KDD 2026 August 9–13, 2026 Jeju, Korea

ACCEPTED

Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining

KDD '26 • August 9–13, 2026 • Jeju Island, Republic of Korea

DOI: 10.1145/3770855.3819051

Beyond Accuracy: Counterfactual Auditing of fMRI Representations

Hyunsuk Chung[*]
hyunsuk.chung.1@unimelb.edu.au
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia

You Bin Lim[*]
youbin2010@snu.ac.kr
Child and Adolescent Psychiatry
Seoul National University Hospital
Seoul, Republic of Korea

Sohyun Kang[*]
sohyunkang@kist.re.kr
Brain Science Institute
Korea Institute of Science and Technology
Seoul, Republic of Korea

Eunice Jiun Chung
jiunchung523@gmail.com
Child and Adolescent Psychiatry
Seoul National University Hospital
Seoul, Republic of Korea

Junhyuk Woo
wjh601@kist.re.kr
Brain Science Institute
Korea Institute of Science and Technology
Seoul, Republic of Korea

Joonho Paik
paikjoonho123@kist.re.kr
Brain Science Institute
Korea Institute of Science and Technology
Seoul, Republic of Korea

Soyeon Caren Han[†]
caren.han@unimelb.edu.au
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia

Kyungreem Han[†]
khan@kist.re.kr
Brain Science Institute
Korea Institute of Science and Technology
Seoul, Republic of Korea

Bung-Nyun Kim[†]
kbn1@snu.ac.kr
Child and Adolescent Psychiatry
Seoul National University Hospital
Seoul, Republic of Korea

[*]These authors contributed equally to this research. [†]Corresponding author.

Abstract

Self-supervised learning is increasingly applied to functional neuroimaging, yet existing evaluations rely mainly on downstream prediction accuracy, offering limited insight into whether learned representations preserve biologically meaningful structure. We propose **Generative Manifold Auditing (GMA)**, a framework that probes the intrinsic structural sensitivity of latent representations via ROI-level counterfactual interventions. Across the multi-site ABIDE benchmark, a denoising autoencoder (DAE) consistently exhibits higher sensitivity to biologically grounded perturbations than contrastive and masked autoencoder (MAE) baselines, which show flattened structural sensitivity despite apparent stability. Applying GMA to Autism Spectrum Disorder (ASD), we observe heterogeneous, lobe-dependent sensitivity deviations relative to typically developing controls, with consistent distributional patterns on an independent Korean Brain Network (KBN) cohort. Longitudinal analysis further reveals a dissociation between behavioral improvement and neural sensitivity dynamics: while all three longitudinal subjects show reduced K-CARS scores after intervention, their neural sensitivity trajectories remain region- and subject-specific. GMA reframes representation evaluation from prediction-centric benchmarks to mechanism-aware scientific auditing.

CCS Concepts

• **Applied computing** → Computational biology.

Keywords

fMRI, intervention, reproducibility, representation, trustworthy

Publication Information

VENUE	PUBLISHER
KDD '26, Volume 2	ACM, New York, NY, USA
DATE	LENGTH
August 9–13, 2026	10 pages
LOCATION	DOI
Jeju Island, Republic of Korea	10.1145/3770855.3819051

ACM Reference Format

Hyunsuk Chung, You Bin Lim, Sohyun Kang, Eunice Jiun Chung, Junhyuk Woo, Joonho Paik, Soyeon Caren Han, Kyungreem Han, and Bung-Nyun Kim. 2026. *Beyond Accuracy: Counterfactual Auditing of fMRI Representations*. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26)*, August 9–13, 2026, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 10 pages. DOI: 10.1145/3770855.3819051.

Beyond Accuracy: Counterfactual Auditing of fMRI Representations

Hyunsuk Chung*
hyunsuk.chung.1@unimelb.edu.au
School of Computing and Information Systems, University of Melbourne
Melbourne, Australia

You Bin Lim*
youbin2010@snu.ac.kr
Child and Adolescent Psychiatry
Seoul National University Hospital
Seoul, Republic of Korea

Sohyun Kang*
sohyunkang@kist.re.kr
Brain Science Institute, Korea Institute of Science and Technology
Seoul, Republic of Korea

Eunice Jiun Chung
jiunchung523@gmail.com
Child and Adolescent Psychiatry
Seoul National University Hospital
Seoul, Republic of Korea

Junhyuk Woo
wjh601@kist.re.kr
Brain Science Institute, Korea Institute of Science and Technology
Seoul, Republic of Korea

Joonho Paik
paikjoonho123@kist.re.kr
Brain Science Institute, Korea Institute of Science and Technology
Seoul, Republic of Korea

Soyeon Caren Han†
caren.han@unimelb.edu.au
School of Computing and Information Systems, University of Melbourne
Melbourne, Australia

Kyungreem Han†
khan@kist.re.kr
Brain Science Institute, Korea Institute of Science and Technology
Seoul, Republic of Korea

Bung-Nyun Kim†
kbn1@snu.ac.kr
Child and Adolescent Psychiatry
Seoul National University Hospital
Seoul, Republic of Korea

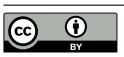
Abstract

Self-supervised learning is increasingly applied to functional neuroimaging, yet existing evaluations rely mainly on downstream prediction accuracy, offering limited insight into whether learned representations preserve biologically meaningful structure. We propose **Generative Manifold Auditing (GMA)**, a framework that probes the intrinsic structural sensitivity of latent representations via ROI-level counterfactual interventions. Across the multi-site ABIDE benchmark, a denoising autoencoder (DAE) consistently exhibits higher sensitivity to biologically grounded perturbations than contrastive and masked autoencoder (MAE) baselines, which show flattened structural sensitivity despite apparent stability. Applying GMA to Autism Spectrum Disorder (ASD), we observe heterogeneous, lobe-dependent sensitivity deviations relative to typically developing controls, with consistent distributional patterns on an independent Korean Brain Network (KBN) cohort. Longitudinal analysis further reveals a dissociation between behavioral improvement and neural sensitivity dynamics: while all three longitudinal subjects show reduced K-CARS scores after intervention, their neural sensitivity trajectories remain region- and subject-specific. GMA reframes representation evaluation from prediction-centric benchmarks to mechanism-aware scientific auditing.

CCS Concepts

• **Applied computing** → **Computational biology**.

* These authors contributed equally to this research.
† Corresponding author.



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3819051

Keywords

fMRI, intervention, reproducibility, representation, trustworthy

ACM Reference Format:
Hyunsuk Chung, You Bin Lim, Sohyun Kang, Eunice Jiun Chung, Junhyuk Woo, Joonho Paik, Soyeon Caren Han, Kyungreem Han, and Bung-Nyun Kim. 2026. Beyond Accuracy: Counterfactual Auditing of fMRI Representations. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3770855.3819051

1 Introduction

Self-supervised learning has become a dominant paradigm for extracting representations from high-dimensional scientific data, including neuroimaging, climate simulations, and molecular systems. In these domains, models are increasingly expected not only to achieve high downstream performance, but also to encode *scientifically meaningful structure* that supports discovery, interpretation, and mechanistic reasoning. However, current evaluation protocols remain largely *prediction-centric*, providing limited insight into whether learned representations preserve underlying system organization or instead become insensitive to structured variation.

This limitation is particularly pronounced in functional neuroimaging. Autism spectrum disorder (ASD) is a neurodevelopmental condition for which early and objective diagnosis remains challenging, as current clinical assessments rely primarily on behavioral observation. Resting-state functional MRI (fMRI) has therefore been explored as a complementary modality, as it enables direct examination of large-scale functional connectivity patterns that are not accessible through behavioral measures alone. Prior studies have consistently reported atypical functional connectivity in ASD, including altered long-range interactions and inefficient network organization at the systems level. Although no single fMRI-based

biomarker has been established, these converging findings suggest that functional network alterations play a substantial role in ASD pathophysiology. Recent machine-learning approaches have leveraged fMRI-derived connectivity features for ASD classification and reported high accuracy under controlled experimental settings. Yet high predictive performance alone does not guarantee that learned representations retain biologically grounded network structure. Objectives that prioritize invariance or aggressive compression may produce representations that appear stable across perturbations while discarding fine-grained structural dependencies that are essential for scientific interpretation. Such representations can perform well on benchmarks while offering limited insight into the underlying biological system, as apparent stability may mask a loss of fine-grained structural sensitivity.

To address this gap, we propose **Generative Manifold Auditing (GMA)**, a framework for evaluating self-supervised representations through their *sensitivity to structured perturbations*. Rather than relying on downstream labels or task-specific accuracy, GMA examines whether a latent representation responds differentially to biologically meaningful changes versus unstructured variation. This perspective shifts evaluation from prediction performance to *representational behavior*, enabling assessment of whether a model preserves organization that aligns with known system structure. Applying this framework to resting-state fMRI, we show that different self-supervised objectives induce markedly different representational properties. Some models yield representations that are highly stable yet largely insensitive to structured perturbations, indicating flattened structural sensitivity. In contrast, generative inductive biases preserve structured sensitivity that generalizes across acquisition sites and datasets, revealing biologically meaningful variation that is invisible to standard metrics. Beyond cross-sectional analysis, GMA also supports longitudinal auditing of representation dynamics. By examining how sensitivity profiles evolve over time, this framework provides a lens for studying whether changes in behavioral outcomes are accompanied by corresponding changes in neural representations, or whether latent deviations persist despite similar clinical trajectories.

Overall, our work makes three contributions:
1) we show that stable self-supervised representations may still fail to preserve biologically meaningful structural sensitivity, **2)** we introduce Generative Manifold Auditing as a general, model-agnostic framework for probing intrinsic structural sensitivity beyond downstream accuracy, and **3)** we demonstrate, across ABIDE and an independent KBN cohort, that structural sensitivity profiles reveal heterogeneous ASD-related network deviations and longitudinal neural dynamics that are not captured by behavioral scores alone.

2 Related Work

fMRI Analysis for ASD. Previous fMRI studies on ASD have extensively investigated functional brain organization using connectivity-based analyses [6, 17, 26]. By modeling functional interactions between brain regions, these studies have enabled group-level comparisons, most commonly at the edge level. Across studies, atypical functional connectivity patterns have been reported in ASD, including altered balance between long-range and short-range interactions and inefficient large-scale network organization. More recent work has extended connectivity analyses to the network level [4, 23], allowing the characterization of large-scale brain organization and its alterations in ASD. In this context, several studies have emphasized the importance of identifying hub nodes, as these highly connected regions play a central role in information integration and may be selectively disrupted in ASD [12, 19].

Deep Learning on Functional Connectomes for ASD. Deep learning for ASD biomarker discovery has increasingly moved from univariate statistical mapping to representation learning over functional connectomes. A prominent line of work uses self-supervised or contrastive objectives to learn discriminative features with limited labels. Momentum Contrast (MoCo) [11] and its graph/fMRI adaptations have been explored for connectome-based disorder detection, including ASD-oriented contrastive frameworks [13, 25, 27]. Despite strong classification performance in some settings, contrastive objectives often encourage invariances that can discard fine-grained topological information, potentially causing biologically structured perturbations to become indistinguishable from unstructured information loss in the latent space. This limitation is central to our motivation: we evaluate representations via intrinsic **structural sensitivity** rather than solely via accuracy.

Generative Representation Learning and Manifold Modeling. Generative objectives provide an alternative inductive bias by forcing models to explain the input distribution rather than only separating classes. Denoising autoencoders (DAEs) [24] learn representations by reconstructing clean inputs from corrupted observations, which can be interpreted as learning manifold structure and topological constraints of valid samples. This perspective has motivated autoencoder-based approaches in neuroimaging for robust feature learning and, in some cases, implicit handling of site variability. In parallel, masked reconstruction objectives such as Masked Autoencoders (MAE) [10] have shown strong scalability in vision, but their application to abstracted connectome vectors can be data-hungry and may induce oversmoothing when the sample size is limited and the structure is not grid-like. Our framework builds on the manifold-learning view of DAEs [24], leveraging reconstruction as a **forensic prior**: the representation is evaluated by whether it preserves biological topology under controlled perturbations.

Perturbation-based Interpretability and Counterfactual Reasoning. Interpretability in neuroimaging increasingly relies on *intervention*-style analyses, probing how model outputs change under structured input modifications. Counterfactual explanation frameworks in brain imaging have demonstrated the utility of "what-if" transformations for interpreting neural classifiers [16]. In connectome analysis, structured perturbations, including masking, lesioning, or edge manipulation, provide a natural mechanism to test whether a model has learned meaningful dependencies. Our **ROI-level selective masking** follows this philosophy: rather than aggregating into lobes, we audit each ROI by masking its incident connectivity pattern and measuring latent displacement relative to a matched random mask baseline. This design yields fine-grained, anatomically grounded auditing signals that follow the intervention-style logic of perturbation-based neuroscience.

Clinical Scores, Longitudinal Assessment, and the Limits of Behavioural Metrics. Clinical scores and behavioural metrics have been widely used to quantify symptom severity and to relate

Table 1: Demographic and clinical characteristics of subjects.

	Control (n = 77)	**ASD** (n = 95)
Age (mean ± s.d.)	8.03 ± 1.8	5.27 ± 1.7
Male (No. (%))	48 (62.3%)	79 (83.2%)
IQ[a] (mean ± s.d. / n)	110 ± 16.4 / 75	65.34 ± 22.7 / 90
ADOS-2[b] relative score (mean ± s.d. / n)	– / 0	5.95 ± 1.51 / 57

[a] IQ, Intellectual Quotient.
[b] ADOS-2, Autism Diagnostic Observation Schedule, 2nd version.

neuroimaging findings to observable outcomes in ASD. Several studies have further employed longitudinal assessments to examine developmental trajectories and treatment-related changes [5, 15]. However, behavioural metrics are often limited by subjectivity, coarse granularity, and reduced sensitivity to subtle neural changes, particularly at the individual level. These limitations highlight the need for neurobiologically grounded and objective markers, motivating the use of network-based longitudinal fMRI analyses to better characterize underlying brain dysfunction.

3 Methodology

In this paper, we formalize **Generative Manifold Auditing (GMA)**, a scientific auditing framework designed to test whether self supervised representations encode biologically grounded functional brain network structure, rather than becoming insensitive to structured network variation. Unlike conventional evaluation protocols that rely solely on downstream classification accuracy, GMA probes the *intrinsic sensitivity* of the latent space to structured, topologically meaningful perturbations.

3.1 Dataset Overview

For this research, we utilized two datasets, **1) ABIDE I Dataset** and **2) Korean Brain Network (KBN) Dataset**. First, the **ABIDE I** dataset is a large-scale multi-site resting-state fMRI repository. Initially, a total of $N = 884$ subjects were identified; however, after applying a rigorous cleaning protocol, excluding subjects with missing phenotypic information and scans with fewer than 80 time points (TRs), a finalized cleaned cohort of $N = 861$ subjects (396 ASD and 465 TD) spanning multiple international acquisition sites was retained for the study. The second dataset is the **Korean Brain Network (KBN) Dataset**. For longitudinal and intervention-based analysis, we employed the KBN dataset, which contains repeated resting-state fMRI scans of children with Autism Spectrum Disorder (ASD) undergoing neurofeedback or behavioral interventions. This dataset provides an independent clinical cohort with substantial heterogeneity in age, scanner protocol, and treatment context. For both datasets, subject-level functional connectivity was represented using AAL-based ROI time series and Fisher z-transformed connectivity features, enabling the same ROI-level auditing protocol across cohorts.

*3.1.1 Participants (**KBN Dataset**).* Participants were recruited from the Department of Child and Adolescent Psychiatry at Seoul National University Children's Hospital and from the Jung-gu Mental Health Welfare Center in Seoul between December 2012 and December 2024. The inclusion criteria for the ASD group were: (1) age between 3 and 10 years, (2) a prior diagnosis of Autism Spectrum Disorder (ASD) based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), confirmed by the Autism Diagnostic Observation Schedule (ADOS or ADOS-2), and (3) drug-naïve for at least four weeks prior to participation, with a cumulative psychiatric medication history of less than one year. Given that intellectual disability is a common comorbidity in ASD, participants with IQ below 70 were also included.

The inclusion criteria for the typically developing (TD) control group were: (1) age between 3 and 10 years, (2) absence of any diagnosable psychiatric disorder as assessed by structured psychiatric interviews and the Schedule for Affective Disorders and Schizophrenia for School-Age Children—Present and Lifetime Version (K-SADS-PL), and (3) IQ above 80. The exclusion criteria applied to both groups were: (1) presence of congenital genetic disorders, (2) history of significant acquired brain injury (e.g., cerebral palsy), (3) diagnosis of epilepsy, neurological disease, or untreated sensory disorder, (4) history of schizophrenia or psychotic disorders, and (5) diagnosis of obsessive–compulsive disorder, major depressive disorder, or bipolar disorder. Additional exclusion criteria for the control group included: (6) language disorder or severe learning disorder indicating profound developmental delay, and (7) any history of psychiatric medication use.

Following screening, a total of 172 participants were included. The ASD group comprised 95 children (mean age = 5.27 years, SD = 1.73, 83.16% male), while the TD comparison group included 77 children (mean age = 8.03 years, SD = 1.76, 62.33% male). Detailed demographic and clinical characteristics are provided in Table 1.

*3.1.2 Imaging Acquisition and Preprocessing (**KBN Dataset**).* All participants were scanned at Seoul National University Hospital using a 3T Tim Trio MRI scanner (Siemens Healthcare, Erlangen, Germany) equipped with a 32-channel head coil. For younger participants unable to remain still, oral chloral hydrate sedation was administered under psychiatrist supervision with caregiver consent and continuous medical monitoring. Due to a scanner protocol update in April 2020, two acquisition parameter sets were used. For scans acquired prior to April 2020 (n = 127), the parameters were: TR = 3000 ms, TE = 40 ms, flip angle = 90°, voxel size = 1.9×1.9×4.0 mm^3, field of view (FOV) = 240 mm, 35 interleaved slices, matrix size = 64×64, and slice thickness = 4.0 mm. For scans acquired after April 2020 (n = 49), the parameters were: TR = 2000 ms, TE = 30 ms, flip angle = 90°, voxel size = 3.0 × 3.0 × 3.0 mm^3, FOV = 220 mm, 48 slices, matrix size = 64 × 64, and slice thickness = 3.0 mm. Functional and T1-weighted structural MRI data were preprocessed using the Configurable Pipeline for the Analysis of Connectomes (C-PAC, v1.8.7) [8]. Structural images were skull-stripped using FSL-BET [21] and registered to the MNI152 (2 mm) template using Advanced Normalization Tools (ANTs) with symmetric diffeomorphic registration (SyN) [1, 2].

Functional preprocessing included distortion correction, motion correction, and slice-timing correction. Spatial smoothing was applied using a 4-mm full-width at half-maximum (FWHM) Gaussian kernel, followed by band-pass filtering in the range of 0.01–0.1 Hz [7]. Nuisance regression incorporated a 24-parameter motion

model [9], global signal regression, mean cerebrospinal fluid signal, and five aCompCor components derived from white matter and cerebrospinal fluid [3], along with second-degree polynomial detrending [14]. Head motion was quantified using framewise displacement, and volumes exceeding 0.2 mm were censored via scrubbing [18]. Functional-to-anatomical co-registration was performed using ANTs with boundary-based registration. Finally, regional time series were extracted by spatially averaging voxel-wise signals within each region of the Automated Anatomical Labeling (AAL) atlas [22], yielding 116 ROI time series per subject.

All preprocessed images underwent visual quality inspection of tissue segmentation outputs (gray matter, white matter, and cerebrospinal fluid). All datasets passed quality control and were included in subsequent analyses.

3.2 Learning the Normative Brain Manifold

The core of GMA is to characterize the manifold of neurotypical brain organization. We train a **Denoising Autoencoder (DAE)** consisting of an encoder $E(\cdot)$ and decoder $D(\cdot)$ using only TD subjects. The model is optimized to reconstruct clean functional connectivity (FC) patterns from corrupted inputs:

$$\mathcal{L}_{\text{REC}} = \mathbb{E}_{\mathbf{x}\sim P_{TD}}\left[\|\mathbf{x} - D(E(\mathbf{x}+\epsilon))\|^2\right], \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Through denoising, the latent space $\mathcal{Z} = E(\mathbf{x})$ is forced to encode the topological constraints of healthy functional brain networks.

3.3 ROI-level Counterfactual Auditing

To audit whether a learned representation preserves biologically grounded network structure, we perform **ROI-level counterfactual auditing** via anatomically guided interventions. Unlike feature- or edge-wise ablations, which treat functional connections as independent variables, our approach operates at the level of *anatomical units*, reflecting the networked nature of the brain. In GMA, ROI-level masking serves as an anatomically structured perturbation operator: by comparing it with sparsity-matched random masking, we test whether the learned representation responds specifically to network-structured disruption rather than generic information removal.

Let $\mathbf{x}_i \in \mathbb{R}^D$ denote the vectorized upper-triangular elements of the FC matrix for subject i, where $D = N(N-1)/2$ and $N = 116$ is the number of ROIs. For each ROI $r \in \{1, \ldots, N\}$, we define an anatomical masking operator $M_r : \mathbb{R}^D \to \mathbb{R}^D$ that removes all edges incident to ROI r:

$$[M_r(\mathbf{x}_i)]_j = \begin{cases} 0, & j \in \mathcal{E}(r), \\ x_{i,j}, & \text{otherwise}, \end{cases} \tag{2}$$

where $\mathcal{E}(r)$ denotes the set of indices corresponding to functional connections incident to ROI r.

This operation simulates a *virtual lesion* of ROI r by functionally disconnecting the region from the rest of the network, while preserving the topological structure of all remaining connections. The original and intervened inputs are then projected into the latent space via a fixed encoder $E(\cdot)$:

$$\mathbf{z}_i = E(\mathbf{x}_i), \qquad \mathbf{z}_i^{(r)} = E(M_r(\mathbf{x}_i)). \tag{3}$$

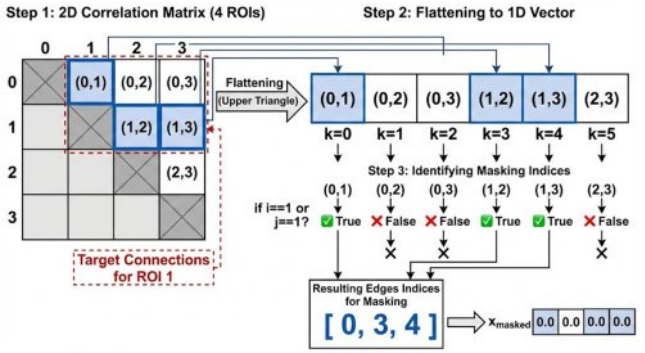


Figure 1: ROI-level counterfactual masking for virtual lesioning. Edges incident to a target ROI are jointly masked in the vectorized functional connectivity representation, simulating node-level disconnection while preserving the remaining network structure.

By comparing these latent representations, we probe how sensitively the learned manifold responds to anatomically meaningful, counterfactual perturbations. This formulation provides an operational definition of network hubs that differs from traditional graph-theoretic centrality. Rather than quantifying hubs based on static connectivity statistics, we identify hubs as regions whose functional disconnection induces disproportionate disruption of the learned normative manifold.

3.4 Structural Integrity and Sensitivity Gaps

For each subject i and ROI r, we quantify the impact of the virtual lesion using the **Structural Integrity Gap**:

$$\Delta_{r,i} = \left\|\mathbf{z}_i - \mathbf{z}_i^{(r)}\right\|_2. \tag{4}$$

A larger $\Delta_{r,i}$ indicates that the representation relies more strongly on the integrity of ROI r to maintain a coherent embedding, suggesting that the region plays a critical role in the learned network structure. However, raw latent displacement alone does not distinguish biologically structured sensitivity from generic vulnerability to information loss. To control for this effect, each ROI-level intervention is paired with a null perturbation that removes the same number of edges at random. Let M_{rand} denote such a randomized masking operator. We define the **Sensitivity Gap** as:

$$\Delta_{r,i}^{\text{sens}} = \|\mathbf{z}_i - E(M_r(\mathbf{x}_i))\|_2 - \|\mathbf{z}_i - E(M_{\text{rand}}(\mathbf{x}_i))\|_2. \tag{5}$$

Larger sensitivity gaps indicate that anatomically guided perturbations induce stronger latent deviations than random perturbations of equal sparsity, reflecting preserved structural awareness. Values near zero indicate limited differential sensitivity, where structured and unstructured perturbations become indistinguishable in the latent space.

Post-hoc Anatomical Aggregation. All interventions are performed strictly at the ROI level, without imposing anatomical priors during training or auditing. For statistical analysis and visualization only, $\Delta_{r,i}$ values are aggregated post hoc into anatomical groups (e.g., Frontal, Parietal, Occipital, Limbic).

3.5 Candidate Sensitivity Marker Discovery via Anatomically Guided Virtual Lesioning

Building on the proposed counterfactual auditing framework, we leverage anatomically guided virtual lesioning as a mechanism for **network-level candidate sensitivity marker discovery**. Rather than attributing importance to individual features or edges, we identify brain regions whose functional integrity is critical for maintaining the normative organization of the learned manifold.

Figure 1 illustrates the virtual lesioning procedure. Starting from a subject-level functional connectivity matrix, we extract the upper-triangular edge representation and identify all connections incident to a target ROI. These edges are simultaneously masked in the vectorized space, effectively simulating a node-level disconnection while preserving the remaining network structure. This anatomically grounded intervention enables controlled, region-specific perturbations without altering global signal statistics.

Let $\mathcal{D}(\cdot)$ denote the decoding or reconstruction function of the generative model trained on typically developing (TD) subjects. For subject i and ROI r, we define the lesion-induced reconstruction deviation as:

$$S_r(x_i) = \|\mathcal{D}(\mathbf{x}_i) - \mathcal{D}(M_r(\mathbf{x}_i))\|_2 . \quad (6)$$

To distinguish ROI-specific structural relevance from nonspecific information loss, we further introduce a matched-random controlled score:

$$S_r^{\text{ctrl}}(x_i) = S_r(x_i) - \mathbb{E}_{M_{\text{rand}} \sim \mathcal{R}_r} \left[\|\mathcal{D}(\mathbf{x}_i) - \mathcal{D}(M_{\text{rand}}(\mathbf{x}_i))\|_2 \right], \quad (7)$$

where $\mathcal{R}_r$ denotes the distribution of random masks that remove the same number of edges as the ROI-level perturbation M_r. This score extends the matched-random control principle from latent sensitivity auditing to ROI-level candidate marker ranking.

Because the model is optimized to reconstruct samples lying on the normative TD manifold, a large deviation $S_r(x_i)$ indicates that disconnecting ROI r substantially disrupts the model's ability to recover a valid, healthy-like functional configuration. Regions that consistently induce large deviations across subjects are thus interpreted as *network-critical components* of normative brain organization. Within this framework, candidate marker identification is decoupled from downstream classification accuracy and explicit label supervision. Instead, candidate markers are explored through *counterfactual interrogation* of the latent representation, guided by anatomical plausibility and network topology. This perspective supports mechanism-aware exploration of regions potentially associated with disease, while reducing reliance on potentially unstable predictive features. Regions showing large and stable $S_r(x_i)$ or $S_r^{\text{ctrl}}(x_i)$ are indicative of functional hub-like behaviour in the normative brain network, with hubness assessed via disruption of global organization rather than raw connectivity statistics.

3.6 Sensitivity Analysis on the KBN Dataset

To examine the robustness and potential clinical relevance of GMA beyond the primary ABIDE analysis, we additionally applied the proposed sensitivity framework to the Korean Brain Network (KBN) dataset. The KBN dataset provides an independent cohort with repeated fMRI scans, allowing us to assess whether ROI-level sensitivity patterns observed in cross-sectional analysis remain meaningful under real-world clinical settings.

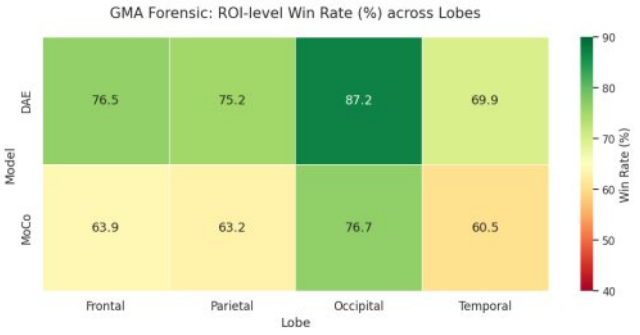


Figure 2: ROI-level structural win rate comparison between DAE and MoCo. Win rate measures how often anatomically guided ROI perturbations induce larger latent displacement than matched random perturbations.

Longitudinal symptom severity was assessed using the **Korean Childhood Autism Rating Scale (K-CARS)** [20]. K-CARS is a 15-item clinician-rated scale evaluating core autistic symptoms, including social deficits, restricted and repetitive behaviors, and sensory hypersensitivity. Scores range from 15 to 60, with a diagnostic cutoff of 30; scores of 30–36.5 indicate mild-to-moderate autism, and scores above 37 indicate severe autism. By jointly considering K-CARS trajectories and longitudinal changes in $\Delta_{r,i}$, we examine whether behavioral improvement and neuro-structural sensitivity exhibit similar or divergent trends over time.

3.7 Evaluation Protocol and Experimental Setup

Baselines and Inductive Biases. We compare models with fundamentally different scientific inductive biases: *1) DAE (Generative; Ours):* MLP-based denoising reconstruction to learn the normative manifold. *2) MoCo (Contrastive):* Instance discrimination maximizing invariance across augmented views. *3) MAE (Masked Autoencoder):* Transformer-based masked reconstruction for tabular FC representations.

Cross-Validation and Site Robustness. To prevent site leakage in multisite fMRI analysis, we employed *5-fold stratified group cross-validation*, ensuring that subjects from the same acquisition site never appear in both training and test folds. Unless otherwise stated, we adopt a *no-harmonization policy* beyond Fisher z-transformation, testing whether models learn *implicit biological invariants* under realistic scanner-induced heterogeneity.

4 Results

4.1 Generative Manifold Auditing: DAE vs. MoCo

We first evaluate whether the proposed Generative Manifold Auditing (GMA) framework can distinguish representations that preserve biologically meaningful network structure from those that do not. We compare a DAE trained on TD subjects against a contrastive baseline (MoCo), using identical ROI-level counterfactual interventions.

Sensitivity Gap Distribution. Across all ROIs, the sensitivity-gap distribution of DAE is systematically shifted toward larger values compared to MoCo. This indicates that anatomically guided ROI

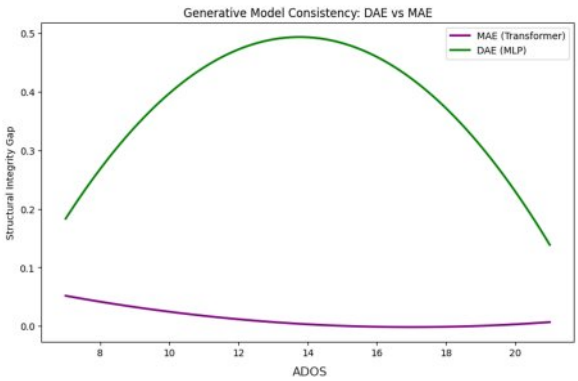


Figure 3: Comparison of structural sensitivity between DAE and MAE. DAE preserves non-linear sensitivity to anatomically meaningful variation, whereas MAE exhibits flattened sensitivity despite apparent stability.

perturbations induce stronger latent deviations than random perturbations of equal sparsity, demonstrating that DAE preserves structural awareness of functional brain topology. In contrast, MoCo exhibits sensitivity gaps concentrated near zero, suggesting that structured and unstructured perturbations are largely indistinguishable in its latent space.

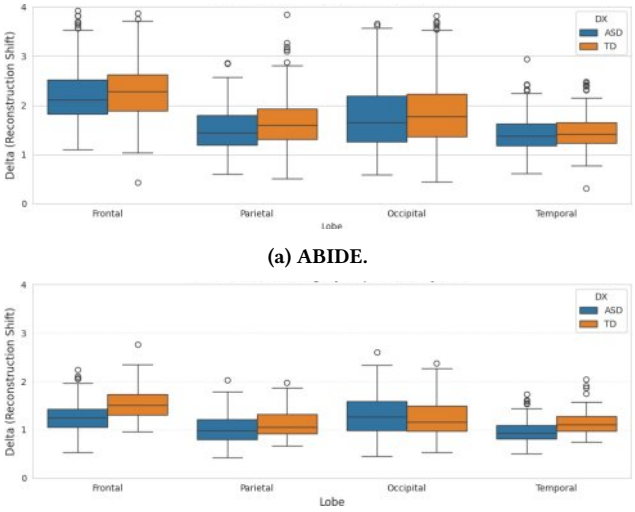
ROI-level Win Rate. To quantify robustness at the regional level, we compute a win rate defined as the proportion of subjects for which anatomically guided perturbations produce larger latent deviations than randomized perturbations (Fig. 2). DAE consistently achieves higher win rates than MoCo across all major anatomical lobes. In particular, the advantage is most pronounced in the Occipital and Parietal lobes, indicating that DAE better preserves fine-grained biological structure in sensory and integrative systems. These results suggest that generative denoising objectives provide a stronger inductive bias for structural sensitivity than contrastive learning in multisite fMRI data.

Quantitatively, DAE achieved a higher mean sensitivity gap than MoCo (0.040 vs. 0.022) and a higher mean win rate (0.67 vs. 0.58), supporting its stronger structured sensitivity under matched perturbation controls.

4.2 When Stability Masks Structural Sensitivity: DAE vs. MAE

We next examine whether reconstruction-based objectives alone are sufficient to preserve biologically meaningful structural sensitivity. To this end, we compare the proposed DAE with a masked autoencoder (MAE), which reconstructs partially observed inputs but lacks an explicit denoising constraint (Fig. 3).

Although MAE achieves high apparent stability under split-half and perturbation tests, this stability reflects a fundamentally different inductive bias. Specifically, MAE exhibits uniformly small sensitivity gaps across ROIs and subjects, indicating that both anatomically guided and randomized perturbations induce similarly small latent deviations. This behavior suggests that MAE learns representations that are invariant to large portions of the input space, resulting in *flattened sensitivity profiles*. While such invariance may be advantageous for compression or representation compactness,



(a) ABIDE.

(b) KBN external validation.

Figure 4: Structural Integrity Gap (Δ) distributions across major lobes on ABIDE and KBN.

it obscures biologically meaningful distinctions between structured network disruptions and unstructured information loss.

In contrast, DAE preserves non-uniform, ROI-specific sensitivity patterns, allowing structured perturbations to propagate differentially through the latent space. This distinction highlights a critical methodological insight: **stability alone is not indicative of biological validity**. Instead, meaningful representations must exhibit structured sensitivity aligned with anatomical organization.

These findings demonstrate that the denoising objective plays a crucial role in preserving structured sensitivity, whereas masked reconstruction without explicit noise modeling can yield representations that appear stable yet are less informative for structured sensitivity auditing. Both models are trained on the same dataset, without additional pretraining or external supervision. This setup isolates the effect of reconstruction design, suggesting that the observed differences in sensitivity are unlikely to arise from data scale, but instead reflect how structural dependencies are encoded during learning.

4.3 Sensitivity Deviations in the ASD Manifold

We next examine whether GMA reveals systematic differences in the distribution of structural sensitivity between ASD and TD populations.

ABIDE Benchmark. On the ABIDE dataset, Structural Integrity Gaps exhibit clear lobe-dependent differences between the ASD and TD groups (Fig. 4-a). Across all examined lobes, the distributions show substantial overlap, with no evidence of uniform or global separation. Instead, group differences manifest as subtle shifts in distributional profiles, with their magnitude and expression varying across anatomical systems.

External Validation on KBN. A similar pattern is observed in the independent KBN cohort (Fig. 4-b). While sensitivity distributions again differ between groups, the direction and prominence of these differences vary across lobes, and overlapping distributions

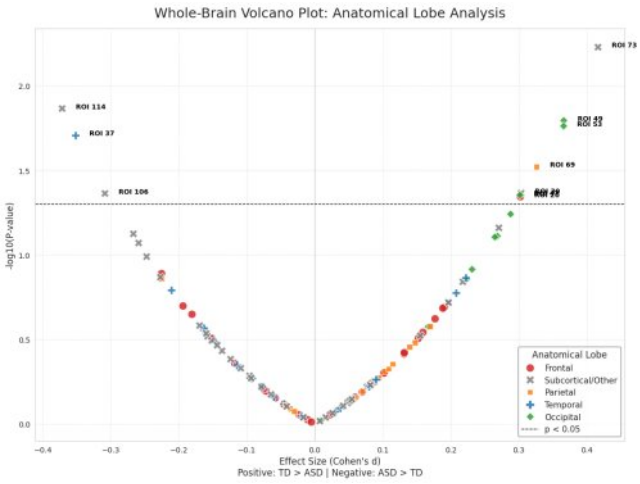


Figure 5: Whole-brain volcano plot of ROI-level structural sensitivity differences between ASD and TD groups on ABIDE. Each point represents one ROI ($N = 116$), with Cohen's d on the x-axis and statistical significance on the y-axis.

remain the dominant characteristic. Accordingly, cross-dataset consistency is reflected in distributional characteristics, rather than in stable effect directions or magnitudes.

Overall, these findings indicate that ASD-related sensitivity differences are best characterized as heterogeneous, lobe-dependent deviations within the learned manifold, rather than as a uniform reduction or deficit. This distributional perspective motivates subsequent analyses that move beyond global summary statistics to region- and subject-specific interrogation.

4.4 Structural Sensitivity Patterns Across ROIs

We further analyze ROI-level sensitivity deviations using whole-brain effect size profiling. Across 116 ROIs, effect sizes form a continuous distribution, with a subset of regions exhibiting large absolute deviations between ASD and TD groups. Rather than attributing distinct pathological mechanisms to the sign of the effect size, we focus on its magnitude as an indicator of deviation from the normative sensitivity profile. ROIs with large absolute effect sizes are those for which structural perturbations lead to pronounced changes in latent representations, indicating heightened sensitivity to network disruption. The identified ROIs are not uniformly distributed across the brain, but show concentration within motor, sensory, and limbic systems. This anatomical structure is revealed through counterfactual auditing, without introducing anatomical priors during training or evaluation.

4.5 Heterogeneous Sensitivity Profiles Across Clinical Severity

We examined whether Structural Integrity Gaps exhibit systematic associations with clinical severity. Across both datasets, linear correlation analyses between integrity gaps and symptom severity scores yield weak and inconsistent effects, with substantial variability across regions. Rather than following a monotonic trend,

Table 2: Representative ROI-level sensitivity markers on the aggregated KBN dataset (effect size: Cohen's d).

ROI	Effect	ROI	Effect
Putamen (L) - 73	**0.88**	Vermis - 114	-0.04
Insula (R) - 30	**0.64**	SMA (R) - 20	-0.11
Occipital Sup (L) - 49	**0.34**	Occipital Inf (L) - 53	-0.15
Hippocampus (L) - 37	**0.31**	Parietal Inf (L) - 69	**-0.30**
Lingual (L) - 47	0.10	Vermis 10 - 106	**-0.30**

sensitivity values form heterogeneous distributions across anatomical systems. Whole-brain effect-size profiling reveals that both positive and negative deviations are distributed across frontal, temporal, parietal, occipital, and subcortical regions, indicating that structural sensitivity differences cannot be explained by severity alone. These findings highlight a key limitation of severity-centric analyses: structural sensitivity does not scale linearly with symptom scores. Instead, GMA captures region- and subject-specific deviations in functional network organization that are largely orthogonal to global clinical severity.

4.6 From Whole-Brain Profiling to Candidate Sensitivity Markers

To translate the whole-brain effect-size landscape (Fig. 5) into interpretable candidates, we summarize a compact set of *candidate sensitivity markers* that exhibit robust group separation on the aggregated KBN cohort. Specifically, we rank ROIs by the magnitude of Cohen's d computed between TD and ASD groups, and report representative regions spanning both directions (positive: TD > ASD; negative: ASD > TD), rather than restricting attention to a single sign. These regions are used as measurement-level indicators of systematic sensitivity variation, not as claims of disease-localized pathology. They provide an anatomically grounded summary of the whole-brain effect-size profile and establish a stable ROI set for subject-level longitudinal tracking.

Matched-random controlled marker analysis. Using the matched random controlled score, we repeated the ROI-ranking analysis to control for nonspecific information loss. The ranking remained stable under this stricter criterion, retaining 8 of the original top-10 ROIs with substantial rank agreement (Spearman $\rho = 0.782$). This supports that the candidate markers reflect ROI-specific structural relevance rather than only the amount of removed connectivity. A visual comparison is shown in Appendix Fig. 8.

4.7 Longitudinal Sensitivity Dynamics Under Intervention

We examined longitudinal sensitivity trajectories using the three KBN subjects with matched pre- and post-intervention imaging. While population-level analyses on ABIDE identified candidate regions based on effect-size ranking, longitudinal analysis requires a different perspective in which the relevant regions are those whose sensitivity profiles exhibit measurable changes within subjects over time. Population-level effect sizes prioritize regions with stable group separation, whereas longitudinal interpretation prioritizes regions that show reliable within-subject modulation under

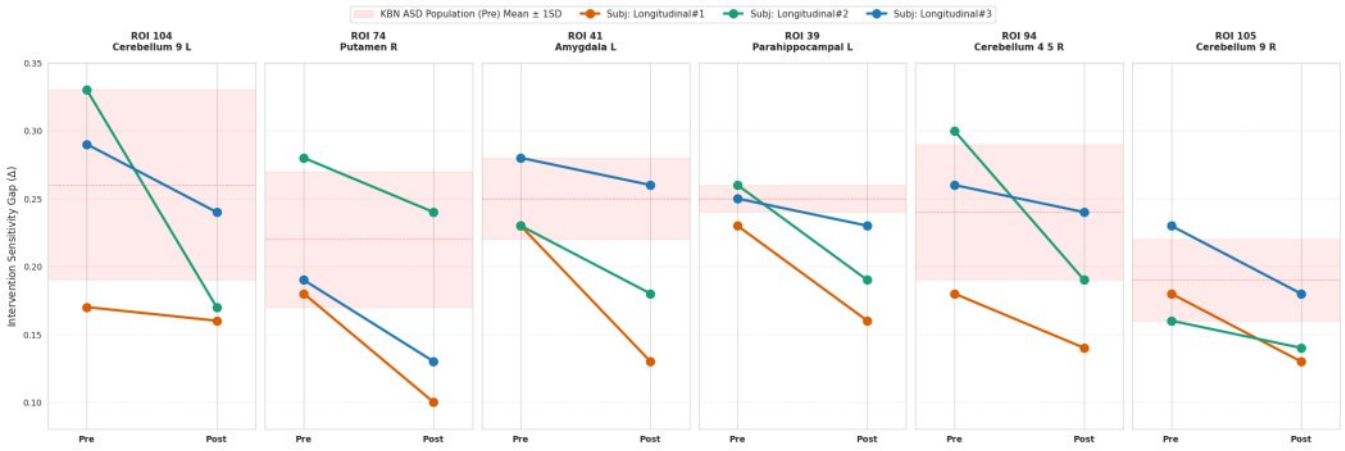


Figure 6: Longitudinal ROI-level Structural Integrity Gaps (Δ) in KBN before and after intervention. Solid lines indicate individual ASD subjects; shading denotes the baseline ASD group mean ± one standard deviation. ROIs show heterogeneous post-intervention sensitivity changes.

Table 3: Longitudinal K-CARS clinical assessments and imaging timepoints for KBN subjects.

Subject	Scale	Clinical Change	MRI Match	V0 Scan Date	V1 Scan Date	V0 K-CARS	V1 K-CARS
Longitudinal#1	K-CARS	Slightly improved	Yes	2018-12-26	2020-01-30	35.5	29.5
Longitudinal#2	K-CARS	Slightly improved	Yes	2018-09-27	2020-01-05	41.0	32.5
Longitudinal#3	K-CARS	Slightly improved	Yes	2015-10-27	2016-11-21	33.5	28.5

intervention; these represent complementary, but not equivalent, criteria.

All three longitudinal KBN subjects showed behavioral improvement as measured by the **Korean Childhood Autism Rating Scale (K-CARS)**. At the neural level, however, Structural Integrity Gaps exhibited heterogeneous, region-specific trajectories (Fig. 6). Several regions, including the right putamen (ROI 74), left amygdala (ROI 41), and parahippocampal cortex (ROI 39), showed post-intervention reductions in the available longitudinal cases, with trajectories moving toward the ASD population mean range. These patterns are consistent with partial stabilization of network-level responses following intervention.

In contrast, cerebellar regions demonstrated marked inter-subject variability. For example, sensitivity in Cerebellum 9 L (ROI 104) and Cerebellum 9 R (ROI 105) decreased substantially in some subjects, while remaining elevated or only partially reduced in others. Similarly, Cerebellum 4/5 R (ROI 94) exhibited persistent deviations in certain individuals despite behavioral improvement. This dissociation indicates that behavioral improvement can occur alongside region-specific persistence of altered neural sensitivity. Together, these findings suggest that GMA enables subject-specific interrogation of latent neural dynamics that are not captured by behavioral assessments alone.

5 Conclusion

We presented **Generative Manifold Auditing (GMA)**, a framework for evaluating self-supervised representations of resting-state fMRI beyond predictive accuracy through anatomically guided interventions. By applying ROI-level perturbations and matched random controls, GMA probes the structural sensitivity of latent spaces and assesses whether learned representations preserve biologically meaningful network organization. Using the ABIDE and KBN cohorts, we showed that denoising-based generative models retain structured sensitivity patterns more clearly than contrastive and masked reconstruction objectives, which may appear stable while exhibiting flattened structural sensitivity. The matched-random controlled marker analysis further showed that candidate ROI-level sensitivity markers remain largely stable after controlling for non-specific information loss, while longitudinal analysis illustrated how GMA can track subject-specific neural sensitivity dynamics beyond behavioral scores alone. These findings highlight the importance of auditing not only what representations predict, but also how they respond to structured, biologically meaningful perturbations. Overall, GMA provides a principled approach for auditing representation validity in networked biological data, with broader implications for intervention-driven analysis in scientific machine learning.

6 Limitations and Ethical Considerations

Limitations. The longitudinal analysis is limited by the small number of ASD subjects with complete pre- and post-intervention fMRI scans in the KBN dataset. This reflects the scarcity of longitudinal resting-state fMRI data in clinical ASD cohorts. Accordingly, the longitudinal results are intended as a demonstration of neural sensitivity dynamics rather than population-level statistical inference, while the main findings are supported by large-scale cross-sectional analyses and independent external validation.

Ethical Considerations. A written informed consent was obtained both from every participant and from their parent and/or legal

guardian after study design was explained. All personal identifiers were removed before analysis, and data were handled in accordance with institutional and national data protection regulations. This study was approved by the Institutional Review Board of Seoul National University Hospital (IRB approval number: 2008-116-1150, 1507-118-690, 1206-054-41) and was performed in accordance with the Declaration of Helsinki.

Acknowledgments

This study was supported by research funds from the National Center for Mental Health, Ministry of Health & Welfare, Republic of Korea (DMHR25E01 to B-NK and KH), the National Research Foundation (NRF) funded by the Korean Government (MSIT) (No. 2020M3E5D9080787 and RS-2024-00397737 to B-NK), and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02217259 to KH and CH).

References

[1] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. 2008. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis* 12, 1 (2008), 26–41.

[2] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage* 54, 3 (2011), 2033–2044.

[3] Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T Liu. 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* 37, 1 (2007), 90–101.

[4] Xia-an Bi, Junxia Zhao, Qian Xu, Qi Sun, and Zhigang Wang. 2018. Abnormal functional connectivity of resting state network detection based on linear ICA analysis in autism spectrum disorder. *Frontiers in physiology* 9 (2018), 475.

[5] G Bussu, Emily JH Jones, T Charman, Mark H Johnson, and JK Buitelaar. 2018. Prediction of autism at 3 years from behavioural and developmental measures in high-risk infants: a longitudinal cross-domain classifier analysis. *Journal of Autism and Developmental Disorders* 48, 7 (2018), 2418–2433.

[6] Heng Chen, Jason S Nomi, Lucina Q Uddin, Xujun Duan, and Huafu Chen. 2017. Intrinsic functional connectivity variance and state-specific under-connectivity in autism. *Human brain mapping* 38, 11 (2017), 5740–5755.

[7] Dietmar Cordes, Victor M Haughton, Konstantinos Arfanakis, Joseph D Carew, Patryk A Turski, Charles H Moritz, Mark A Quigley, and Mary E Meyerand. 2001. Frequencies contributing to functional connectivity in the cerebral cortex in resting-state data. *American Journal of Neuroradiology* 22, 7 (2001), 1326–1333.

[8] Cameron Craddock, Sharad Sikka, Brian Cheung, Ranjeet Khanuja, Satrajit S Ghosh, Chaogan Yan, Qingyang Li, Daniel Lurie, Joshua Vogelstein, Randal Burns, et al. 2013. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). *Front Neuroinform* 42, 10.3389 (2013).

[9] Karl J Friston, Sarah Williams, Richard Howard, Richard S J Frackowiak, and Robert Turner. 1996. Movement-related effects in fMRI time-series. *Magnetic Resonance in Medicine* 35, 3 (1996), 346–355.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 16000–16009.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 9729–9738.

[12] Christopher L Keown, Michael C Datko, Colleen P Chen, Jose Omar Maximo, Afrooz Jahedi, and Ralph-Axel Müller. 2017. Network organization is globally atypical in autism: a graph theory study of intrinsic functional connectivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 2, 1 (2017), 66–75.

[13] Yicheng Leng, Syed Muhammad Anwar, Islem Rekik, Sen He, and Eung-Joo Lee. 2025. Self-supervised graph transformer with contrastive learning for brain connectivity analysis towards improving autism detection. In *2025 IEEE 22nd international symposium on biomedical imaging (ISBI)*. IEEE, 1–5.

[14] Torben E Lund, Kristoffer H Madsen, Karam Sidaros, Wen-Lin Luo, and Thomas E Nichols. 2006. Non-white noise in fMRI: Does modelling have an impact? *NeuroImage* 29, 1 (2006), 54–66.

[15] Shreyas Mahapatra, Edward Khokhlovich, Samantha Martinez, Benjamin Kannel, Stephen M Edelson, and Andrey Vyshedskiy. 2020. Longitudinal epidemiological study of autism subgroups using Autism Treatment Evaluation Checklist (ATEC) score. *Journal of autism and developmental disorders* 50, 5 (2020), 1497–1508.

[16] T. Matsui et al. 2022. Counterfactual Explanation of Brain Activity Classifiers Using Image-To-Image Transfer by Generative Adversarial Network. *Frontiers in Neuroinformatics* (2022).

[17] Sarah K Noonan, Frank Haist, and Ralph-Axel Müller. 2009. Aberrant functional connectivity in autism: evidence from low-frequency BOLD signal fluctuations. *Brain research* 1262 (2009), 48–63.

[18] Jonathan D Power, Anish Mitra, Timothy O Laumann, Abraham Z Snyder, Bradley L Schlaggar, and Steven E Petersen. 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* 84 (2014), 320–341.

[19] Lu Qian, Yao Wang, KangKang Chu, Yun Li, ChaoYong Xiao, Ting Xiao, Xiang Xiao, Ting Qiu, YunHua Xiao, Hui Fang, et al. 2018. Alterations in hub organization in the white matter structural network in toddlers with autism spectrum disorder: A 2-year follow-up study. *Autism Research* 11, 9 (2018), 1218–1228.

[20] Min Sup Shin and Yung-Hee Kim. 1998. Standardization study for the Korean version of the Childhood Autism Rating Scale: Reliability, validity and cut-off score. *Korean Journal of Clinical Psychology* (1998).

[21] Stephen M Smith. 2002. Fast robust automated brain extraction. *Human Brain Mapping* 17, 3 (2002), 143–155.

[22] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicole Delcroix, Bernard Mazoyer, and Marc Joliot. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* 15, 1 (2002), 273–289.

[23] Marjolein Verly, Judith Verhoeven, Inge Zink, Dante Mantini, Lukas Van Oudenhove, Lieven Lagae, Stefan Sunaert, and Nathalie Rommel. 2014. Structural and functional underconnectivity as a negative predictor for language in autism. *Human brain mapping* 35, 8 (2014), 3602–3615.

[24] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning (ICML)*. 1096–1103.

[25] Xiaochuan Wang, Yuqi Fang, Qianqian Wang, Pew-Thian Yap, Hongtu Zhu, and Mingxia Liu. 2025. Self-supervised graph contrastive learning with diffusion augmentation for functional MRI analysis and brain disorder detection. *Medical image analysis* 101 (2025), 103403.

[26] David E Welchew, Chris Ashwin, Karim Berkouk, Raymond Salvador, John Suckling, Simon Baron-Cohen, and Ed Bullmore. 2005. Functional disconnectivity of the medial temporal lobe in Asperger's syndrome. *Biological psychiatry* 57, 9 (2005), 991–998.

[27] Zuohao Yin, Feng Xu, Yue Ma, Shuo Huang, Kai Ren, and Li Zhang. 2025. MAMVCL: Multi-Atlas Guided Multi-View Contrast Learning for Autism Spectrum Disorder Classification. *Brain Sciences* 15, 10 (2025), 1086.

A Additional Analyses

This appendix reports additional diagnostic checks supporting the robustness of the proposed sensitivity auditing results. We examine two complementary issues: whether the latent sensitivity summary is primarily explained by cognitive ability, and whether ROI-level candidate marker rankings remain stable after a matched-random control that accounts for nonspecific information removal.

Cognitive ability confound check. Figure 7 provides an additional check for whether cognitive ability acts as a dominant confounding factor in the latent sensitivity analysis. Because the ABIDE cohort contains substantial site heterogeneity, a single pooled association between full-scale IQ (FIQ) and the latent sensitivity summary could obscure site-specific variation. We therefore summarize the relationship separately within each acquisition site. The site-wise correlations are weak and heterogeneous, with no stable positive or negative pattern across sites. This suggests that cognitive ability may contribute to local variation in some sites, but does not provide a consistent explanation for the main sensitivity patterns. This site-wise analysis is intended as a conservative diagnostic check rather than a definitive causal adjustment. The absence of a consistent cross-site association supports the use of latent sensitivity as a complementary signal beyond cognitive ability.

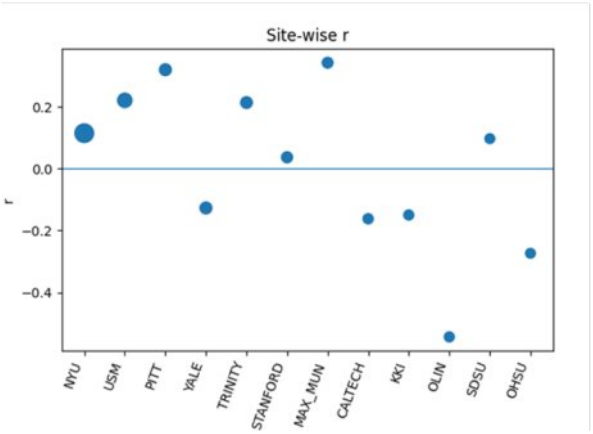


Figure 7: Site-wise correlation between FIQ and latent sensitivity. Correlations are weak and heterogeneous across acquisition sites.

Matched-random controlled marker consistency. Figure 8 examines whether the candidate ROI-level marker rankings reflect anatomically structured sensitivity rather than generic vulnerability to information removal. In the raw analysis, ROI effects are estimated from anatomically guided perturbations that remove all connections incident to a target ROI. However, a large response could in principle arise from the number of removed connections rather than from ROI-specific structural relevance. To control for this possibility, we compare raw Cohen's d values with matched-random controlled Cohen's d values, where each ROI perturbation is contrasted against random masks that remove the same number of edges. The close alignment between raw and controlled effects indicates that the ranking structure is largely preserved under this stricter control.

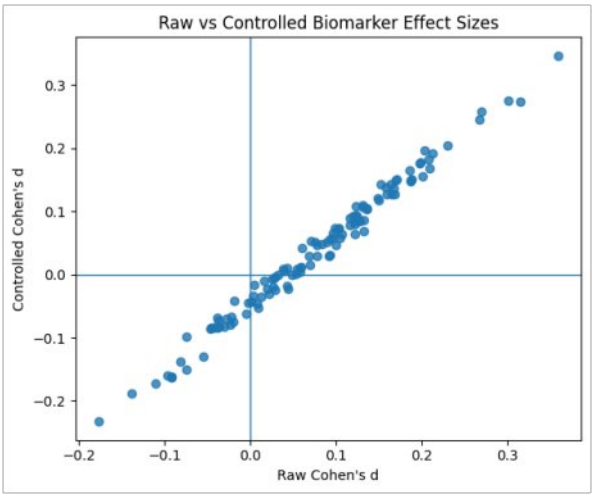


Figure 8: Consistency between raw and matched-random controlled ROI effect sizes. The near-linear alignment indicates that ROI-level rankings are largely preserved.

Taken together, these appendix analyses support the robustness of the proposed auditing results. The FIQ analysis suggests that latent sensitivity is not consistently explained by cognitive ability across sites, while the matched-random comparison shows that candidate marker rankings remain stable under perturbation-size control.