Site Confounding in Multi-Site Autism Classification from Resting-State fMRI

A Technical Report on ASD versus Typical-Control Classification in the ABIDE I Preprocessed Derivative (CPAC / CC200), Contrasting Subject-Stratified and Leave-Site-Out Cross-Validation

July 11, 2026

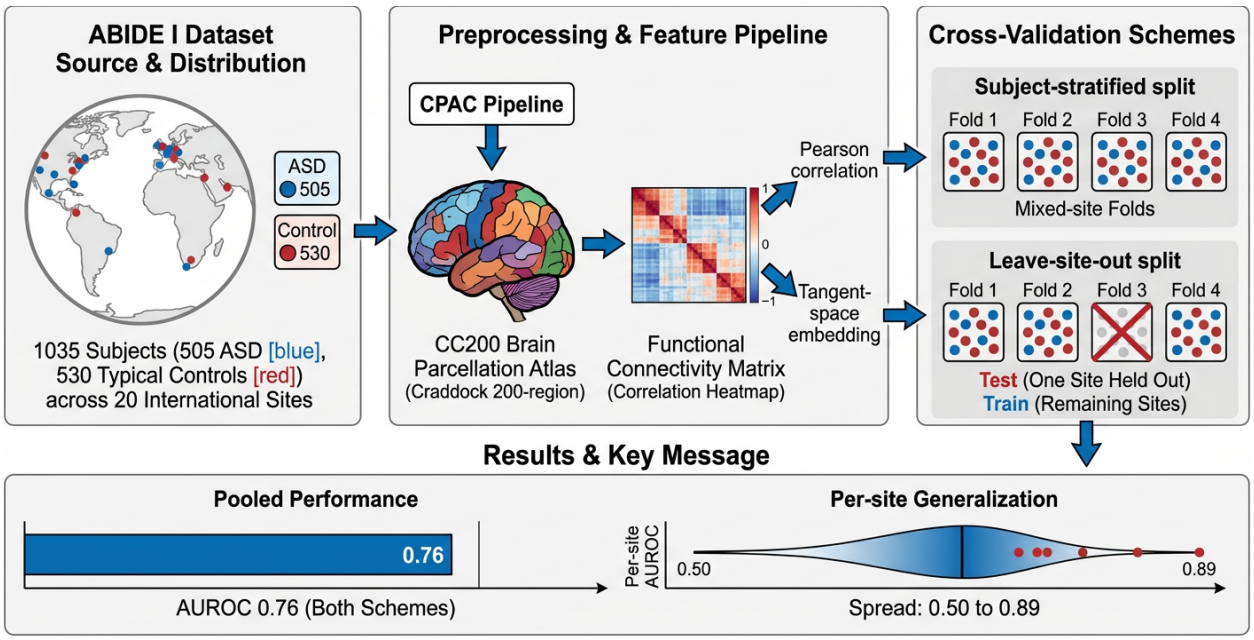

Figure 1: Graphical abstract. Resting-state functional MRI from 1,035 participants (505 autism spectrum disorder [ASD], 530 typical controls [TC]) across 20 site labels of the ABIDE I preprocessed release (CPAC pipeline, CC200 parcellation) was reduced to functional-connectivity features (Pearson correlation and tangent-space embedding). Regularized linear classifiers were evaluated under two schemes: a subject-stratified split, in which every site is represented in both training and test partitions, and a leave-site-out split, in which an entire site is held out. Aggregate (pooled) performance is essentially preserved between the two schemes (AUROC ≈ 0.76), but per-site generalization varies widely (site-wise AUROC 0.50–0.89), revealing that the cost of site confounding manifests as *variance* across scanners rather than a uniform drop in the mean.

Abstract

Machine-learning classification of autism spectrum disorder (ASD) from resting-state functional MRI (rs-fMRI) is complicated by the aggregation of data across many acquisition sites, where scanner, protocol, and cohort differences confound the diagnostic signal. Using the ABIDE I preprocessed release from the Preprocessed Connectomes Project, we built a binary classifier for

ASD versus typical control (TC) on the Configurable Pipeline for the Analysis of Connectomes (CPAC) derivative `rois_cc200` (band-pass filtered, no global signal regression) parcellated with the Craddock 200 (CC200) atlas. Functional connectivity was represented both by vectorized Pearson correlation matrices (19,900 features) and by tangent-space embeddings of the connectivity manifold (20,100 features), and classified with ℓ_2-regularized logistic regression and a linear support-vector classifier. We report the area under the receiver operating characteristic curve (AUROC), balanced accuracy, and macro-averaged F_1 under two evaluation schemes: subject-stratified 5-fold cross-validation and leave-site-out (LSO) cross-validation across the 20 site labels ($N = 1035$; 505 ASD, 530 TC). The best configuration (Tangent + L2 logistic regression) achieved a subject-stratified AUROC of 0.760 ± 0.014 and, when evaluated with LSO, a pooled AUROC of 0.761. Across all four model configurations the pooled AUROC changed by less than 0.5% between schemes, and balanced accuracy and macro-F_1 were, if anything, marginally higher under LSO. This aggregate stability, however, conceals substantial heterogeneity: site-wise LSO AUROC for the best model ranged from 0.50 (chance, MAX_MUN) to 0.89 (KKI), with a standard deviation of 0.10, and site-wise performance was not significantly related to site sample size ($r = 0.23$, $p = 0.32$). We conclude that a population-level linear functional-connectivity signature for ASD generalizes across sites at the aggregate level, but that generalization to any individual unseen site remains unreliable—the practical signature of site confounding in this setting is elevated between-site variance rather than a collapse of the mean. All code and per-subject predictions under both schemes are released to support reproducibility.

Contents

1 Introduction

1.1 Clinical and neurobiological context

Autism spectrum disorder (ASD) is a heterogeneous neurodevelopmental condition defined by persistent difficulties in social communication and interaction together with restricted, repetitive patterns of behavior, interests, or activities [1]. It is common and its identified prevalence continues to rise: the most recent surveillance from the U.S. Centers for Disease Control and Prevention Autism and Developmental Disabilities Monitoring Network estimated that approximately one in 36 eight-year-old children met criteria for ASD in 2020, with a pronounced male predominance and substantial geographic variation across monitoring sites [2]. Diagnosis remains fundamentally behavioral, anchored by expert clinical judgment and standardized instruments such as the Autism Diagnostic Observation Schedule [3]. Because such assessment is time-consuming and requires specialized expertise, there has been sustained interest in whether objective neurobiological measures might complement diagnosis, stratify heterogeneity, or illuminate mechanism.

Resting-state functional MRI (rs-fMRI), which characterizes the temporal correlation structure of spontaneous blood-oxygen-level-dependent fluctuations, has been central to this effort. An influential body of work frames ASD as a disorder of distributed brain networks rather than of any single region. The cortical "underconnectivity" theory proposed that ASD is characterized by reduced long-range, particularly frontal–posterior, functional coordination, on the basis of task-based fMRI and corpus-callosum morphometry [4, 5], and early resting-state work reported diminished anterior–posterior coupling within the default-mode network [6]. The default-mode network has since become a focal point, with evidence for atypical within- and between-network connectivity that maps onto social-cognitive and self-referential difficulties [7]. Yet the literature is far from unanimous: reports of both hypo- and hyperconnectivity coexist, patterns depend on age and the specific networks examined, and methodological choices in preprocessing and parcellation materially influence conclusions [8, 9]. This tension has motivated a shift away from small, single-site studies toward large, openly shared, multi-site resources capable of supporting reproducible, adequately powered analyses.

1.2 ABIDE and the Preprocessed Connectomes Project

The Autism Brain Imaging Data Exchange (ABIDE I) is the canonical open resource of this kind. It aggregated structural and resting-state functional MRI together with rich phenotypic data from more than 1,100 individuals with ASD and typical controls (TC) contributed by roughly 17 international laboratories, establishing multi-site rs-fMRI as a practical substrate for autism research [10] and later extended by ABIDE II [11]. To lower the barrier to entry and to reduce analytic variability, the Preprocessed Connectomes Project (PCP) released preprocessed ABIDE I derivatives generated by standardized pipelines—including the Configurable Pipeline for the Analysis of Connectomes (CPAC) [12]—and provided region-of-interest time series under several published brain atlases, among them the Craddock 200 (CC200) functional parcellation derived by spatially constrained spectral clustering [13]. These derivatives underpin much of the subsequent machine-learning literature on ABIDE, from early large-scale connectivity classifiers [14] and systematic benchmarking of feature representations and confound handling [15], through deep-learning models [16] and graph-based approaches that treat the cohort as a population graph [17], to sparse cross-cohort biomarkers [18].

1.3 Site as a confounder

A recurring and sobering theme of this literature is that classification performance on ABIDE is modest and difficult to generalize. Naive multi-site classifiers reach only about 60% accuracy [14], and even carefully engineered pipelines plateau near 66–67% under cross-site evaluation [15]. The central obstacle is that acquisition *site* is a powerful confounder. Sites differ in scanner hardware, field strength, pulse sequences, and acquisition parameters, and they differ equally in *who* they scan—in age distributions, sex ratios, symptom severity, and the balance of cases to controls [19]. These "measurement" and "sampling" biases produce systematic, non-biological structure in the connectivity features. A model can exploit this structure to identify *which site* a scan came from, and because the case/control ratio varies by site, learning site can spuriously inflate apparent diagnostic accuracy. Head motion—which differs systematically between clinical and control groups—compounds the problem by injecting reproducible, spatially structured artifact into functional connectivity [20, 21]. Statistical harmonization methods such as ComBat, originally developed for genomic batch effects [22] and adapted to neuroimaging measures [23, 24], can attenuate measurement bias, but they cannot manufacture generalization to a scanner never seen during training.

The way performance is estimated therefore determines what a reported number means. Under a *subject-stratified* split, subjects from every site appear in both the training and test partitions, so a model may legitimately (or illegitimately) rely on site-specific regularities that are shared across the split. Under a *leave-site-out* (LSO) split, an entire site is withheld for testing, so the model must transfer to a new scanner and a new cohort—a setting that mirrors real-world deployment and is a far more stringent test of a genuinely diagnostic signal. The gap between these two estimates is a direct, interpretable measure of how much a model's apparent skill depends on site-bound information, and the broader machine-learning-for-science literature has repeatedly warned that failing to separate such structure between training and test data is a leading cause of over-optimistic, non-reproducible findings [25, 26].

1.4 Objectives

This report builds an ASD-versus-TC classifier on a named, reproducible ABIDE I derivative and rigorously contrasts the two evaluation regimes. Our specific objectives are: (i) to specify exactly which preprocessed derivative, atlas, and pipeline are used, and to profile the demographic and site structure of the sample; (ii) to construct two complementary functional-connectivity feature representations (Pearson correlation and tangent-space embedding) and classify them with regularized linear models; (iii) to report AUROC, balanced accuracy, and macro-F_1 under both subject-stratified and leave-site-out cross-validation; (iv) to quantify precisely how much performance changes between the two schemes, both in aggregate and site by site; and (v) to release per-subject predictions and code. Our central empirical finding—that pooled performance is largely preserved across schemes while between-site variance is large—nuances the common expectation of a uniform "performance drop" and clarifies where the true cost of site confounding lies.

2 Methods

Figure 2 summarizes the end-to-end analysis pipeline, from the ABIDE I preprocessed derivative through feature extraction to the two evaluation schemes. All analyses were implemented in Python using `nilearn` [27] for connectivity estimation and `scikit-learn` [28] for classification and cross-validation. Random seeds were fixed (seed = 42) throughout for reproducibility.

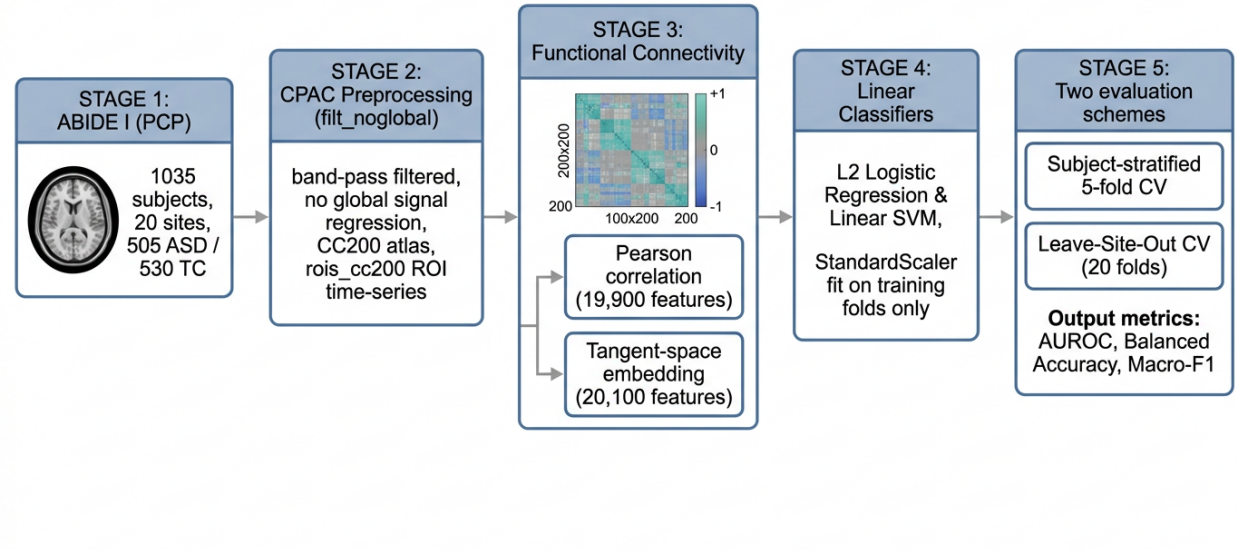

Figure 2: Analysis pipeline. The ABIDE I preprocessed derivative (CPAC pipeline, `filt_noglobal` strategy, CC200 atlas) supplies regional resting-state time series for 1,035 subjects. Functional connectivity is estimated as vectorized Pearson correlation (19,900 features) and as a tangent-space embedding of the connectivity manifold (20,100 features). Regularized linear classifiers (ℓ_2 logistic regression and linear SVM) are trained with feature standardization fit on training data only, and evaluated under two cross-validation schemes reporting AUROC, balanced accuracy, and macro-F_1.

2.1 Data source and derivative selection

We used the ABIDE I preprocessed release distributed by the Preprocessed Connectomes Project, retrieved programmatically through the `nilearn` ABIDE fetcher. We selected a single, explicitly named derivative to eliminate ambiguity: the **CPAC** pipeline [12] with the **`filt_noglobal`** nuisance strategy—that is, with band-pass temporal filtering applied and *without* global signal regression—and the **CC200** (Craddock 200) functional parcellation [13]. The corresponding per-subject artifact is the region-of-interest mean time-series file `{FILE_ID}_rois_cc200.1D`, a matrix of T time points by 200 regions. Each subject's diagnostic label (`DX_GROUP`: 1 = ASD, 2 = TC) and acquisition site (`SITE_ID`) were taken from the PCP phenotypic table `Phenotypic_V1_0b_preprocessed1.csv`. Band-pass filtering without global signal regression is a widely used and conservative default; global signal regression remains controversial because it can introduce spurious anticorrelations and is entangled with motion, so we retained the non-GSR strategy.

2.2 Cohort definition and quality control

Starting from the phenotypic table, we retained every subject for whom the CC200 time-series file was present, non-empty, and correctly shaped (exactly 200 ROI columns). Any non-finite values in a time series were replaced with zero before connectivity estimation; in practice these were negligible. The resulting analysis cohort comprised **1,035 subjects** (505 ASD, 530 TC) distributed across **20 site labels**. The 20 labels correspond to the roughly 17 contributing laboratories, several of which contributed more than one acquisition sample and are recorded under separate identifiers in the PCP release (for example `LEUVEN_1`/`LEUVEN_2`, `UCLA_1`/`UCLA_2`, and `UM_1`/`UM_2`); we treat each such label as a distinct "site" for the leave-site-out analysis because it corresponds to a distinct acquisition context. The full per-site composition is given in Table 1 and Figure 3.

Table 1: Sample composition by acquisition site. Counts of ASD and TC subjects, mean age (years), and sex distribution. Sites are ordered by total sample size. The final row summarizes the full cohort. Age is mean ± SD at scan.

Site	ASD	TC	Total	Age (ASD/TC)	% Male
NYU	75	100	175	14.7 / 15.7	79.4
UM_1	53	53	106	12.8 / 14.1	76.4
UCLA_1	41	31	72	13.1 / 13.3	86.1
USM	46	25	71	23.5 / 21.3	100.0
YALE	28	28	56	12.8 / 12.7	71.4
PITT	29	27	56	19.0 / 18.9	85.7
MAX_MUN	24	28	52	26.1 / 24.6	92.3
KKI	20	28	48	10.0 / 10.0	75.0
TRINITY	22	25	47	16.8 / 17.1	100.0
CALTECH	19	18	37	27.4 / 28.0	78.4
SDSU	14	22	36	14.7 / 14.2	80.6
UM_2	13	21	34	14.9 / 16.7	94.1
OLIN	19	15	34	16.5 / 16.7	85.3
LEUVEN_2	15	19	34	13.9 / 14.2	76.5
SBL	15	15	30	35.0 / 33.7	100.0
LEUVEN_1	14	15	29	21.9 / 23.3	100.0
CMU	14	13	27	26.4 / 26.9	77.8
OHSU	12	14	26	11.4 / 10.1	100.0
UCLA_2	13	13	26	12.7 / 12.3	92.3
STANFORD	19	20	39	10.0 / 10.0	79.5
All sites	**505**	**530**	**1,035**	**17.1 / 16.8**	**84.8**

The demographic profile makes the confounding problem concrete. Site sample sizes span an order of magnitude, from 26 (OHSU, UCLA_2) to 175 (NYU). Mean age ranges from roughly 10 years (KKI, STANFORD, OHSU) to 35 years (SBL). Several sites are entirely male (USM, TRINITY, SBL, LEUVEN_1, OHSU), while others include appreciable numbers of female participants. The ASD:TC ratio is also site-dependent, ranging from control-heavy (NYU: 75 ASD versus 100 TC) to case-heavy (USM: 46 ASD versus 25 TC; UCLA_1: 41 versus 31). Because diagnosis, age, sex, and site co-vary in this way, any classifier trained on pooled data can partly "solve" the diagnostic task by proxy through site- or demographic-linked signal—precisely the mechanism the leave-site-out analysis is designed to expose.

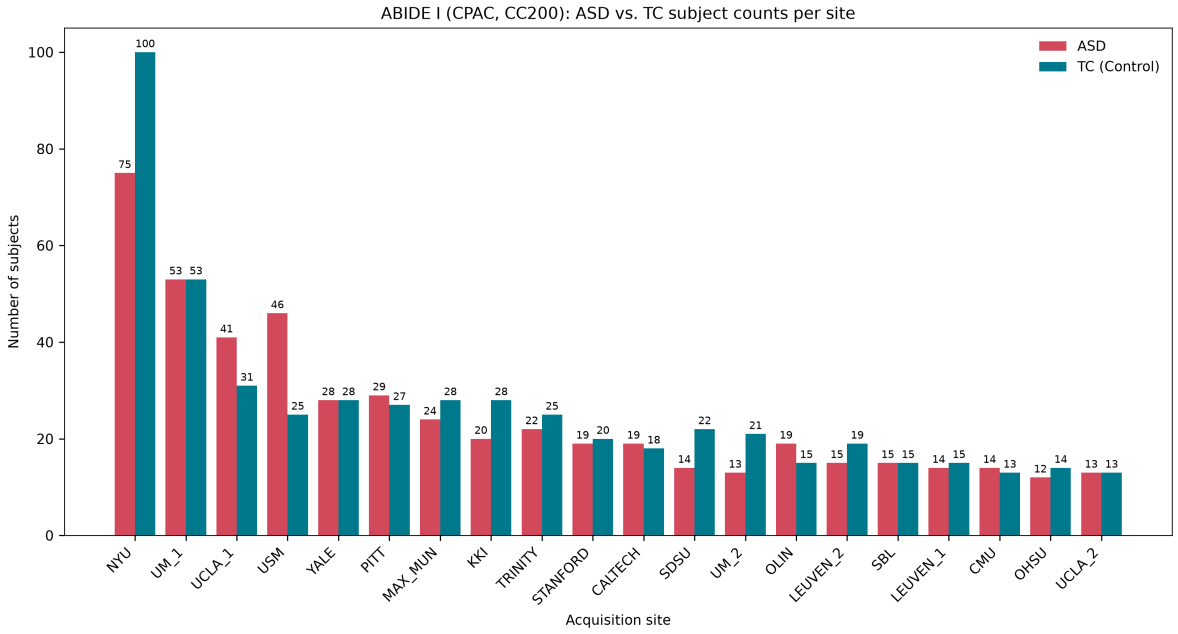

Figure 3: ASD versus TC subject counts across the 20 ABIDE I site labels (CPAC/CC200 cohort, $N = 1035$). Site sample sizes are highly imbalanced and the case/control ratio varies markedly by site, illustrating the coupling between acquisition site and diagnostic composition that makes site a strong confounder.

2.3 Feature extraction: functional connectivity

For each subject we estimated a 200×200 functional-connectivity matrix from the CC200 regional time series and derived two complementary vectorized feature representations.

Pearson correlation. We computed the Pearson correlation between every pair of regional time series and vectorized the upper triangle (excluding the diagonal), yielding $\binom{200}{2} = 19{,}900$ features per subject. This is the classical, directly interpretable connectivity representation, in which each feature is the strength of coupling between a specific pair of brain regions.

Tangent-space embedding. Functional-connectivity matrices are symmetric positive-definite and live on a curved Riemannian manifold, so Euclidean operations on raw correlations are geometrically distorted. Following the group-covariance framework of Varoquaux et al. [26] and its benchmarking by Dadi et al. [29], we projected each subject's covariance matrix into the tangent space at the group geometric mean and vectorized the result (including the diagonal), yielding $\binom{201}{2} = 20{,}100$ features. Tangent-space parametrization consistently ranks among the strongest connectivity representations for predictive modeling and typically outperforms raw correlations [29], which motivates its inclusion alongside the Pearson baseline. As detailed in Section 4.5, the tangent reference (group mean) was estimated once on the full cohort; we discuss the implications of this choice for the leave-site-out estimate.

2.4 Classifiers

We deliberately restricted attention to regularized *linear* classifiers, which the benchmarking literature identifies as strong, stable, and appropriately conservative for high-dimensional connectivity features on samples of this size [15, 29]. Two estimators were used: (i) ℓ_2-regularized logistic regression (`LogisticRegression`, `C`= 1.0, L-BFGS solver, up to 5,000 iterations), which yields calibrated class probabilities; and (ii) a linear support-vector classifier (ℓ_2 penalty, squared-hinge loss, `C`= 1.0, up to 10,000 iterations), whose signed distance to the separating hyperplane serves as the decision score. Crossing two feature representations with two classifiers gives four model configurations. To prevent information leakage, feature standardization (`StandardScaler`) was fit on the training partition of each fold only and then applied to the corresponding test partition; standardization was never fit on held-out data.

2.5 Evaluation schemes

The two cross-validation schemes are contrasted schematically in Figure 4.

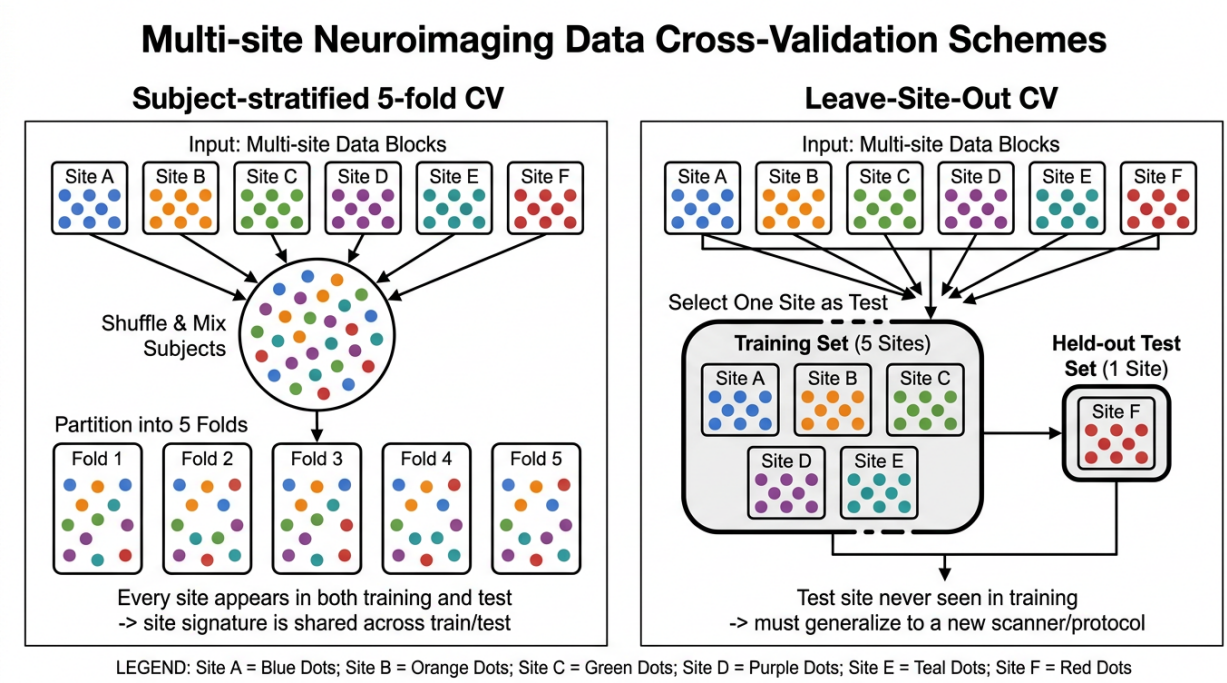

Figure 4: The two cross-validation schemes. *Left:* subject-stratified 5-fold CV mixes subjects from all sites, so every site is represented in both training and test partitions and any site-bound regularity is shared across the split. *Right:* leave-site-out CV withholds an entire site for testing, forcing the model to generalize to a scanner and cohort it has never seen. The gap between the two estimates measures how much apparent performance depends on site-bound information.

Subject-stratified split. We used stratified 5-fold cross-validation (`StratifiedKFold`, shuffled, seed = 42), which preserves the overall ASD:TC ratio in each fold but ignores site membership. Subjects from a given site are distributed across all folds, so the training set always contains examples from every site present in the test set. This is the standard, optimistic evaluation regime widely reported in the literature.

Leave-site-out (LSO) split. We used leave-one-group-out cross-validation (`LeaveOneGroupOut`) with `SITE_ID` as the grouping variable, producing 20 folds. In each fold, all subjects from one site are held out for testing while the model is trained on the remaining 19 sites. No subject from the test site—and, critically, no information about that site's scanner or cohort—is available during training. This directly emulates deployment to a new imaging center.

2.6 Performance metrics

For every configuration and scheme we report three complementary metrics. **AUROC** (area under the receiver operating characteristic curve) is a threshold-free measure of ranking quality and is insensitive to class prevalence, making it comparable across sites of differing composition. **Balanced accuracy**, the mean of sensitivity and specificity, corrects for class imbalance at the default decision threshold. **Macro-F_1**, the unweighted mean of the per-class F_1 scores, summarizes precision/recall balance treating both classes equally. For the subject-stratified scheme we report the mean ± standard deviation across the 5 folds. For the LSO scheme we report both the *pooled* metric (computed once on the concatenated out-of-fold predictions across all subjects, so that each subject contributes exactly one prediction made by a model that never saw that subject's site) and the *site-wise* metric (computed within each held-out site and then averaged across sites). The pooled metric is the fair, dataset-level analogue of the stratified estimate; the site-wise summary exposes between-site heterogeneity. Out-of-fold predictions were retained for every subject under both schemes, enabling the per-subject deliverable described in Section 3.5.

3 Results

3.1 Subject-stratified classification

Under subject-stratified 5-fold cross-validation, all four configurations classified ASD versus TC well above chance (Table 2). AUROC ranged from 0.735 (Pearson + linear SVM) to 0.760 (both tangent-based models). The tangent-space representation modestly outperformed raw Pearson correlation, consistent with prior benchmarking [29]: the Tangent + L2 logistic regression reached an AUROC of 0.760 ± 0.014, a balanced accuracy of 0.673 ± 0.036, and a macro-F_1 of 0.672 ± 0.035, while the tangent linear SVM achieved a comparable AUROC of 0.760 ± 0.010 with slightly higher balanced accuracy (0.678 ± 0.012). These values sit squarely within the 0.70–0.76 AUROC band that characterizes well-conducted ABIDE classification and match the ∼66–67% accuracy ceiling reported by careful pipelines [14, 15]. The pooled out-of-fold ROC curves for the best model per representation are shown in Figure 5.

Table 2: Subject-stratified 5-fold cross-validation. Mean ± SD across folds for the four model configurations. Best value per metric in **bold**.

Feature	Classifier	AUROC	Balanced accuracy	Macro-F_1
Pearson	Logistic regression (ℓ_2)	0.747 ± 0.012	0.676 ± 0.016	0.676 ± 0.015
Pearson	Linear SVM (ℓ_2)	0.735 ± 0.015	0.666 ± 0.018	0.665 ± 0.017
Tangent	Logistic regression (ℓ_2)	$\mathbf{0.760 \pm 0.014}$	0.673 ± 0.036	0.672 ± 0.035
Tangent	Linear SVM (ℓ_2)	$\mathbf{0.760 \pm 0.010}$	$\mathbf{0.678 \pm 0.012}$	$\mathbf{0.678 \pm 0.013}$

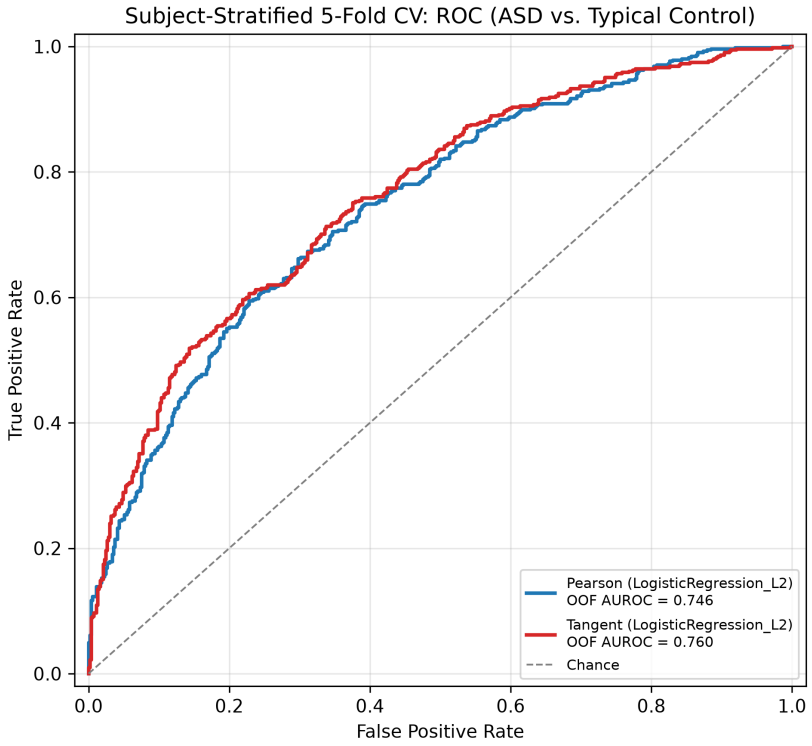

Figure 5: Subject-stratified ROC curves. Pooled out-of-fold ROC for the best classifier per feature representation. The tangent-space embedding (red, AUROC = 0.760) modestly exceeds Pearson correlation (blue, AUROC = 0.746); both clearly exceed chance (dashed diagonal).

3.2 Leave-site-out classification

Under leave-site-out cross-validation, in which every prediction is made by a model that never saw the test subject's site, pooled performance was strikingly similar to the stratified estimates (Table 3). The Tangent + L2 logistic regression reached a pooled AUROC of 0.761, a pooled balanced accuracy of 0.684, and a pooled macro-F_1 of 0.684. The tangent linear SVM was essentially identical (pooled AUROC 0.760), and even the raw-Pearson models held up (0.747 and 0.732). Pooled ROC curves are shown in Figure 6. That the aggregate signal survives a strict site hold-out indicates that at least part of the learned functional-connectivity signature reflects diagnosis-linked structure that is shared across scanners, rather than being entirely site-bound artifact.

Table 3: Leave-site-out cross-validation. Pooled out-of-fold metrics (each subject predicted by a model blind to its site) and site-wise metrics (computed within each held-out site, then averaged ± SD across the 20 sites). Best pooled value per metric in **bold**.

Feature / Classifier		AUROC	Balanced acc.	Macro-F_1
Pearson / Logistic reg.	pooled	0.747	0.685	0.685
	site-wise	0.745 ± 0.085	0.677 ± 0.073	0.670 ± 0.075
Pearson / Linear SVM	pooled	0.732	0.678	0.678
	site-wise	0.733 ± 0.091	0.671 ± 0.085	0.665 ± 0.087
Tangent / Logistic reg.	pooled	**0.761**	0.684	0.684
	site-wise	0.765 ± 0.104	0.671 ± 0.091	0.655 ± 0.106
Tangent / Linear SVM	pooled	0.760	**0.687**	**0.687**
	site-wise	0.757 ± 0.105	0.674 ± 0.093	0.659 ± 0.103

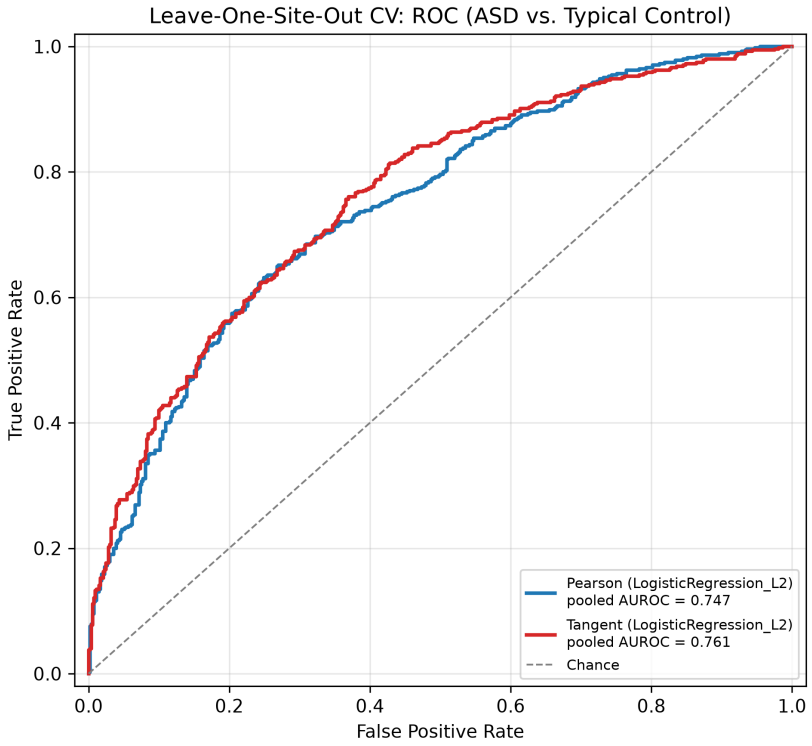

Figure 6: Leave-site-out ROC curves. Pooled out-of-fold ROC for the best classifier per feature representation under leave-site-out cross-validation. Pooled AUROC (tangent = 0.761, Pearson = 0.747) is virtually unchanged from the subject-stratified estimates in Figure 5.

3.3 Contrasting the two schemes: quantifying the change

The user's central question is how much performance drops when moving from a subject-stratified to a leave-site-out split. The direct, aligned comparison of stratified means against LSO-pooled metrics is presented in Table 4 and visualized in Figure 7. The answer, for this derivative and these models, is unexpected but clear: **at the aggregate level there is essentially no drop**. For the Tangent + L2 logistic regression, pooled AUROC changed by −0.0003 (a relative change of −0.04%),

balanced accuracy by +0.011 (+1.7%, i.e. LSO was slightly *higher*), and macro-F_1 by +0.012 (+1.7%). Across all four configurations the change in pooled AUROC lay within ±0.5% (from +0.43% for Pearson + SVM to −0.04% for tangent + logistic regression), and balanced accuracy and macro-F_1 were consistently a percentage point or two higher under LSO rather than lower. In other words, the strict site hold-out did not degrade dataset-level discrimination at all.

Table 4: Performance change from subject-stratified to leave-site-out. Stratified mean versus LSO-pooled metric, with absolute and relative change. A *positive* relative change denotes a drop under LSO; *negative* values (in parentheses) denote that LSO exceeded the stratified estimate. The best model (Tangent + L2 logistic regression) is highlighted.

Configuration	Metric	Stratified	LSO (pooled)	Abs. change	Rel. change
Pearson + LR	AUROC	0.747	0.747	+0.000	(−0.04%)
Pearson + LR	Balanced acc.	0.676	0.685	+0.009	(−1.34%)
Pearson + LR	Macro-F_1	0.676	0.685	+0.009	(−1.40%)
Pearson + SVM	AUROC	0.735	0.732	−0.003	+0.43%
Pearson + SVM	Balanced acc.	0.666	0.678	+0.013	(−1.88%)
Pearson + SVM	Macro-F_1	0.665	0.678	+0.013	(−1.95%)
Tangent + LR	**AUROC**	**0.760**	**0.761**	**+0.000**	**(−0.04%)**
Tangent + LR	**Balanced acc.**	**0.673**	**0.684**	**+0.011**	**(−1.67%)**
Tangent + LR	**Macro-F_1**	**0.672**	**0.684**	**+0.012**	**(−1.72%)**
Tangent + SVM	AUROC	0.760	0.760	−0.000	+0.01%
Tangent + SVM	Balanced acc.	0.678	0.687	+0.008	(−1.25%)
Tangent + SVM	Macro-F_1	0.678	0.687	+0.009	(−1.31%)

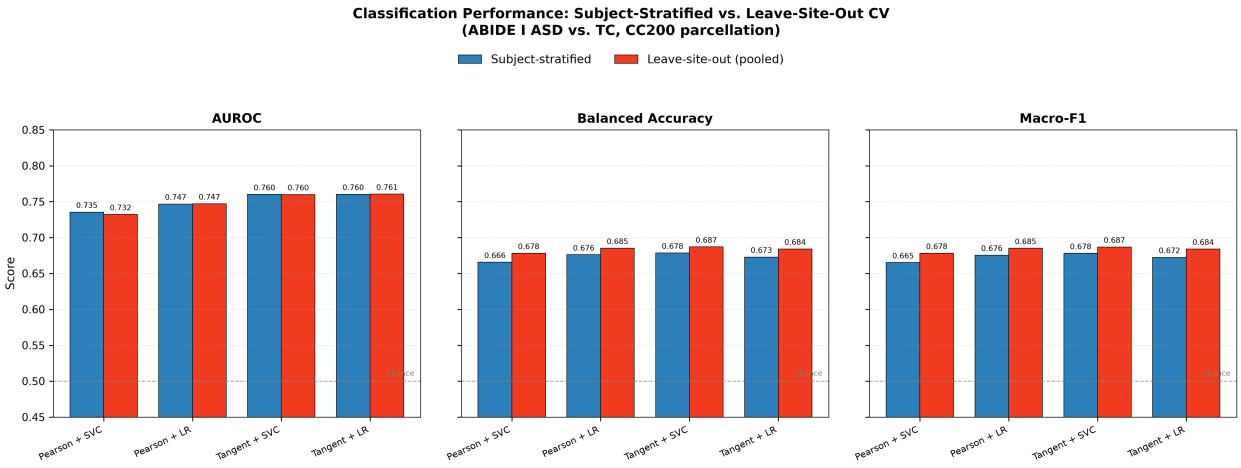

Figure 7: Subject-stratified versus leave-site-out (pooled) performance for all four model configurations, across AUROC, balanced accuracy, and macro-F_1. The near-perfect overlap of the blue (stratified) and red (LSO-pooled) bars shows that aggregate performance is preserved under a strict site hold-out; if anything, balanced accuracy and macro-F_1 edge slightly upward under LSO.

3.4 Where the cost of site confounding appears: between-site variance

The stability of the pooled metrics does not mean site is irrelevant—it means the cost of site confounding is expressed as *variance across sites* rather than as a shift in the mean. The site-wise

columns of Table 3 already hint at this: while the mean site-wise AUROC for the best model is 0.765, its standard deviation across sites is 0.10, roughly seven times the fold-to-fold standard deviation seen under stratification. Figure 8 plots each held-out site's AUROC against its sample size for the Tangent + L2 logistic regression. Site-wise LSO AUROC ranged from 0.50—indistinguishable from chance, at MAX_MUN—to 0.89 at KKI, with other low-performing sites including SBL (0.59), LEUVEN_1 (0.63), and OHSU (0.64), and other strong sites including LEUVEN_2 (0.88), UCLA_2 (0.86), USM (0.85), and UM_1 (0.85). Balanced accuracy showed the same spread, from 0.43 (MAX_MUN, i.e. below chance) to 0.81 (UCLA_2).

Crucially, this heterogeneity was *not* explained by how much data a site contributed. The correlation between a held-out site's sample size and its LSO AUROC was weak and statistically non-significant (Pearson $r = +0.23$, $p = 0.32$; OLS slope $\approx 6.9 \times 10^{-4}$ AUROC units per additional subject). The largest site (NYU, $n = 175$) achieved a solid but unremarkable AUROC of 0.82, while some of the smallest sites were among both the best (UCLA_2, $n = 26$, AUROC 0.86) and the worst (OHSU, $n = 26$, AUROC 0.64). Performance on a new site is therefore governed less by training-set size than by how representative the training sites are of the held-out site's particular scanner, protocol, and cohort—for instance MAX_MUN, an adult sample acquired on distinctive hardware with a wide age range, is poorly served by a training set dominated by pediatric and adolescent cohorts.

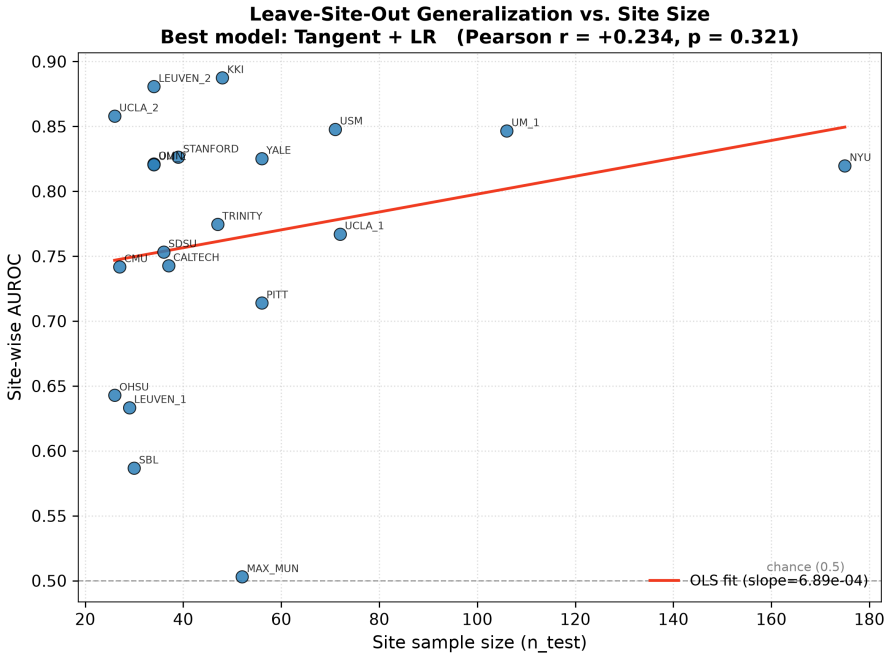

Figure 8: Site-wise leave-site-out generalization versus site sample size for the best model (Tangent + L2 logistic regression). Each point is one held-out site. Site-wise AUROC spans from chance (0.50, MAX_MUN) to 0.89 (KKI). The relationship with sample size is weak and non-significant ($r = +0.23$, $p = 0.32$; red OLS trend line), indicating that generalization to a new site depends on cohort/scanner representativeness rather than merely on how much training data is available.

3.5 Per-subject predictions

For the single best-performing model (Tangent + L2 logistic regression), we retained the out-of-fold prediction and decision score for every one of the 1,035 subjects under *both* schemes, and

merged them into a unified table, `per_subject_predictions.csv`. Each row records the subject identifier (`SUB_ID`), site (`SITE_ID`), true label, and, for each scheme, the predicted label and the model's probability of ASD. Subject-identifier alignment was verified across the phenotype table and both feature matrices before merging, guaranteeing that every prediction corresponds to the correct subject. This artifact supports fine-grained downstream analysis—for example, identifying subjects that both schemes misclassify (candidate label-noise or atypical cases) versus subjects that only the leave-site-out model misses (candidates for site-driven error)—and is released alongside the code.

4 Discussion

4.1 Principal findings

We set out to classify ASD versus typical control on a named, reproducible ABIDE I derivative (CPAC pipeline, `filt_noglobal` strategy, CC200 parcellation) and to quantify how estimated performance changes between a subject-stratified and a leave-site-out evaluation. Three findings stand out. First, regularized linear classifiers on functional connectivity achieve AUROC around 0.76, with tangent-space embedding slightly ahead of raw Pearson correlation—numbers fully consistent with the established ABIDE ceiling [14–16]. Second, and contrary to the common expectation that leave-site-out evaluation should sharply deflate performance, *pooled* AUROC, balanced accuracy, and macro-F_1 were essentially unchanged—within roughly one to two percent, and if anything marginally higher under LSO. Third, the aggregate stability conceals large between-site heterogeneity: site-wise LSO AUROC ranged from chance to 0.89 (SD $\approx$ 0.10), unrelated to site sample size.

4.2 Interpreting the (absent) aggregate drop

Why does the anticipated performance collapse not materialize in the pooled metrics? The most parsimonious explanation is that these linear models are *not* primarily exploiting a shared, cross-site site-signature to solve the diagnostic task. Had the stratified estimate been inflated by leakage of site-bound structure across the split, removing that structure via a site hold-out would have caused a marked drop; the absence of such a drop implies that the discriminative signal the models rely on is, to a useful degree, genuinely diagnosis-linked and transferable across scanners. This is a comparatively reassuring result: it suggests that a modest but real population-level functional-connectivity signature of ASD exists and survives strict site separation. It is also consistent with the benchmarking view that simple linear models on well-chosen connectivity representations are robust and resist overfitting to nuisance structure [26, 29].

At the same time, the result should not be over-read. The absolute performance is modest (AUROC $\approx$ 0.76; balanced accuracy $\approx$ 0.68), far from anything clinically deployable, and it reflects a population-average effect rather than an individual-level biomarker [1, 9]. The observation that balanced accuracy and macro-F_1 were slightly *higher* under LSO most plausibly reflects the mechanics of the two schemes rather than a genuine benefit of holding out sites: the pooled LSO metric aggregates predictions from 20 models each trained on $\sim$95% of the data, whereas the stratified metric averages five folds each trained on 80%, and small differences in effective training-set size and threshold behavior can produce differences of this magnitude.

4.3 The real cost of site: variance, not bias

The finding that matters most for practice is the between-site variance. A pooled AUROC of 0.76 is a statement about the dataset as a whole; it is *not* a promise about any particular clinic. Our leave-site-out analysis shows that if one were to deploy this classifier at a genuinely new site, the expected AUROC could be as high as 0.89 or as low as 0.50, and one cannot predict which merely from the amount of training data available [19]. This is the concrete, operational signature of site confounding in this setting: it inflates the *uncertainty* of any deployment estimate rather than uniformly biasing the average. Sites that are demographically or technically atypical relative to the training pool—most starkly MAX_MUN, an older, wide-age-range adult sample—are precisely those where the model degrades to chance. This pattern aligns with the decomposition of site effects into measurement and sampling bias [19] and with the broader recognition that head-motion and cohort differences inject structured, site-specific variance into connectivity [20, 21]. It also reinforces a methodological lesson emphasized across machine-learning-for-science: a single pooled cross-validation number, however carefully computed, can obscure the fact that generalization is highly uneven, and reporting the *distribution* of held-out-site performance is essential to honest appraisal [25, 26].

4.4 Comparison with the literature

Our stratified and pooled-LSO AUROCs ($\sim$0.75–0.76) are in close agreement with the reproducible-biomarker benchmark of Abraham et al. [15], which reported cross-validated accuracies near 66–67% on ABIDE, and exceed the 60% of the early leave-one-out replication of Nielsen et al. [14]. Deep-learning approaches such as Heinsfeld et al. [16] reported accuracies in the $\sim$70% range, and population-graph methods [17] somewhat higher, but comparisons across studies are complicated by differing subsets, atlases, and—most importantly—evaluation protocols. Our contribution is less a new accuracy record than a controlled, transparent contrast of the two evaluation regimes on one fixed derivative, which isolates the effect of the evaluation choice itself. The very high accuracies occasionally reported in the ASD literature (>90%) are, in light of our and others' analyses, most plausibly explained by evaluation choices that allow site- or subject-bound information to leak across the train/test boundary [25].

4.5 Limitations

Several limitations qualify these conclusions. **First and most important**, the tangent-space embedding was fit once on the full cohort in the feature-extraction stage: the group-mean reference matrix used for the tangent projection was estimated from all 1,035 subjects, including, under any given LSO fold, the held-out site. This constitutes a mild form of information leakage that could modestly inflate the tangent LSO estimates. Reassuringly, the raw-Pearson representation—which involves no group-level fitting and therefore no such leakage—exhibits the same qualitative pattern (pooled AUROC of 0.747 under both schemes), which indicates that our central conclusion is not an artifact of the tangent reference; nonetheless, a fully rigorous protocol would re-estimate the tangent reference within each training partition, and we flag this for future work. **Second**, the classifier used the default decision threshold of 0.5 for balanced accuracy and macro-F_1; per-site threshold calibration was not performed and might improve or further destabilize site-wise results. **Third**, we did not apply explicit harmonization (e.g. ComBat [22–24]) or motion-based subject exclusion beyond the PCP defaults; while the `filt_noglobal` strategy includes nuisance

regression, residual motion and site effects certainly remain. **Fourth**, we confined the analysis to a single atlas (CC200) and pipeline (CPAC); results can shift with parcellation granularity and preprocessing strategy. **Fifth**, ASD is profoundly heterogeneous [7, 8], and a single binary label collapses meaningful clinical variation; the ABIDE sample is also predominantly male, limiting generalization to females.

4.6 Future directions

The natural next steps follow directly from the limitations. Tangent references and all feature-level fitting should be nested strictly within LSO training folds to remove the residual leakage noted above. Harmonization approaches that explicitly separate measurement from sampling bias [19], or predictive ComBat variants designed for out-of-sample sites, should be evaluated for their ability to shrink the between-site variance we observe rather than merely to raise the pooled mean. Because generalization failures concentrate in demographically atypical sites, domain-adaptation and site-invariant representation-learning methods—and, more prosaically, matching or re-weighting the training pool to the target cohort—are promising. Finally, reporting practice should follow suit: for any multi-site neuroimaging classifier intended for eventual deployment, the full distribution of leave-site-out performance, not a single pooled figure, should be treated as the primary result.

4.7 Conclusion

On the ABIDE I CPAC/CC200 derivative, regularized linear classifiers distinguish ASD from typical controls with an AUROC of about 0.76, and this aggregate performance is preserved almost exactly when moving from a subject-stratified to a leave-site-out evaluation—an encouraging sign that the learned functional-connectivity signature is genuinely, if modestly, diagnostic rather than a site artifact. The important caveat is that this pooled stability masks large, sample-size-independent variability across sites, with held-out-site AUROC ranging from chance to 0.89. The practical cost of site confounding here is therefore not a uniform loss of accuracy but a large uncertainty in how a model will perform at any specific new center. Honest evaluation of multi-site neuroimaging classifiers must report that distribution, and building models whose *worst-case* site performance is acceptable—not merely whose average is—remains the central challenge for clinical translation.

Data and Code Availability

The ABIDE I preprocessed derivatives are openly available from the Preprocessed Connectomes Project (`http://preprocessed-connectomes-project.org/abide/`). All analysis code (six sequential Python scripts, `01_download_and_profile.py` through `06_package_results.py`, plus a master runner `run_all.sh`), the aggregated cross-validation metrics, the per-site metrics, and the unified per-subject prediction table (`per_subject_predictions.csv`, 1,035 rows, both schemes) are released with this report. The pipeline is deterministic (fixed seed = 42) and reproduces every figure and table herein.

References

[1] Catherine Lord, Mayada Elsabbagh, Gillian Baird, and Jeremy Veenstra-Vanderweele. Autism spectrum disorder. *The Lancet*, 392(10146):508–520, 2018. doi: 10.1016/S0140-6736(18)31129-2.

[2] Matthew J. Maenner, Zachary Warren, Ashley Robinson Williams, Esther Amoakohene, Amanda V. Bakian, Deborah A. Bilder, Maureen S. Durkin, Robert T. Fitzgerald, Sarah M. Furnier, Michelle M. Hughes, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2020. *MMWR Surveillance Summaries*, 72(2):1–14, 2023. doi: 10.15585/mmwr.ss7202a1.

[3] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H. Cook Jr, Bennett L. Leventhal, Pamela C. DiLavore, Andrew Pickles, and Michael Rutter. The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30(3):205–223, 2000. doi: 10.1023/A:1005592401947.

[4] Marcel Adam Just, Vladimir L. Cherkassky, Timothy A. Keller, Rajesh K. Kana, and Nancy J. Minshew. Functional and anatomical cortical underconnectivity in autism: Evidence from an fMRI study of an executive function task and corpus callosum morphometry. *Cerebral Cortex*, 17(4):951–961, 2007. doi: 10.1093/cercor/bhl006.

[5] Marcel Adam Just, Timothy A. Keller, Vicente L. Malave, Rajesh K. Kana, and Sashank Varma. Autism as a neural systems disorder: A theory of frontal-posterior underconnectivity. *Neuroscience & Biobehavioral Reviews*, 36(4):1292–1313, 2012. doi: 10.1016/j.neubiorev.2012.02.007.

[6] Vladimir L. Cherkassky, Rajesh K. Kana, Timothy A. Keller, and Marcel Adam Just. Functional connectivity in a baseline resting-state network in autism. *NeuroReport*, 17(16):1687–1690, 2006. doi: 10.1097/01.wnr.0000239956.45448.4c.

[7] Aarthi Padmanabhan, Charles J. Lynch, Marie Schaer, and Vinod Menon. The default mode network in autism. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2(6):476–486, 2017. doi: 10.1016/j.bpsc.2017.04.004.

[8] Jocelyn V. Hull, Lisa B. Dokovna, Zachary J. Jacokes, Carinna M. Torgerson, Andrei Irimia, and John Darrell Van Horn. Resting-state functional connectivity in autism spectrum disorders: A review. *Frontiers in Psychiatry*, 7:205, 2017. doi: 10.3389/fpsyt.2016.00205.

[9] Roma A. Vasa, Stewart H. Mostofsky, and Joshua B. Ewen. The implications of brain connectivity in the neuropsychology of autism. *Neuropsychology Review*, 24(1):16–31, 2014. doi: 10.1007/s11065-014-9250-0.

[10] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X. Castellanos, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Susan Y. Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014. doi: 10.1038/mp.2013.78.

[11] Adriana Di Martino, David O'Connor, Bosi Chen, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Joshua H. Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, et al. Enhancing

studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, 4:170010, 2017. doi: 10.1038/sdata.2017.10.

[12] R. Cameron Craddock, Sharad Sikka, Brian Cheung, Ranjeet Khanuja, Satrajit S. Ghosh, Chao-Gan Yan, Qingyang Li, Daniel Lurie, Joshua Vogelstein, Randal Burns, Stanley Colcombe, Maarten Mennes, Clare Kelly, Adriana Di Martino, Francisco X. Castellanos, and Michael Milham. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (C-PAC). *Frontiers in Neuroinformatics*, 7, 2013. doi: 10.3389/conf.fninf.2013.09.00042.

[13] R. Cameron Craddock, G. Andrew James, Paul E. Holtzheimer III, Xiaoping P. Hu, and Helen S. Mayberg. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8):1914–1928, 2012. doi: 10.1002/hbm.21333.

[14] Jared A. Nielsen, Brandon A. Zielinski, P. Thomas Fletcher, Andrew L. Alexander, Nicholas Lange, Erin D. Bigler, Janet E. Lainhart, and Jeffrey S. Anderson. Multisite functional connectivity MRI classification of autism: ABIDE replication study. *Frontiers in Human Neuroscience*, 7:599, 2013. doi: 10.3389/fnhum.2013.00599.

[15] Alexandre Abraham, Michael P. Milham, Adriana Di Martino, R. Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gaël Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, 147:736–745, 2017. doi: 10.1016/j.neuroimage.2016.10.045.

[16] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R. Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17:16–23, 2018. doi: 10.1016/j.nicl.2017.08.017.

[17] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease. *Medical Image Analysis*, 48:117–130, 2018. doi: 10.1016/j.media.2018.06.001.

[18] Noriaki Yahata, Jun Morimoto, Ryuichiro Hashimoto, Giuseppe Lisi, Kazuhisa Shibata, Yuki Kawakubo, Hitoshi Kuwabara, Miho Kuroda, Takashi Yamada, Fukuda Megumi, Hiroshi Imamizu, José E. Náñez Sr, Hidehiko Takahashi, Yasumasa Okamoto, Kiyoto Kasai, Nobumasa Kato, Yuka Sasaki, Takeo Watanabe, and Mitsuo Kawato. A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nature Communications*, 7:11254, 2016. doi: 10.1038/ncomms11254.

[19] Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, Naohiro Okada, Hirotaka Yamagata, Koji Matsuo, Ryuichiro Hashimoto, Go Okada, Yuki Sakai, Jun Morimoto, Jin Narumoto, Yasuhiro Shimada, Kiyoto Kasai, Nobumasa Kato, Hidehiko Takahashi, Yasumasa Okamoto, Saori C. Tanaka, Mitsuo Kawato, Okito Yamashita, and Hiroshi Imamizu. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLOS Biology*, 17(4):e3000042, 2019. doi: 10.1371/journal.pbio.3000042.

[20] Jonathan D. Power, Kelly A. Barnes, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. Spurious but systematic correlations in functional connectivity MRI networks arise

from subject motion. *NeuroImage*, 59(3):2142–2154, 2012. doi: 10.1016/j.neuroimage.2011.10.018.

[21] Jonathan D. Power, Mark Plitt, Timothy O. Laumann, and Alex Martin. Ridding fMRI data of motion-related influences: Removal of signals with distinct spatial and physical bases in multiecho data. *Proceedings of the National Academy of Sciences*, 115(9):E2105–E2114, 2018. doi: 10.1073/pnas.1720985115.

[22] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007. doi: 10.1093/biostatistics/kxj037.

[23] 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, Robert T. Schultz, Ragini Verma, and Russell T. Shinohara. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161:149–170, 2017. doi: 10.1016/j.neuroimage.2017.08.047.

[24] Jean-Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T. Shinohara. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120, 2018. doi: 10.1016/j.neuroimage.2017.11.024.

[25] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100857, 2023. doi: 10.1016/j.patter.2023.100857.

[26] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017. doi: 10.1016/j.neuroimage.2016.10.038.

[27] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8:14, 2014. doi: 10.3389/fninf.2014.00014.

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.48550/arXiv.1201.0490.

[29] Kamalaker Dadi, Mehdi Rahim, Alexandre Abraham, Darya Chyzhyk, Michael Milham, Bertrand Thirion, and Gaël Varoquaux. Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, 192:115–134, 2019. doi: 10.1016/j.neuroimage.2019.02.062.