($p < 0.001$) are considerable in number. This result supports our goal to combine datasets to increase sample size and obtain a high power in statistical analysis. Next, in Fig. 6 (right), we increase the difficulty of our problem by randomly sub-sampling for each scanner group such that the intersection of support is minimized. In such an extreme case, our method attains a better $\mathcal{M}$ metric relative to the Naïve method, thus justifying the applicability to situations where there is a mismatch of support across the sites. Lastly, we inspect the performance on the quantitative metrics on the entire dataset in Table 2. All metrics Δ_{Eq}, Adv and $\mathcal{M}$ improve relative to the baselines with a small drop in the $\mathcal{ACC}$.

ADCP dataset. This experiment's data was collected as part of the NIH-sponsored Alzheimer's Disease Connectome Project (ADCP) [1, 25]. It is a two-center MRI, PET, and behavioral study of brain connectivity in AD. Study inclusion criteria for AD / MCI (Mild Cognitive Impairment) patients consisted of age $55-90$ years who retain decisional capacity at initial visit, and meet criteria for probable AD or MCI. MRI images were acquired at three sites. The three sites differ primarily in terms of the patient demographics. We inspect the quantitative results of this experiment in Tab. 2 and place the qualitative results in the appendix § A.4,A.5. The table reveals considerable improvements in all our metrics relative to the Naïve method.

Limitations. Currently, our formulation assumes that the to-be-pooled imaging datasets are roughly similar – there is definitely a role for new developments in domain alignment to facilitate deployment in a broader range of applications. Secondly, larger latent space dimensions may cause compute overhead due to matrix exponential parameterization. Finally, algorithmic improvements can potentially simplify the overhead of the two-stage training.

5. Conclusions

Retrospective analysis of data pooled from previous / ongoing studies can have a sizable influence on identifying early disease processes, not otherwise possible to glean from analysis of small neuroimaging datasets. Our development based on recent results in equivariant representation learning offers a strategy to perform such analysis when covariates/nuisance attributes are not identically distributed across sites. Our current work is limited to a few such variables but suggests that this direction is promising and can potentially lead to more powerful algorithms.

Acknowledgments The authors are grateful to Vibhav Vineet (Microsoft Research) for discussions on the causal diagram used in the paper. Thanks to Amit Sharma (Microsoft Research) for the conversation on their MatchDG project. Special thanks to Veena Nair and Vivek Prabhakaran from UW Health for helping with the ADCP dataset. Research supported by NIH grants to UW CPCP (U54AI117924), RF1AG059312, Alzheimer's Disease Connectome Project (ADCP) U01 AG051216, and RF1AG059869, as well as NSF award CCF 1918211. Sathya Ravi was also supported by UIC-ICR start-up funds.

References

[1] Nagesh Adluru, Veena A Nair, Vivek Prabhakaran, Shi-Jiang Li, Andrew L Alexander, and Barbara B Bendlin. Geodesic path differences in neural networks in the alzheimer's disease connectome project: Developing topics. *Alzheimer's & Dementia*, 16:e047284, 2020. 8

[2] Paul S Aisen, Jeffrey Cummings, Clifford R Jack, John C Morris, Reisa Sperling, Lutz Frölich, Roy W Jones, Sherie A Dowsett, Brandy R Matthews, Joel Raskin, et al. On the path to 2025: understanding the alzheimer's disease continuum. *Alzheimer's research & therapy*, 9(1):1–10, 2017. 11, 12

[3] Aditya Kumar Akash, Vishnu Suresh Lokhande, Sathya N Ravi, and Vikas Singh. Learning invariant representations using inverse contrastive loss. *arXiv preprint arXiv:2102.08343*, 2021. 2

[4] Jesper LR Andersson, Mark Jenkinson, Stephen Smith, et al. Non-linear registration, aka spatial normalisation fmrib technical report tr07ja2. *FMRIB Analysis Group of the University of Oxford*, 2(1):e21, 2007. 12

[5] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2

[6] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. 2

[7] Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020. 4, 5

[8] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. 4

[9] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020. 3

[10] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 7

[11] Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017. 7

[12] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14340–14349, 2021. 2

[13] David S. Dummit and Richard M. Foote. *Abstract algebra*. Wiley, 3rd ed edition, 2004. 5

[14] Simon F Eskildsen, Pierrick Coupé, Vladimir S Fonov, Jens C Pruessner, D Louis Collins, Alzheimer's Disease Neuroimaging Initiative, et al. Structural imaging biomarkers of alzheimer's disease: predicting disease progression. *Neurobiology of aging*, 36:S23–S31, 2015. 2

[15] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012. 7, 11

[16] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170, 2017. 1

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[18] Avishek Ghosh and Ramchandran Kannan. Alternating minimization converges super-linearly for mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1093–1103. PMLR, 2020. 6

[19] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013. 1, 12

[20] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19:513–520, 2006. 3

[21] Marshall Hall. *The theory of groups*. Courier Dover Publications, 2018. 5

[22] Moritz Hardt and Benjamin Recht. *Patterns, predictions, and actions: A story about machine learning*. https://mlstory.org, 2021. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 7

[24] Mark Hunacek. Lie groups, lie algebras, and representations: An elementary introduction, by brian hall. pp. 351.£ 50. 2003. isbn 0 387 401229 (springer-verlag). *The Mathematical Gazette*, 89(514):149–151, 2005. 4

[25] Gyujoon Hwang, Cole John Cook, Veena A Nair, Andrew L Alexander, Piero G Antuono, Sanjay Asthana, Rasmus Birn, Cynthia M Carlsson, Guangyu Chen, Dorothy Farrar Edwards, et al. Ic-p-161: Characterizing structural brain alterations in alzheimer's disease patients with machine learning. *Alzheimer's & Dementia*, 14(7S_Part_2):P135–P136, 2018. 8

[26] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008. 1, 2

[27] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012. 12

[28] Anthony W Knapp and AW Knapp. *Lie groups beyond an introduction*, volume 140. Springer, 1996. 4

[29] Yujia Li, Kevin Swersky, and Richard Zemel. Learning unbiased features. *arXiv preprint arXiv:1412.5244*, 2014. 4, 7, 8

[30] Jingqin Luo, Folasade Agboola, Elizabeth Grant, Colin L Masters, Marilyn S Albert, Sterling C Johnson, Eric M McDade, Jonathan Vöglein, Anne M Fagan, Tammie Benzinger, et al. Sequence of alzheimer disease biomarker changes in cognitively normal adults: A cross-sectional study. *Neurology*, 95(23):e3104–e3116, 2020. 1

[31] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021. 2, 3, 7

[32] Ronak Mehta, Rudrasis Chakraborty, Yunyang Xiong, and Vikas Singh. Scaling recurrent models via orthogonal approximations in tensor trains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10571–10579, 2019. 5

[33] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017. 7, 8

[34] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2, 11

[35] Prashant Pandey, Mrigank Raman, Sumanth Varambally, and Prathosh AP. Domain generalization via inference-time label-preserving target projections. *arXiv preprint arXiv:2103.01134*, 2021. 2

[36] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 3

[37] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 2

[38] Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. An online method for distributionally deep robust optimization, 2020. 3

[39] Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985. 2

[40] Florian Schaefer and Anima Anandkumar. Competitive gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 7

[41] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 2

[42] Jon M Selig. Cayley maps for se (3). In *12th International Federation for the Promotion of Mechanism and Machine Science World Congress*, page 6. London South Bank University, 2007. 5

[43] Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017. 3

[44] Anja Soldan, Corinne Pettigrew, Anne M Fagan, Suzanne E Schindler, Abhay Moghekar, Christopher Fowler, Qiao-Xin Li, Steven J Collins, Cynthia Carlsson, Sanjay Asthana, et al. Atn profiles among cognitively normal individuals and longitudinal cognitive outcomes. *Neurology*, 92(14):e1567–e1579, 2019. 1

[45] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019. 2

[46] Chung-Piaw Teo, Jay Sethuraman, and Wee-Peng Tan. Gale-shapley stable marriage problem revisited: Strategic issues and applications. *Management Science*, 47(9):1252–1267, 2001. 3

[47] Paul M Thompson, Jason L Stein, Sarah E Medland, Derrek P Hibar, Alejandro Arias Vasquez, Miguel E Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, et al. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior*, 8(2):153–182, 2014. 1

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[49] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 6, 7, 8, 11

[50] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8978–8987, 2021. 2

[51] Zhuoran Yang, Yufeng Zhang, Yongxin Chen, and Zhaoran Wang. Variational transport: A convergent particle-basedalgorithm for distributional optimization. *arXiv preprint arXiv:2012.11554*, 2020. 3

[52] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013. 2

[53] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 2

[54] Deli Zhao, Jiapeng Zhu, and Bo Zhang. Latent variables on spheres for autoencoders in high dimensions. *arXiv preprint arXiv:1912.10233*, 2019. 4

[55] Hao Henry Zhou, Vikas Singh, Sterling C Johnson, Grace Wahba, Alzheimer's Disease Neuroimaging Initiative, et al. Statistical tests and identifiability conditions for pooling and

analyzing multisite datasets. *Proceedings of the National Academy of Sciences*, 115(7):1481–1486, 2018. 2, 3, 7, 8

[56] Hao Henry Zhou, Yilin Zhang, Vamsi K. Ithapu, Sterling C. Johnson, Grace Wahba, and Vikas Singh. When can multi-site datasets be pooled for regression? Hypothesis tests, ℓ_2-consistency and neuroscience applications. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4170–4179. PMLR, 06–11 Aug 2017. 1

A. Appendix

A.1. Proofs of theoretical results

In this section, we will provide the proofs of Lemma 4 and Lemma 5 discussed in the main paper.

Lemma. *Given two latent space representations $\boldsymbol{\ell}_i, \boldsymbol{\ell}_j \in \mathbf{S}^{n-1}$, and the corresponding cosets $g_i H = \tau(\boldsymbol{\ell}_i)$ and $g_j H = \tau(\boldsymbol{\ell}_j)$, $\exists! g_{ij} = g_j g_i^{-1} \in G$ such that $\boldsymbol{\ell}_j = g_{ij} \cdot \boldsymbol{\ell}_i$.*

Proof. Given $g_i H = \tau(\boldsymbol{\ell}_i)$ and $g_j H = \tau(\boldsymbol{\ell}_j)$, we use $g_{ij} = g_i g_j^{-1} \in G$ such that, $g_j H = g_{ij} g_i H$.

Now using the equivariance fact (3) , we get,

$$\begin{aligned} & g_j H = g_{ij} g_i H \\ \implies & \tau(\boldsymbol{\ell}_j) = g_{ij} \tau(\boldsymbol{\ell}_i) \\ \implies & \tau(\boldsymbol{\ell}_j) = \tau(g_{ij} \cdot \boldsymbol{\ell}_i) \end{aligned}$$

Now as τ is an identification, i.e., a diffeomorphism, we get $\boldsymbol{\ell}_j = g_{ij} \boldsymbol{\ell}_i$. Note that $\mathbf{S}^{n-1}$ is a Riemannian homogeneous space and the group G acts transitively on $\mathbf{S}^{n-1}$, i.e., given $\mathbf{x}, \mathbf{y} \in \mathbf{S}^{n-1}$, $\exists g \in G$ such that, $\mathbf{y} = g \cdot \mathbf{x}$. Hence from $\boldsymbol{\ell}_j = g_{ij} \boldsymbol{\ell}_i$ and the transitivity property we can conclude that g_{ij} is unique. □

Lemma. *For a $\tau : \mathcal{L} \to G/H$ as defined above, and a mapping $b : \mathcal{L} \to \mathcal{Z}$, the function $\Phi : \mathcal{L} \to \mathcal{Z}$ defined by*

$$\Phi(\boldsymbol{\ell}) = \tau(\boldsymbol{\ell}) \cdot b\left(\tau(\boldsymbol{\ell})^{-1} \cdot \boldsymbol{\ell}\right) \tag{9}$$

is G-equivariant, i.e., $\Phi(g \cdot \boldsymbol{\ell}) = g \Phi(\boldsymbol{\ell})$.

Proof. Let $\boldsymbol{\ell} \in \mathcal{L}$. Consider the Φ mapping of $g \cdot \boldsymbol{\ell}$, that is $\Phi(g \cdot \boldsymbol{\ell}) = \tau(g \cdot \boldsymbol{\ell}) \cdot b\left(\tau(g \cdot \boldsymbol{\ell})^{-1} \cdot \boldsymbol{\ell}\right)$.

Using the fact (3) from the main paper, we have $\tau(g \cdot \boldsymbol{\ell}) = g\tau(\boldsymbol{\ell})$ and $\tau(g \cdot \boldsymbol{\ell})^{-1} = \tau(\boldsymbol{\ell})^{-1} g^{-1}$. Substituting these in $\Phi(g \cdot \boldsymbol{\ell})$, we get

$$\begin{aligned} \Phi(g \cdot \boldsymbol{\ell}) &= g\tau(\boldsymbol{\ell}) \cdot b\left(\tau(\boldsymbol{\ell})^{-1} g^{-1} g \cdot \boldsymbol{\ell}\right) \\ &= g\tau(\boldsymbol{\ell}) b\left(\tau(\boldsymbol{\ell})^{-1} \cdot \boldsymbol{\ell}\right) \\ \text{Thus, } \Phi(g \cdot \boldsymbol{\ell}) &= g\Phi(\boldsymbol{\ell}) \end{aligned}$$

□

A.2. Details on Evaluation Metrics

Recall from Section 4 of the paper, our discussion on three metrics – $\boldsymbol{\Delta_{Eq}}$, $\boldsymbol{Adv}$ and $\boldsymbol{\mathcal{M}}$. While $\boldsymbol{\Delta_{Eq}}$ and $\boldsymbol{\mathcal{M}}$ are variants of distance measure on the latent space, $\boldsymbol{Adv}$ assesses the ability to predict the nuisance attributes from the latent representation (and is therefore probabilistic in nature). Observe that $\boldsymbol{\Delta_{Eq}}$ and $\boldsymbol{\mathcal{M}}$ are (euclidean) distance measures and could be very different depending on the normalization of the vectors. For our purposes of evaluating these latent vectors/features in downstream tasks, we perform a simple feature normalization in order to obtain $0-1$ latent vectors given by,

$$\tilde{z}_i = \frac{z_i - \min(z_i)}{\max(z_i) - \min(z_i)}. \tag{10}$$

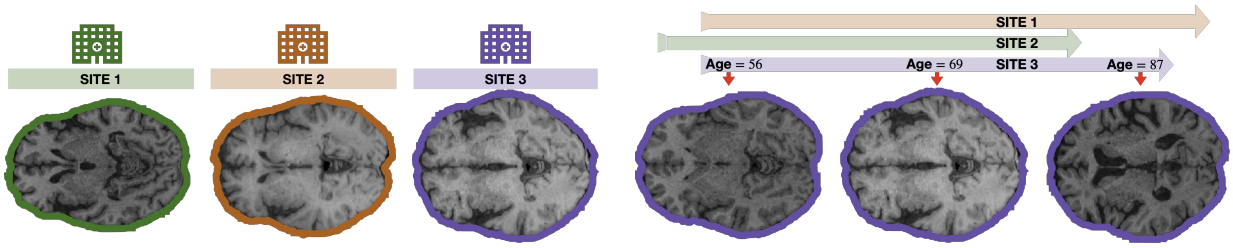
Our feature normalization is composed of two steps: (i) centering – the numerator in (10) ensures that the mean of z (along its coordinates) is 0; and (ii) scale – the denominator projects the features z on the sphere at origin with radius $\|z_i\|_\infty^{\geq} = \max(z_i) - \min(z_i) \geq 0$. Note that our scaling step can be thought of as the usual projection in a special case: when z_i is guaranteed to be nonnegative (for example, when z_i represent activations), then $\|z_i\|_\infty^{\geq}$ simply corresponds to a lower bound of the usual infinity norm, $\|z\|_\infty$ (hence projection on a scaled ℓ_∞ ball). We adopt this normalization only to compute $\boldsymbol{\Delta_{Eq}}$ and $\boldsymbol{\mathcal{M}}$ measures, and not for model training.

For computing the $\boldsymbol{Adv}$ measure, we follow [49] to train an adversarial neural network predicting the nuisance attributes. We use a three-layered fully connected network with batch normalization and train for 150 epochs. [34] uses similar architecture for the adversaries with different hidden layers of $0, 1, 2, 3$. We found that a three-layer adversary is powerful enough to predict the nuisance attributes and hence we use it to report the $\boldsymbol{Adv}$ measure.

A.3. Understanding ADNI dataset

Dataset. The data was downloaded from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. ADNI was set up with an objective to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD) using serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers. We have three imaging protocol (scanner) types in the dataset, namely, GE, Siemens and Phillips. The count of samples AD/CN in each of these imaging protocols are provided in Table 3. An example illustration (borrowed from [2]) of using different scanner on the images is shown in Figure 8.

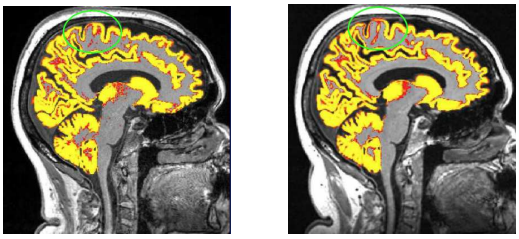
Preprocessing. All images were first normalized and skull-stripped using Freesurfer [15]. A linear (affine) registra-



(a) Variation due to scanner for particular age group.

(b) Variation due to covariates (age) in scanner 3.

Figure 7. **Sample Images from ADCP dataset.** **(a)** MRI images on control subjects from the ADCP dataset for different **sites** in the age group 70-80. **(b)** Images obtained from Site 3 for three extreme **age groups**. The gantt chart on top of the image indicates the respective age range in the other sites.



(a) GE (b) Siemens

Figure 8. **Scanner effects on images.** Two imaging protocols are shown: (a) Siemens, (b) GE. The yellow region is the cortical ribbon segmentation, and the green circle shows that the imaging protocol from different manufacturers have an effect on the scan. Image borrowed from [2].

tion was performed to register each image to MNI template space.

A.4. Understanding ADCP dataset

Participants. The data for ADCP was collected through an NIH-sponsored Alzheimer's Disease Connectome Project (ADCP) U01 AG051216. The study inclusion criteria for AD (Alzheimer's disease) / MCI (Mild Cognitive Impairment) patients consisted of age between 55-90 years, willing and able to undergo all procedures, retains decisional capacity at initial visit, meets criteria for probable AD or meets criteria for MCI.

Scanners. MRI images were acquired at three distinct sites on GE scanners. T1-weighted structural images were acquired using a 3D gradient-echo pulse sequence (repetition time (TR) = 604 ms, echo time (TE) = 2.516 ms, inversion time = 1060 ms, flip angle = 8°, field of view (FOV) = 25.6 cm, 0.8 mm isotropic). T2-weighted structural images were acquired using a 3D fast spin-echo sequence (TR = 2500 ms, TE = 94.398 ms, flip angle = 90°, FOV = 25.6 cm, 0.8 mm isotropic).

Preprocessing. The Human Connectome Project (HCP) minimal preprocessing pipeline version 3.4.0 [19] was followed for data processing. This pipeline is based on FMRIB Software Library [27]. Next, the T1w and T2w images are aligned, a B1 (bias field) correction is performed, and the subject's image in native structural volume space is registered to MNI space using FSL's FNIRT [4]. Only T1w images in the MNI space were used for further analysis and experiments.

Data Statistics. We plot the distributions of several attributes in this dataset conditioned on the site. In Figure 10, we show that the values of age and cognitive scores differ across the three sites in this dataset. Cognitive scores are computed based on an test assigned to the patients. Higher scores indicate higher cognitive operation in the patient. Table 4 shows the sample counts for target variable of prediction AD (Alzheimer's disease) and Control group.

A.5. Visualizing the latent space

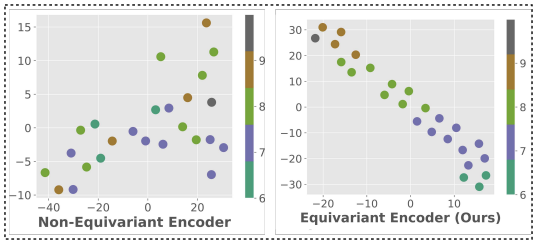
In the paper Figure 4, we have seen the latent space $\tau(\ell)$ for the samples in the ADNI and the Adult datasets. Here, we will see similar qualitative results for the German and the ADCP dataset in Figure 9 of the supplement. In the plots, the latent representations for a non-equivariant encoder are stretched thoughout the latent space. In contrast, the representations of an equivariant encoder, for a discretized value of **Age**, are localized to specific regions. Further, these representations have a monotonic behaviour with respect to the values of **Age**.

Table 3. Sample counts for ADNI dataset

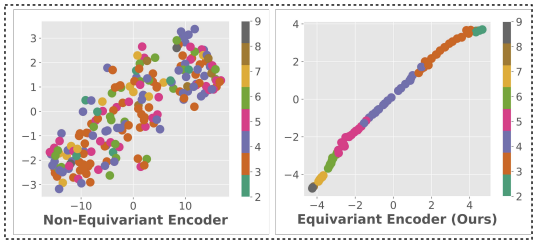
Imaging Protocol	AD	CN
Manufacturer=GE Medical Systems	44	78
Manufacturer=Philips Medical Systems	32	50
Manufacturer=Siemens	83	162

Table 4. Sample counts for ADCP dataset

	AD	Control	Female	Male
site 1	10	39	29	20
site 2	10	33	30	13
site 3	5	19	14	10

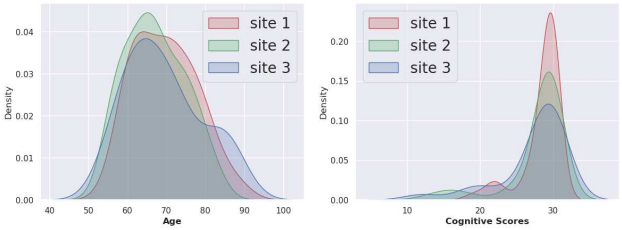


(a) ADCP Dataset



(b) German Dataset

Figure 9. **t-SNE plots of latent representations of $\tau(\ell)$** . For both ADCP **(left)** and German **(right)**, the the latent vectors of the equivariant encoder are evenly distributed with respect to the age covariate value. The non-equivariant space is generated from the naïve pooling model. Different colors denote the discretized set of **age** covariate value present in the data.



(a) Age (b) Cognitive Score

Figure 10. **Distribution of attributes in the ADCP dataset.** On the **left** we observe the distribution of age for the three different sites present in the ADCP dataset. On the **right**, we see the distribution of the cognitive scores. The cognitive scores are computed based on a test that assesses executive function. Higher scores indicate higher level of cognitive flexibility. Both age and cognitive scores are observed to vary across the sites.

Listing 1. Residual Block

```
BatchNorm3d
Swish
Conv3d
BatchNorm3d
Swish
Conv3d
```

Listing 2. Fully Connected Block

```
AdaptiveAvgPool3d
Flatten
Dropout
Linear
BatchNorm1d
Swish
Dropout
Linear
```

A.6. Hyper-parameters and NN Architectures

For tabular datasets such as German and Adult, our encoders and decoders comprise of fully connected networks and a hidden layer of 64 nodes. The dimension of the quotient latent space $\tau(\ell_i)$ is 30. Adam is used as a default optimizer and the learning rate is adjusted based on the validation set.

Imaging datasets like ADNI and ADCP require 3D convolutions and a ResNet architecture as the backbone. The last layer is used to describe the quotient space $\tau(\ell_i)$. We present the residual and the fully connected block below. Detailed architectures can be viewed in the code.

A.7. Scaling factors

Recall from the Algorithm 1 of the main paper that our loss function for each stage comprises of reconstruction and prediction losses in addition to the objectives concerning equivariance and invariance. These multi-objective loss functions require scaling factors that upweight one objective over the other. These scaling factors group up as hyperparameters for the Algorithm. In our experiments, it was observed that the results were robust to a range of scaling factor choices. For the results reported in Table 1 of the paper, they were identified through cross-validation. Here we provide an example for the scaling factors used for the Adult dataset, please refer to the bash scripts available in the code for the scaling factors of other datasets.

- **Stage one: Equivariance to Covariates**
  - Equivariance Loss L_{stage1}
    Scaling Factor : 1.0
  - Reconstruction Loss $\sum_i \|X_i - \mathfrak{D}(\mathfrak{E}(X_i))\|$
    Scaling Factor : 0.02
- **Stage two: Invariance to Site**
  - Invariance Loss $\mathcal{MMD}$
    Scaling Factor : 0.1
  - Prediction Loss $\|Y - h(\Phi(\ell))\|^2$
    Scaling Factor : 1.0
  - Reconstruction Loss $\|\ell - \Psi(\Phi(\ell))\|^2$
    Scaling Factor : 0.1

We refer the reader to Algorithm 1 and Section 3 of the main paper for the details on the notations used above.