Explainable AI techniques for dynamic functional brain imaging: validation and analysis of E/I imbalance in autism

Anthony Strock[1,*], *Trang-Anh E. Nghiem*[1], *Alex De Lecea*[5], *Atif Hassan*[6], *Srikanth Ryali*[1], *Vinod Menon*[1,2,3,4,*]

[1] Department of Psychiatry & Behavioral Sciences, Stanford University, CA 94305, USA.

[2] Department of Neurology & Neurological Sciences, Stanford University, CA 94305, USA.

[3] Wu Tsai Neurosciences Institute, Stanford University, CA 94305, USA.

[4] Graduate School of Education, Stanford University, CA 94305, USA.

[5] Department of Mathematics, University of Washington, Seattle, WA 98195, USA

[6] Department of Artificial Intelligence, IIT Kharagpur, India

[*] Address correspondence to:

astrock@stanford.edu; menon@stanford.edu

Abstract

Deep neural networks are increasingly crucial for analysing dynamic functional brain imaging data, offering unprecedented accuracy in distinguishing brain activity patterns across health and disease. However, they often function as black boxes obscuring the neurobiological features driving classifications between groups. This study systematically investigates explainable AI (xAI) methods to address this challenge, employing two complementary simulation approaches: recurrent neural networks for controlled parameter exploration, and The Virtual Brain for biophysically realistic whole-brain modelling. These simulations generate fMRI datasets with known regional alterations in excitation/inhibition (E/I) balance, mimicking mechanisms implicated in psychiatric and neurological disorders. Our comprehensive validation demonstrates that Integrated Gradients and DeepLift successfully identify ground-truth affected regions across challenging conditions, including high noise (-10dB SNR), low prevalence (1% of regions), and subtle E/I alterations. This performance remains robust across three different attribution methods and baseline choices, establishing the reliability of xAI for functional neuroimaging analysis. Critically, successful cross-species validation using both human (68-region) and mouse (426-region) connectomes demonstrates the approach's ability to detect mechanistic alterations across different scales of brain organization. Application to the multisite ABIDE resting-state fMRI dataset (N=834) reveals that regions within the default mode network, particularly the posterior cingulate cortex and precuneus, most clearly differentiated children with autism from neurotypical controls. The convergence between these empirical findings and our biophysical simulations of E/I imbalance provides computational support for mechanistic theories of E/I imbalance in autism while demonstrating how xAI can bridge cellular-level mechanisms with clinical biomarkers. This work establishes a framework for reliable interpretation of deep neural

network models in functional neuroimaging, with implications for understanding brain disorders and developing targeted brain stimulation strategies.

Author Summary

Our research tackles a key challenge in brain imaging analysis: making complex artificial intelligence methods more transparent and biologically meaningful. When scientists and clinicians use AI to analyse functional brain scans, these powerful tools can accurately discriminate conditions like autism, but they typically don't explain which brain regions are involved or why. This creates a "black box" problem that limits our understanding of brain disorders. We developed and tested methods that reveal which brain areas contribute most to AI decisions when analysing brain activity. Using both computer simulations and real patient data, we demonstrate that these methods can reliably identify regions where the balance between excitation and inhibition is disrupted - a mechanism thought to underlie autism and other conditions. Our findings specifically highlight the posterior cingulate cortex and precuneus, key regions of the "default mode network" involved in self-directed and social cognition. This work contributes to bridging the gap between cellular-level mechanisms and whole-brain patterns observed in clinical scans, potentially helping scientists develop more targeted treatments and better understand the biological basis of brain disorders.

Introduction

Deep neural networks (DNNs) are increasingly central to human brain research, offering powerful tools for scientific discovery and clinical applications (1, 2, 3, 4, 5). These models have demonstrated remarkable success in identifying patterns and making predictions from complex neuroimaging data (3, 4, 5). However, a fundamental challenge persists: DNN models typically function as "black boxes", where understanding which neurobiological features drive model decisions remains elusive. Despite their high prediction accuracy, these models offer little insight into how specific patterns of brain activity contribute to classification or what neural mechanisms might underlie these patterns This lack of interpretability poses a significant obstacle in neuroscience, where identifying meaningful biomarkers and predictors of clinical symptoms is essential for advancing our understanding of brain function and dysfunction. The inability to interpret these models limits their utility for developing mechanistic insights and targeted therapeutic strategies, particularly in complex disorders where understanding the biological basis of model decisions is crucial. Here we develop and validate a framework for making DNN models interpretable in the context of functional neuroimaging, demonstrating how the recently developed explainable AI (xAI) methods can bridge cellular-level mechanisms with whole-brain dynamics observed in clinical populations.

While deep learning has revolutionized structural MRI analysis (6, 7, 8) through advances in segmentation, classification, and disease prediction, its application to functional MRI has progressed more slowly due to the inherent complexity of its temporal dynamics. Traditional approaches to functional MRI analysis have relied heavily on pre-engineered features, focusing primarily on static inter-regional connectivity measures or predefined activation

patterns (9, 10, 11). Recent attempts to incorporate neural networks, including graph-based approaches (12, 13) and surface-based convolutional networks (10), still largely reduce temporal dynamics to static representations. While these approaches have provided valuable insights into brain organization, they often fail to capture the full spatiotemporal complexity of brain dynamics, including state transitions (14) and time-varying connectivity patterns (15). This reduction of temporal information represents a crucial limitation, as growing evidence suggests that dynamic properties of brain activity carry important information about both normal brain function and disease states (16).

To overcome these limitations, we recently developed a novel spatiotemporal DNN (stDNN) that analyses fMRI time series data without feature engineering (3, 4, 17). Unlike conventional approaches, stDNNs learn discriminative features directly from raw temporal signals, preserving the rich dynamics of brain activity patterns. This approach has demonstrated remarkable success across multiple domains in clinical neuroscience. In studies of sex differences, stDNNs identified replicable, behaviorally relevant patterns of functional brain organization that generalize across independent cohorts (17). When applied to autism spectrum disorder, these models discovered robust brain signatures that predict both diagnosis and social communication symptom severity with high accuracy (4), establishing their utility for patient stratification. Furthermore, in a multi-cohort study of the 22q11.2 deletion syndrome, stDNNs revealed reproducible functional brain signatures associated with psychosis risk (3), demonstrating their ability to detect clinically relevant brain-behavior relationships in brain disorders.

However, despite their predictive power, these models face a crucial limitation: like other deep learning approaches, they function as “black boxes”, leaving unclear which patterns of

brain activity and connectivity drive their predictions. This interpretability gap is particularly problematic for clinical applications, where understanding the biological basis of model decisions is essential for developing targeted interventions (15, 18). For instance, while stDNNs can accurately identify individuals with autism, two critical challenges remain: first, identifying which brain circuits drive these classifications, and second, validating that the identified circuits accurately reflect underlying neurobiological mechanisms rather than algorithmic artifacts. This validation is crucial because feature attribution methods themselves must be rigorously tested before we can confidently use them to guide mechanistic understanding or treatment strategies. The challenge reflects a fundamental tension in clinical machine learning between achieving high predictive accuracy and gaining mechanistic insights that can inform therapeutic development (18, 19).

Recent advances in the field of Explainable Artificial Intelligence (xAI) offer promising solutions to this interpretability problem. Among various approaches, Integrated Gradients (IG) (20) have emerged as particularly promising, providing a mathematically rigorous framework for attributing model predictions to input features. IG accumulates gradients along a path from baseline to input to capture cumulative feature importance (20). The method satisfies important theoretical properties such as sensitivity and implementation invariance. Although these methods have been successfully applied in computer vision applications, applying these methods to fMRI time series data poses unique challenges, such as choosing appropriate baselines and validating biological relevance.

Besides IG, several other feature attribution methods have been developed for explaining deep learning model decisions. DeepLIFT (Deep Learning Important FeaTures) compares neuron activations to a reference state and computes importance scores based on activation

differences (21). Input X Gradient (IXG) provides a first-order approximation of feature importance by element-wise multiplication of input and gradient values (22). Guided Backpropagation (GB) modifies the standard backpropagation to produce clearer visualizations by suppressing negative gradients (23). While these methods have shown promise in computer vision and natural language processing, their ability to robustly extract features from spatiotemporal functional neuroimaging data remains unknown.

To address the challenges in validating feature attribution methods in neuroimaging, here we develop a comprehensive validation framework using two complementary simulation approaches. First, we employ recurrent neural networks (RNNs) to generate synthetic fMRI data with precisely controlled parameters. Second, we utilized The Virtual Brain (TVB) simulator (24) to create biophysically informed simulations incorporating realistic brain connectivity from both human (25) and mouse connectomes (26). This dual approach enables validation across different scales and species while maintaining known ground truth, bridging the gap between abstract RNN models and biologically plausible neural dynamics.

In both approaches, we systematically investigated how various neurophysiological factors – including noise levels, prevalence of brain regions affected by excitation/inhibition (E/I) imbalance, and severity of E/I alterations – impact feature attribution in neuroimaging analyses. By manipulating these parameters, we could simulate different scenarios of E/I imbalance, which is crucial for understanding neurological disorders such as autism (27) and schizophrenia (28). Our comprehensive framework allows us to assess the sensitivity and specificity of feature attribution methods in detecting E/I-related changes across a range of controlled conditions.

Our study has five key objectives, each addressing critical challenges in applying feature attribution methods to neuroimaging. First, we validate these methods using carefully constructed RNN simulations with known ground truth, testing for robustness by systematically varying the prevalence of regions with E/I imbalance (1-10% of nodes), severity of these alterations (δ = 0.1-0.5), and signal-to-noise ratios (-10dB to 10dB). Second, we extend this validation using The Virtual Brain simulator, which incorporates biologically realistic connectomes from both human (68 regions) and mouse (426 regions) brains, allowing us to test feature attribution at different anatomical scales and across species. Third, we perform the first systematic comparison of multiple feature attribution methods (Integrated Gradients, DeepLIFT, Input X Gradient, Guided Backprop) in the context of neuroimaging, addressing a significant gap even in simpler domains like structural imaging. Fourth, we examine how baseline choice affects feature attribution in dynamic brain imaging - a crucial methodological question unique to time series data that has not been previously addressed. Finally, we apply these validated methods to the multisite autism brain imaging data exchange (ABIDE) resting-state fMRI dataset (N=834) to identify key brain regions distinguishing individuals with autism from controls, testing whether our simulation-based validation translates to real-world clinical applications.

A key challenge in clinical neuroscience research is in developing biologically grounded simulations that can validate analytical methods while maintaining clear links to testable hypotheses about brain dysfunction. This challenge is particularly acute when trying to bridge cellular-level mechanisms with whole-brain dynamics observable in clinical populations (29). To address this gap, we designed neurobiological simulations based on the excitation/inhibition (E/I) imbalance theory, which has emerged as a fundamental framework for understanding various neurological and psychiatric disorders (30). E/I imbalance has been

extensively implicated in both autism and schizophrenia pathophysiology (30, 31). This dysregulation manifests through altered GABAergic signalling and regional changes in neurotransmitter concentrations, as demonstrated by converging evidence from cellular studies and magnetic resonance spectroscopy (32). In autism specifically, disrupted E/I balance appears to be a key mechanism underlying core symptoms (27, 33). The evidence spans multiple scales of investigation: from cellular studies showing altered GABAergic signalling in animal models, to human neuroimaging revealing regional GABA concentration changes (34, 35).

Critically, these E/I alterations show anatomical specificity, particularly affecting nodes of the default mode network (DMN) (36), encompassing the posterior cingulate cortex (PCC), precuneus (PCun), angular gyrus (AnG) and ventromedial prefrontal cortex (PFC). This network plays a central role in in self-referential processing, social cognition, and autobiographical memory retrieval - cognitive functions that are notably impaired in autism spectrum disorders and schizophrenia (36, 37, 38).

This framework provides an ideal test case for validating feature attribution methods for several reasons. First, E/I imbalance represents a well-defined cellular mechanism that can be explicitly modelled in neural simulations (39). Second, its effects manifest across multiple spatial scales, from synaptic alterations to large-scale network dysfunction observable in fMRI (40). Third, the anatomical specificity of these alterations, particularly within the DMN, provides clear predictions for feature attribution methods to identify.

By selectively perturbing E/I balance in specific brain regions implicated in autism, particularly DMN nodes like the posterior cingulate cortex, we simulated disease-like brain

dynamics with known ground truth alterations. This approach allows us to systematically test whether feature attribution methods can reliably identify affected regions across different scales and species, providing a crucial bridge between cellular mechanisms and clinical biomarkers.

Results

RNN simulations and validation

We first validated feature attribution methods using recurrent neural network (RNN) simulations, which allowed systematic control over key parameters while maintaining temporal dynamics characteristic of fMRI data. These simulations generated synthetic fMRI time series from two groups: one with typical network dynamics and another with hyperexcitable nodes, mimicking localized E/I imbalance.

Model design and parameter space

The RNN model generated time series data for 100 regions, with adjustable parameters controlling: signal-to-noise ratio (SNR: -10dB to 10dB), prevalence of affected regions (1-10% of nodes), and severity of hyperexcitability ($\delta = 0.1$-0.5). Network connectivity was defined using small-world topology to reflect brain-like architecture, while regional dynamics were modelled using nonlinear activation functions convolved with a canonical hemodynamic response (see Methods).

Classifier performance

Our stDNN classifier demonstrated robust performance across simulation conditions (**Fig S1**). Even at high noise levels, with 5% prevalence, and small hyperexcitability (δ = 0.1), the classifier achieved perfect discrimination (accuracy = 100%) (-10dB SNR) (**Fig S1A**). Performance remained robust down to 1% prevalence, though convergence required more training epochs (**Fig S1B**). The classifier successfully detected group differences across hyperexcitability levels above δ = 0.1, with faster convergence at higher severities (**Fig S1C**).

Feature attribution analysis

IG successfully identified hyperexcited nodes across all simulation conditions (**Fig 2**).To establish a consistent threshold across feature attribution methods and simulation parameters, we normalized all attributions by dividing it by its maximum value across ROIs. With a normalized attribution threshold of 0.5, IG achieved perfect detection (F1 score = 100%) of affected network nodes even under challenging conditions: high noise (-10dB SNR, **Fig 2A-C**), low prevalence (1%, **Fig 2D-F**), and subtle hyperexcitability (δ = 0.1, **Fig 2G-I**). Remarkably, IG's performance remained stable across these conditions, suggesting robust feature detection even with minimal signal.

These RNN simulation results established that IG can reliably identify affected brain regions across a wide range of simulation conditions.

Human TVB Simulations

We next validated feature attribution methods using biophysically informed simulations with The Virtual Brain (TVB), incorporating realistic human brain connectivity and dynamics.

These simulations provided a more stringent test of our approach while maintaining known ground truth alterations in specific brain regions.

Biophysical model implementation

Using structural connectivity from human diffusion imaging (68-region Desikan-Killiany atlas(41)), we simulated whole-brain dynamics with explicit modelling of excitatory and inhibitory populations in each region. We created two groups by introducing E/I imbalance in one group through altered inhibitory synaptic conductance ($\Delta Qi = 0.1\text{-}0.5nS$): first targeting only bilateral posterior cingulate cortex (PCC), then progressively including precuneus (PCun) and angular gyrus (AnG) to test effects of anatomical extent on classification and IG performance.

Classifier performance

The stDNN classifier achieved robust discrimination between groups across simulation conditions (**Fig S2**). For large imbalance in PCC-only, classification accuracy remained above 90% for SNR levels down to -10dB (**Fig S2A**). Convergence time decreased with increasing anatomical extent of E/I imbalance, with fastest classification when alterations included PCC, PCun, and AnG (**Fig S2B**). The classifier showed sensitivity to E/I imbalance severity, with reliable discrimination (accuracy > 90%) for $\Delta Qi \geq 0.3nS$ (**Fig S2C**).

Feature attribution results

IG successfully identified regions with altered E/I balance across conditions (**Fig 3**). With PCC-only alterations, IG showed high attributions specifically localized in the bilateral PCC across SNR levels (**Fig 3A-C**). When additional regions were affected, IG accurately detected the expanded anatomical extent of alterations (**Fig 3D-F**), maintaining precise localization

288 even as the number of affected regions increased. Attribution patterns remained stable across
289 E/I imbalance severities (**Fig 3G-I**), though stronger alterations (ΔQi = 0.5nS) produced more
290 pronounced attribution scores. Importantly, regions without simulated E/I alterations
291 consistently showed attribution scores below our normalized attribution threshold of 0.5,
292 demonstrating high specificity in feature detection.
293
294 These TVB simulation results demonstrate that feature attribution methods can reliably
295 identify biologically realistic E/I alterations in the context of whole-brain dynamics and
296 authentic human brain connectivity. The successful detection of affected regions across
297 varying anatomical distributions and E/I imbalance severities supports the method's potential
298 utility for identifying circuit alterations in clinical populations.
299
300 ***Mouse TVB simulations***
301
302 To test the generalizability of our approach across species and scales, we conducted parallel
303 analyses using the mouse connectome. Unlike human diffusion imaging-based connectivity,
304 the Allen Brain Atlas provides invasively validated, directional connectivity data across 426
305 regions, offering higher spatial resolution and more direct measurement of anatomical
306 connections.
307
308 *Model implementation*
309 We adapted our TVB simulations to the mouse brain, maintaining the same biophysical
310 parameters for excitatory and inhibitory populations while incorporating mouse-specific
311 structural connectivity. E/I imbalance was introduced in the retrosplenial cortex (RSC), the

key DMN hub in rodents analogous to the human PCC. We systematically varied inhibitory conductance (ΔQi = 0.1-0.5nS) to parallel our human simulations.

Classifier performance

The stDNN classifier showed robust but distinct performance patterns with mouse data (**Fig S3**). Classification accuracy remained above 90% for moderate noise levels (0dB SNR) with RSC-only alterations (**Fig S3A**). Performance improved when E/I imbalance extended to additional regions (RSC+Cg, RSC+Cg+PrL; **Fig S3B**), though requiring more training epochs than for human simulations. The classifier maintained sensitivity to E/I imbalance severity, achieving reliable discrimination for ΔQi ≥ 0.3nS (**Fig S3C**).

Feature attribution results

IG successfully identified affected regions of the mouse brain, though with some species-specific patterns (**Fig 4**). With RSC-only alterations, IG showed enhanced attributions precisely localized in all RSC subregions only for SNR levels down to 0dB (**Fig 4A-C**). When E/I imbalance extended to additional regions, IG accurately detected the broader anatomical distribution (**Fig 4D-F**), maintaining precision and recall above 80%. Hierarchical attribution strengths, varying across subregions, were observed here due to more structurally distinct regions affected here than in human TVB simulations. This pattern held across E/I imbalance severities (**Fig 4G-I**), suggesting robust detection of regional hierarchies in the mouse brain.

The successful cross-species validation, particularly the detection of analogous regions (RSC/PCC) despite different scales and connectivity patterns, demonstrates the robustness of our feature attribution approach. The higher spatial resolution of the mouse connectome

337 suggests the potential relevance of the stDNN and IG approach to high-dimensional data and
338 to reveal additional insights about subregional specificity while maintaining reliable detection
339 of E/I alterations.
340
341 ***ABIDE fMRI dataset analysis***
342
343 After validating our approach using simulated data, we applied these methods to the Autism
344 Brain Imaging Data Exchange (ABIDE) dataset to identify brain regions that reliably
345 distinguish individuals with autism from neurotypical controls. This multi-site dataset
346 (N=834; 419 autism, 415 controls) provided a robust test of our methods in real-world
347 clinical data.
348
349 *Classification performance*
350 Our stDNN achieved significant classification accuracy across the ABIDE dataset, reliably
351 distinguishing autism from control participants. Performance remained stable across different
352 site splits and cross-validation folds, demonstrating robustness to data collection variability.
353 Notably, classification accuracy was maintained even with highly heterogeneous
354 symptomatology across autistic individuals and multi-site imaging data (4).
355
356 *Feature attribution analysis*
357 Feature attribution analysis revealed a consistent set of discriminative brain regions (**Fig 5**).
358 IG identified the bilateral posterior cingulate cortex (PCC) and right precuneus as the regions
359 most strongly contributing to classification, with attribution scores above our normalized
360 attribution threshold of 0.5. This finding was remarkably stable across multiple analyses as
361 described below.

The convergence between our empirical findings and simulation results provides computational support for the role of E/I imbalance in autism, particularly within the DMN. The identification of these regions is consistent with previous studies showing altered DMN function in autism and adds mechanistic insight by suggesting these alterations may reflect underlying E/I imbalance.

Comparison of feature attribution methods

To establish the reliability of our approach, we systematically compared multiple feature attribution methods across both simulated and empirical data. We evaluated four methods: Integrated Gradients (IG), DeepLIFT, Input X Gradient (IXG), and Guided Backpropagation (GB), each representing different theoretical approaches to explaining model decisions.

RNN simulations

In our controlled RNN simulations, IG, DeepLIFT, and IXG showed remarkably consistent performance (**Fig 6A**). These methods demonstrated high correlations in their attribution patterns ($r > 0.99$) and achieved similar accuracy as measured by F1score in identifying hyperexcitable nodes (**Fig S4A**). In contrast, GB showed weaker correlations with other methods ($r <= 0.42$) and impaired F1 scores (**Fig S4A**) due to reduced precision (**Fig S4B**) but not recall (**Fig S4C**) in detecting affected regions.

Human TVB simulations

In human TVB data, IG, DeepLIFT, and IXG maintained strong correlations ($r > 0.97$, **Fig 6B**) in identified regions with affected E/I balance. Weaker correlations were seen

between attribution profiles of GB and other methods (r <= 0.39). Poorer F1 scores (**Fig S4D**) at identifying affected nodes due to lower precision (**Fig S4E**) but not recall (**Fig S4F**) was also observed for GB compared to other methods.

Mouse TVB simulations

Similar patterns emerged in mouse TVB simulations, where these methods showed high agreement (r > 0.99, **Fig 6C**) despite the increased anatomical complexity of the mouse connectome. In contrast, GB showed weaker correlations with other methods (r = 0.77) and reduced F1 score due to low precision in detecting affected regions (**Fig S4G-I**).

ABIDE fMRI data

In the ABIDE dataset, IG, DeepLIFT, and IXG identified highly consistent sets of discriminative regions (r > 0.98, **Fig 6D**). All three methods highlighted the bilateral PCC and right precuneus as key regions distinguishing autism from controls. GB showed markedly different attribution patterns (r < 0.06), suggesting potential limitations in its clinical applications.

These results demonstrate that IG, DeepLIFT, and IXG provide consistent and reliable feature attribution across different experimental conditions and scales of brain organization, while GB shows divergent patterns and reduced accuracy, suggesting it may be less suitable for neuroimaging applications.

Impact of baseline choice in Integrated Gradients

The choice of baseline represents a critical methodological consideration when applying Integrated Gradients to neuroimaging data. Unlike static image analysis where a black (zero) image serves as a natural baseline, the optimal reference point for dynamic brain data is not immediately obvious. We therefore systematically evaluated multiple baseline choices including the zero, mean, median across the ROIs of the input timeseries, and random across our simulation frameworks and empirical data.

RNN simulations

In RNN simulations, zero, mean, median, and random baselines showed remarkably high correlations in their attribution patterns (r => 0.99, **Fig 7A**) and achieved equivalently good performance in identifying affected regions (**Fig S5A-C**).

Human TVB simulations

This consistency extended to the more complex TVB simulations, where these three baselines maintained strong correlations despite added biological complexity (r >= 0.94 for human, **Fig 7B**). The preservation of attribution patterns across baseline choices suggests robust feature detection rather than methodological artifacts. In contrast, random baselines showed weaker correlations with other methods (r < 0.49) and reduced F1 scores in detecting affected regions (**Fig S5D**) due to both impaired precision (**Fig S5E**) and recall (**Fig S5F**) compared to other baseline choices.

Mouse TVB simulations

On higher-dimensional mouse TVB simulations yielded high correlations between mean and median baselines (r = 0.93) as well as between zero and random normal baselines (r =0.82). Weaker correlations (r <= 0.64) were found between other baseline pairs (**Fig 7C**). As with

RNN simulations, no significant differences across baseline choices were found as to their accuracy scores at identifying affected regions (**Fig S5G-I**).

ABIDE fMRI data

Baseline robustness was further confirmed in empirical data, with zero, mean, and median baselines showing highly correlated attribution patterns in the ABIDE dataset ($r > 0.99$, **Fig 7D**). Importantly, random baselines consistently showed poorer performance across all conditions, with significantly lower correlations with other baselines (typically $r <= 0.96$).

These results demonstrate that zero, mean, and median baselines all provide stable and consistent feature attribution in neuroimaging applications, but random baselines should be avoided due to their poor reproducibility.

Discussion

DNN models have emerged as powerful tools for analyzing functional MRI data, offering unprecedented accuracy in predicting cognitive states and clinical outcomes (3, 4). However, their inherent complexity and "black box" nature have limited our ability to determine which neuroanatomical features and neurophysiological mechanisms drive their predictions. This opacity poses a significant challenge in neuroscience, where interpretability is crucial for advancing our understanding of brain function and dysfunction. To address this, we investigated whether xAI techniques, particularly Integrated Gradients (IG), could reliably identify relevant brain regions underlying accurate prediction by stDNNs applied to fMRI data. Unlike conventional approaches that rely on pre-engineered features, stDNNs operate

directly on high dimensional fMRI time series data, making the interpretation of their decision-making particularly challenging.

Through a comprehensive validation framework combining recurrent neural network models and whole-brain simulations informed by empirical biophysics and connectivity, we demonstrated that IG can robustly identify brain regions with altered excitation/inhibition (E/I) balance. Our validation analysis showed consistent performance across multiple attribution methods with their different baselines, and simulation approaches, establishing the reliability of these techniques for analyzing fMRI data. Critically, this robust performance extended from simplified neural models to complex whole-brain simulations and across species.

The successful application of these validated methods to empirical fMRI data from the ABIDE dataset revealed key regions within the default mode network that distinguish individuals with autism from neurotypical controls. This convergence between computational predictions and clinical findings demonstrates the potential of xAI to bridge the gap between cellular-level mechanisms giving rise to E/I imbalance and whole-brain dynamics observable in clinical populations. Together, these results demonstrate that xAI methods can provide reliable, biologically meaningful insights into the neural basis of neurodevelopmental disorders while maintaining the advantages of deep learning approaches.

Validation of feature attribution methods using RNN simulations

We first employed RNNs to generate synthetic fMRI data under controlled conditions. These RNN simulations provided a simple framework for systematic validation of feature

attribution methods while maintaining key properties of fMRI time series. The RNN model allowed precise control over signal-to-noise ratio, prevalence of affected regions, and severity of alterations in node excitability (**Fig 2**). The performance of IG proved remarkably robust across these varied conditions. IG maintained perfect detection (F1 score = 1) of affected regions even with high noise (-10dB SNR), low prevalence (1% of regions), and subtle alterations in E/I balance ($\delta = 0.1$) (**Fig 2**). This systematic validation established the wide reliability of IG for detecting altered network nodes across a range of conditions.

Validation of feature attribution methods using human connectome and biophysical whole-brain simulations

Building on the RNN based validation, we employed The Virtual Brain simulator to test feature attribution methods in a more biologically constrained context. TVB simulations incorporated realistic brain connectivity from human and mouse connectomes, allowing us to model E/I imbalance within the context of actual brain network architecture (**Fig 3**). We used TVB to generate two classes of datasets: one representing a neurotypical population and another mimicking individuals with autism. The model incorporated structural connectivity data from the Human Connectome Project (HCP), widely used in numerous studies (25). This approach allowed us to simulate whole-brain dynamics that more closely approximate real fMRI data. The key difference between the two simulated groups was the introduction of excitation/inhibition (E/I) imbalances in specific regions of the default mode network (DMN) for the clinical autism-like group. This design was informed by work on autism (42) and aligns with the growing body of evidence pointing to altered DMN dynamics in autism (36).

509 IG successfully identified regions where we introduced E/I imbalance, particularly in the
510 posterior cingulate cortex and related DMN regions (**Fig 3**). This performance remained
511 robust across different severity levels of E/I imbalance and numbers of affected regions.
512 Across levels of E/I imbalance prevalence and severity, IG achieved near-perfect feature
513 identification (F1 scores = 1) at high noise levels (-10dB SNR). This robustness to noise is
514 particularly crucial in fMRI analysis, where signal quality can vary significantly between
515 individuals due to biophysical and motion artefacts.
516
517 Consistent with our statistical model results, IG also demonstrated high sensitivity to
518 different prevalence rates of affected regions. It successfully identified relevant inputs even at
519 low prevalence rates of 1%, underscoring its potential to detect subtle group differences. This
520 capability is especially relevant for heterogeneous conditions like autism, where affected
521 brain regions may vary across individuals.
522
523 **Validation of feature attribution methods using mouse connectome and biophysical**
524 **whole-brain simulations**
525
526 To test the generalizability of our findings across species and scales, we conducted parallel
527 analyses using mouse connectome data. The mouse analyses offered several unique
528 advantages. Unlike the human HCP DTI-based atlas, the Allen Brain Atlas provides
529 experimentally validated, high-resolution directional connectivity data based on anterograde
530 viral injections and whole-brain imaging across 426 regions in the mouse brain (26), offering
531 a detailed view of axonal projections and inter-regional connections. This higher granularity
532 (426 versus 68 regions in humans) and direct measurement of anatomical connectivity

provided a stringent test of our feature attribution methods at a finer scale of brain organization.

Using this detailed mouse connectome, we generated two groups of simulated data: one modelling typical wild-type dynamics and another modelling autism-like alterations based on empirical findings from mouse models (40, 43). Specifically, we introduced E/I imbalances in the retrosplenial cortex, which serves as the key DMN hub in rodents analogous to the posterior cingulate cortex in humans (38). This parallel between the retrosplenial cortex and posterior cingulate cortex is supported by both anatomical studies and functional evidence showing similar patterns of connectivity and task-related deactivation (44, 45, 46, 47, 48, 49).

Our analyses revealed striking similarities in the performance of feature attribution methods across species. IG successfully identified altered E/I balance in the retrosplenial cortex with precision comparable to its performance in human simulations (**Fig 4**). However, we also observed some species-specific differences - while IG achieved 100% precision in identifying affected regions.

The successful cross-species validation has important implications for translational research. First, it demonstrates that our feature attribution approach can detect mechanistic alterations despite substantial differences in brain size, organization, and measurement scale. Second, the consistency between mouse and human findings provides computational support for the use of mouse models to study E/I imbalance in autism and related disorders. Finally, the higher resolution of the mouse connectome allowed us to validate our methods in higher-dimensional data at a scale closer to the underlying circuit mechanisms, strengthening the biological plausibility of our findings.

558

559 **Comparing feature attribution methods**

560

561 A critical challenge in applying deep learning to neuroimaging is determining whether
562 different feature attribution methods yield consistent and biologically meaningful results.
563 Surprisingly, despite their widespread use in computer vision and other domains, there has
564 been no systematic comparison of feature attribution methods even for simpler brain imaging
565 applications such as those involving static structural MRI images. This gap is particularly
566 notable for dynamic functional neuroimaging, where the temporal complexity of the data
567 poses additional challenges.

568

569 We evaluated four prominent feature attribution methods: Integrated Gradients (IG),
570 DeepLIFT (Deep Learning Important FeaTures), Input X Gradient (IXG), and Guided
571 Backpropagation (GB). These methods were selected based on several key criteria: (a) they
572 can be applied to any pre-trained DNN model, offering broad applicability across various
573 network architectures and tasks; (b) these methods require only one pass of the
574 backpropagation algorithm, making them computationally inexpensive compared to
575 approaches that necessitate complex input perturbations; and (c) they effectively address
576 common challenges in gradient-based attribution, such as gradient saturation and
577 discontinuities of gradient arising because of activation functions like sigmoid and ReLU
578 units. Each method approaches the attribution problem differently: IG accumulates gradients
579 along a path from baseline to input (20), DeepLIFT compares activation patterns to a
580 reference state (21), IXG directly examines input-gradient products (22), and GB modifies
581 the backpropagation process to enhance visualization (21). These methods were chosen

because they represent distinct theoretical approaches to feature attribution and have shown success in other domains (50).

Our systematic comparison revealed several key findings **(Fig 5**). First, we observed remarkably high correlations ($r > 0.95$) between IG, DeepLIFT, and IXG attribution patterns across all our simulated conditions - from simple RNN models to complex TVB simulations. This consistency held true for both human and mouse connectomes, suggesting these methods are capturing fundamental features of the network dynamics rather than methodological artifacts.

In contrast, Guided Backpropagation showed notably weaker correlations with other methods ($r < 0.4$ in human data, $r < 0.77$ in mouse data) and reduced accuracy in identifying ground truth affected regions in our simulations. This divergence likely reflects GB's focus on positive contributions to the gradient, which may not account for complex interactions in brain networks. These findings align with theoretical work suggesting GB may not satisfy important attribution method axioms (20).

The relative performance of these methods was also remarkably consistent across different simulation conditions. In both RNN and TVB simulations, as well as ABIDE fMRI data, IG, DeepLIFT, and IXG displayed strongly correlated attribution profiles. On simulated time series, these methods maintained high precision in identifying regions with altered E/I balance across various noise levels, prevalence rates, and severity of E/I imbalance. This robustness across conditions strengthens confidence in their reliability for neuroimaging applications.

These findings have several important implications. First, they establish that multiple theoretically-grounded attribution methods can robustly and reliably identify relevant features in complex neuroimaging data. Second, they suggest that the choice between IG, DeepLIFT, and IXG may not be critical fMRI applications. Third, they indicate that GB, despite its popularity in computer vision, may be less suitable for neuroimaging analyses where capturing complex network interactions is crucial.

Impact of baseline choice on feature attribution in dynamic brain imaging

A critical and previously unexamined question in applying feature attribution methods to neuroimaging is the choice of baseline for computing attributions. This represents a key contribution of our study, as the baseline issue is particularly challenging for dynamic functional brain imaging. Unlike static image analysis where a black (zero) image serves as a natural baseline, or structural brain imaging where baseline choices are more straightforward, the optimal baseline for temporally complex fMRI signals is not obvious.

The baseline choice is theoretically crucial for Integrated Gradients, as it defines the reference point from which feature importance is accumulated along the integration path(20). In fMRI analysis, several candidates present themselves: a zero baseline (analogous to the black image in computer vision), the mean or median of the time series (capturing central tendency), or random noise (testing attribution robustness). Each choice embodies different assumptions about what constitutes an "absence of signal" in brain activity data.

We systematically evaluated these baseline choices across our simulation frameworks. In the RNN simulations, zero, mean, and median baselines showed remarkably high correlations in

their attribution patterns (r > 0.99, **Fig 6A**) and achieved equivalent performance in identifying affected regions. This consistency extended to the more complex TVB simulations, where these three baselines maintained strong correlations (r > 0.94 for human, **Fig 6B**; r > 0.64 for mouse, **Fig 6C**) despite the added biological complexity. High correlations between attribution profiles across baselines were also found in the ABIDE fMRI data (r > 0.99). This convergence in real-world data, where ground truth is unknown, provides strong support for the biological relevance of the underlying attribution patterns.

These findings have several important implications. First, they demonstrate that IG's performance in fMRI analysis is robust across reasonable baseline choices, reducing one source of methodological uncertainty. Second, they suggest that zero, mean, or median baselines can all serve as valid reference points for computing attributions in fMRI data. Third, they caution against the use of random baselines, which may introduce unnecessary variability into the attribution process. Our results provide clear guidance for future studies applying feature attribution methods to fMRI data, ensuring reproducible and interpretable results in neuroimaging applications.

Alignment between human TVB simulations and empirical ABIDE fMRI data

We applied our feature attribution methods to both human TVB simulations with controlled excitability and E/I imbalance, and to the ABIDE fMRI dataset. This dual approach aimed to identify brain regions that most reliably distinguish individuals with autism from neurotypical controls, while also providing a computational framework to interpret these findings. Our analysis revealed several key insights that both validate our approach and enhance our understanding of the neurobiology of autism.

In the ABIDE fMRI data, our feature attribution analysis consistently identified regions within the DMN as key drivers of discrimination between autism and controls, with particularly strong contributions from the bilateral posterior cingulate cortex and right precuneus (**Fig 5**). The robustness of these findings was demonstrated across multiple feature attribution methods (IG, DeepLIFT, and IXG), which showed highly correlated attribution patterns ($r > 0.95$, **Fig 6D**). This consistency across methods, in a real-world dataset where the ground truth is unknown, provides support for the biological relevance of these attributions. Importantly, these findings were stable across different baseline choices ($r > 0.96$, **Fig 7D**), suggesting they reflect genuine biological features rather than methodological artifacts.

Our TVB simulations, informed by empirical findings, demonstrated that E/I imbalance in the PCC leads to network-level alterations detectable by our feature attribution methods (**Fig 3**). The emergence of the same regions in our empirical data analysis suggests that altered brain arising from E/I imbalance in these DMN nodes could indeed play a role in autism pathophysiology. This alignment between computational predictions and clinical findings provides a mechanistic framework supporting the E/I imbalance theory of autism, while also pointing to specific anatomical loci for this dysfunction.

Several aspects of our findings align with and extend previous research. The prominent role of the PCC in our analysis is consistent with prior studies showing altered DMN function in autism (36, 51), reduced GABA concentrations in this region (52), and correlations between PCC dysfunction and social cognitive deficits (51). However, our approach provides new

insight into how these regional alterations may drive broader network dysfunction observable in fMRI data.

The high attribution scores for the precuneus, particularly in the right hemisphere, also align with previous findings but suggest a more specific role in autism pathophysiology than previously recognized. This region's involvement in self-referential processing and social cognition (38) makes it a plausible locus for autism-related alterations, and our results suggest it may be particularly important for distinguishing autism from typical development.

These findings have several important clinical implications. First, they suggest that feature attribution methods could help develop more objective biomarkers for autism, potentially aiding in diagnosis or treatment monitoring. Second, the specificity of our findings to particular DMN nodes might help guide the development of targeted interventions to alter E/I imbalance. Third, the robustness of these findings across different attribution methods and baselines suggests they could be reliable enough for clinical applications.

Conclusion

Our study demonstrates that explainable AI methods, including Integrated Gradients, can reliably identify biologically meaningful features in fMRI data analyzed by deep learning models. Through systematic validation using both recurrent neural networks and biophysically detailed whole-brain simulations using human and mouse connectomes, we identify methods can robustly detect brain regions affected by altered excitation/inhibition balance across various experimental conditions and species. This validation framework represents a key advance in applying feature attribution methods to dynamic neuroimaging

data, addressing crucial methodological questions about baseline choice and cross-method consistency.

The convergence between computational predictions and analysis of the empirical fMRI data provides compelling evidence linking cellular-level E/I imbalance to whole-brain dynamics in autism. Our finding that multiple feature attribution methods consistently identify key nodes of the default mode network, particularly the posterior cingulate cortex and precuneus, strengthens confidence in these regions' role in autism pathophysiology. Importantly, the successful cross-species validation using mouse connectomes demonstrates the potential for bridging insights between animal models and human studies.

More broadly, our work establishes a general framework for using explainable AI to test mechanistic hypotheses about brain dysfunction in neuropsychiatric disorders. The ability to link cellular-level alterations to network-level biomarkers opens new possibilities for understanding disease mechanisms and developing targeted interventions. This work represents an important step toward more interpretable and biologically informed applications of artificial intelligence in clinical neuroscience.

Materials and Methods

We validate the accuracy of the relevant inputs retrieved by feature attribution methods using two different simulation models for synthesizing fMRI timeseries. The first simulation model is a simple recurrent neural network model where the differences between two classes is simulated using different functional connectivity between the brain regions as well as the local excitatory/inhibitory (E/I) imbalances in some subset of brain regions. The primary

purpose of this simple model is to investigate the performance of the feature attribution methods as a function of prevalence rates and SNRs and to find the optimal thresholds on the feature attribution scores to separate the most relevant brain features from the irrelevant ones. The second model is a more realistic biophysical model for autism which is implemented using The Virtual Brain stimulator. The differences between the two classes is created by introducing differences in the local E/I imbalances in some set of brain regions. These regions are chosen based on our recent findings (4, 42) in the autism population. We simulate timeseries of nodes (or ROIs) consisting of two classes which differ either with respect to the connectivity of ROIs in the two classes or differ with respect to the local excitability of some chosen nodes. We then train a classifier to predict the class label given the timeseries of the nodes. Finally, we examine how each feature attribution method discovers important nodes or features in each class. We hypothesize that these feature attribution methods would correctly identify the nodes that differ in functional connectivity or and their local excitability between the classes.

Simulated fMRI data with a recurrent neural network model

We simulated the whole-brain fMRI timeseries by the time series of 100 nodes or regions of interest (ROIs). To generate simulated fMRI data of two classes, we assumed that the difference between the two classes lies in their local excitability on some chosen nodes. We assumed that the network consists of 100 nodes defined using an adjacency matrices A of a randomly sampled connected Watts–Strogatz small-world graphs (53), each of the 100 nodes joined with its $k = 10$ nearest neighbors in a ring topology with a probability $p = 0.1$ of rewiring. The local level of excitability for the two classes, W_0 and W_1, are defined as $W_k = \Lambda_k A^T$, where $\Lambda_0 = \frac{1}{\rho} I_n$, $\Lambda_1 = \begin{pmatrix} \left(\frac{1+\delta}{\rho}\right) I_a & 0 \\ 0 & \frac{1}{\rho} I_{n-a} \end{pmatrix}$, I_a and I_{n-a} are respectively the $a \times a$

and the $(n-a) \times (n-a)$ the identity matrices, $n = 100$ is the total number of nodes, a is the number of nodes affected by the difference in local excitability, ρ is the spectral radius of the matrix A, and $\delta > 0$ measures the amplitude of the difference between the two classes. We simulate several fMRI datasets with varying degrees of differences in their local excitability. We quantify these differences using (1) a prevalence rate that describes the ratio of the number of nodes that differ in local excitability to the overall number of nodes (i.e. the ratio $\frac{a}{n}$), and (2) the severity of the hyper-excitability defined as the ratio of local difference in excitability in affected nodes to the level of excitability (i.e. the ratio $\frac{\frac{1+\delta}{\rho} - \frac{1}{\rho}}{\frac{1}{\rho}} = \delta$). We hypothesize that the difficulty of finding the discriminating nodes increases when the prevalence rate or the hyper-excitability level drops.

We simulated fMRI data using a Recurrent Neural Network (RNN) convolved with a canonical hemodynamic response function (HRF), which can be described with the following dynamical system:

$$x_t = tanh(Wx_{t-1} + \epsilon_t) \tag{1}$$

$$y = HRF * x + \omega \tag{2}$$

Where $W \in \{W_0, W_1\}$ is the recurrent weight matrix, y_t represents the BOLD fMRI signal of multiple regions of interests (ROIs) at time t, x_t represents hidden neural variables giving rise to the BOLD signal at time t, $\epsilon_t = N(0, \sigma_\epsilon)$ represents the intrinsic neural noise, $\omega_t = N(0, \sigma_\omega)$ represents the measurement noise which is independent across regions, HRF is a canonical hemodynamic response function with a sampling period of 1s (54).

We chose σ_ω to achieve some fixed level of Signal to Noise Ratio (SNR) as follows:

$$\sigma_\omega = \frac{\rho}{10^{0.1*SNR}} \tag{3}$$

where σ_ω is the variance of the noise and ρ is the maximal variance of the signal across ROIs that we empirically computed after simulating x_t dynamics. An example of this simulated data for each of the two classes over time can be seen in **Fig 1**.

We investigated the performance of various feature attribution methods with the prevalence rates ranging from 1 to 10% in steps of 1%, the hyper-excitability amplitude ranging from 0.0 to 0.5 in steps of 0.1, and the SNR ranging from -10dB to 10dB in steps of 10dB.

Simulated fMRI data with The Virtual Brain Simulator

We used a more realistic biophysical model to generate synthetic fMRI datasets for two classes representing the neurotypical and autism populations to investigate the performance of the feature attribution methods. Specifically, we simulated fMRI data using a model of whole-brain dynamics where biophysical parameters can be varied at the regional level. The model was implemented in The Virtual Brain (TVB) simulator (24, 55) in previous work (56). Briefly, each simulated region comprises two neural populations, one of excitatory (E) and one of inhibitory (I) neurons, whose mean firing rates across E and I populations are described by the following differential equations:

$$\tau \frac{dx_i^E}{dt} = -x_i^E + F_E(Q_E(x_i^E + J_i), x_i^I) + \nu_i^E \tag{4}$$

$$\tau \frac{dx_i^I}{dt} = -x_i^I + F_I\left(Q_E(x_i^E + RJ_i), x_I^i\right) + \nu_i^I \tag{5}$$

$$J_i(t) = \sum_{j \neq i} W_{ij} x_j^E \left(t - \frac{D_{ij}}{s}\right) \tag{6}$$

$$y_i = HRF * (r x_i^E + (1 - r) x_i^I) + \omega_i \tag{7}$$

Where x_i^E and x_i^I are respectively the firing rates of excitatory and inhibitory populations in region i, J_i represents the input into region i of other regions long-range excitatory connections between regions, $R = 1.4$ is the ratio between the conductance of long-range

excitatory-to-excitatory and excitatory-to-inhibitory synapses, W and D characterize respectively the number and the length of white-matter fibres between regions between regions both empirically derived from synaptic tracing in the mouse brain, $s = 3m/s$ represents the speed of transmission between regions, $Q_E = 1\ nS$ and Q_I represents respectively the synaptic strength of excitation and inhibition, ν_E and ν_I are standard normal integration noise, the transfer functions F_E and F_I which relates each population's output rate to the input it receives from itself and other populations were fitted to reproduce the input-output relations of integrate-and-fire neurons in spiking simulations (57), y represents the BOLD fMRI signal of multiple regions of interests (ROIs), $\omega_i = N(0, \sigma_\omega)$ represents the measurement noise in region i, HRF is a canonical hemodynamic response function. As previously we chose σ_ω to achieve some fixed level of Signal to Noise Ratio (SNR) as follows:

$$\sigma_\omega = \frac{\rho}{10^{0.1*SNR}} \tag{8}$$

where σ_ω is the variance of the noise and ρ is the maximal variance of the signal across ROIs that we empirically compute after simulation.

For the human brain, tract weights and lengths were obtained with diffusion tensor imaging of a healthy adult's brain (56) (data publicly available at https://zenodo.org/records/4263723). The connectome was partitioned according to the commonly used 68-ROI Desikan-Killiany-Tourville atlas (58). For the mouse brain, the structural connectome was taken from a tracer-derived structural connectome of the mouse brain from the Allen Brain Institute (26) with a 426-ROI parcellation. Inter-regional synaptic delays were modelled based on the matrix of tract lengths from the same source (26). Synaptic weights and delays were taken to be the same from the left and right brain hemispheres.

825

826 The datasets for the autism class was generated by introducing E/I imbalance in a subset of
827 ROIs. Specifically, the biophysical parameter Q_I capturing inhibitory synaptic conductance
828 was modified specifically in regions of the default mode network where excitation/inhibition
829 imbalance has been observed in humans and rodent models of autism and hypothesized to
830 underlie symptoms. For the human brain, Q_I was altered bilaterally progressively in the
831 posterior cingulate cortex, in the precuneus (PCun) and in the angular gyrus (AnG). In the
832 mouse brain, Q_I was locally modified progressively in the bilateral agranular, ventral, and
833 dorsal subdivisions of the retrosplenial cortex (RSC), in the cingulate cortex (Cg) and in the
834 prelimbic cortex (PrL).

835

836 We simulated time series of 100 s for all ROIs for each connectome. This procedure was
837 repeated with 100 different random seeds that affected initial conditions and random noise
838 realizations. To simulate fMRI data, the firing rate time series were then convolved with a
839 canonical hemodynamic response function with TR of 1s. To test the robustness of the xAI
840 methods across different noise levels, in both human and mouse models, measurement noise
841 was added to the ROI timeseries generated using the above simulation models. To quantify
842 the impacts of severity of E/I imbalance, inhibitory conductance in affected regions was
843 varied between 4.5 nS and 4.9 nS to obtain severity varying from ΔQi = 0.1nS to ΔQi =
844 0.5nS between affected and unaffected regions. To investigate the effects of prevalence,
845 imbalance was implemented in different subsets of DMN regions of varying size – in the
846 human model, we compared imbalance bilaterally in the PCC only, in the PCC and PCun
847 only, and in the PCC, PCun and AnG, while in the mouse model, we compared imbalance
848 bilaterally in the RSC only, the RSC and Cg only, and the RSC, Cg, and PrL.

849

850 ***ABIDE dataset***

851 We leveraged neuroimaging and phenotypic data from the Autism Brain Imaging Data
852 Exchange (ABIDE; http://fcon_1000.projects.nitrc.org/indi/abide/) (1, 2). The dataset
853 consists 419 subjects with autism and 415 typically developing subjects from 14 sites. The
854 criterion for the subject selection is given in (4).

855

856 ***stDNN classifier model for the simulated fMRI datasets***

857 We used a two-layer 1D convolutional neural network model for classifying the two groups.
858 Each class consists of 400 subjects where each subject is defined by a matrix of dimensions
859 $M \times 100$ where M is number of nodes and 100 is the length of timeseries (i.e. 1 minute and
860 40 seconds for a TR of 1s). Our stDNN model consists of two one-dimensional convolutional
861 block layers, a temporal averaging operation, a dropout layer, and a sigmoid layer as shown
862 in **Fig 1C**. Each convolutional block consists of a 1D convolutional layer followed by a
863 rectified linear activation layer. Each convolutional block is followed by a maxpool layer
864 which reduces the temporal dimension by 2. We used 80% of the data to train the model, the
865 remaining 20% for testing the model. The code for the model implemented in PyTorch is
866 available at https://github.com/scsnl/Strock_Nghiem_bioRxiv_2025.

867

868 ***Feature attribution methods***

869 We obtained the discriminating features from each simulated subject by using a feature
870 attribution method which determines how much each nodes or ROI contributes to the
871 decision made by our classifier for this subject. The simulated fMRI data being temporal,
872 each of these feature attribution methods provides us with a value for each ROI and for each
873 time step of simulation. We obtain the influence of an ROI by taking the absolute value of the
874 median of the feature attribution across time steps. To evaluate the performance of the

attribution method we normalize the attribution by dividing it by the maximum threshold across ROI, and compare ROI with attribution above 0.5 with ROI affected E/I imbalance using standard recall, precision and F1 metrics. We first evaluated the performance of Integrated Gradient (IG) (20) on the simulated datasets. We examined how IG performance may depend on the choice of baseline which is required for evaluating the path integral for the feature attribution. It is not clear what an optimal baseline is for the neuroimaging applications. Therefore, we investigated several baselines including zero, random, mean, and median of all fMRI time series across the subjects. We finally compared the performance of other gradient base feature attribution methods such as DeepLIFT (DL) (21), Input X Gradient (IXG) (22), Guided Backprop (GB) (23) implemented in the Captum python library.

Integrated Gradients

IG directly uses the gradient of the classifier with respect to the input ROIs to define the feature attribution. This method can be used for any already trained DNN model with a single backpropagation pass. It integrates this gradient over a path in the fMRI input space going to the subject's simulated fMRI data from a baseline of the same dimension. Formally, IG attribution method can be written as:

$$\phi(x) = (x - x') \times \int_{\alpha=0}^{1} \frac{\partial F\big(x' + \alpha(x - x')\big)}{\partial x_i} d\alpha \tag{9}$$

where x is the original input, x' is the chosen baseline from which we start our integration, F represents the function associating fMRI input to its clinical category, and ϕ is the function giving the feature attribution score for a given node x. IG overcomes the problems associated with the simple gradient based approaches such as the noisy gradients and gradient saturation. As IG requires a baseline, we compared the feature attribution obtained for various baselines: (1) the median of the signal across nodes and samples, (2) the mean of the signal across nodes

and samples, (3) the default all-zero baseline, and (4) a random sample drawn from a normal distribution centred about zero with the variance reflecting the variance of the input.

Deep learning important features (DeepLIFT)

As IG, DeepLIFT is feature attribution method that can applied to any pretrained DNN model with a single backpropagation pass. This method computes the importance of each input in the prediction of the output. This importance for each input is computed in terms of the difference between the output for a reference or the baseline and the given input. As in the IG, the baseline needs to be chosen based on the domain. It is developed to address the issues with the basic gradient based feature attribution approaches. As in the case of IG, we evaluate the performance of DeepLIFT for all the baselines as defined above.

Input X Gradient (IXB)

IXB is also a gradient based attribution method used for obtaining feature attributions. This is a very basic method where gradient of the input with respect to the target is computed and then the gradient is multiplied with the input to obtain the feature attribution scores. If both the gradient of the input and also its magnitude is high then that feature's attribution is high. Both IG and DeepLIFT methods are refinements over this approach.

Guided Backpropagation (GB)

This method is developed for explaining the decisions taken by a CNN model by computing the gradient of the output with respect to the input. This method requires a modification of the gradient in the ReLU layers where the gradient to these layers are made zero if the input to the ReLU during the forward pass is negative. This approach is applicable to only the convolutional networks.

6. References

1. Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. Mol Psychiatry. 2019;24(11):1583-98.
2. Kriegeskorte N, Golan T. Neural network models and deep learning. Curr Biol. 2019;29(7):R231-R6.
3. Supekar K, de Los Angeles C, Ryali S, Kushan L, Schleifer C, Repetto G, et al. Robust and replicable functional brain signatures of 22q11.2 deletion syndrome and associated psychosis: a deep neural network-based multi-cohort study. Mol Psychiatry. 2024;29(10):2951-66.
4. Supekar K, Ryali S, Yuan R, Kumar D, de Los Angeles C, Menon V. Robust, Generalizable, and Interpretable Artificial Intelligence-Derived Brain Fingerprints of Autism and Social Communication Symptom Severity. Biol Psychiatry. 2022;92(8):643-53.
5. Vieira S, Pinaya WH, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neurosci Biobehav Rev. 2017;74(Pt A):58-75.
6. Colliot O. Machine Learning for Brain Disorders. 2023.
7. Wang J, Knol MJ, Tiulpin A, Dubost F, de Bruijne M, Vernooij MW, et al. Gray Matter Age Prediction as a Biomarker for Risk of Dementia. P Natl Acad Sci USA. 2019;116(42):21213-8.
8. Zaharchuk G, Gong E, Wintermark M, Rubin D, Langlotz CP. Deep Learning in Neuroradiology. AJNR Am J Neuroradiol. 2018;39(10):1776-84.
9. Gadgil S, Zhao Q, Pfefferbaum A, Sullivan EV, Adeli E, Pohl KM. Spatio-Temporal Graph Convolution for Resting-State fMRI Analysis. Med Image Comput Comput Assist Interv. 2020;12267:528-38.

10. Ngo GH, Khosla M, Jamison K, Kuceyeski A, Sabuncu MR. Predicting individual task contrasts from resting-state functional connectivity using a surface-based convolutional network. Neuroimage. 2022;248:118849.

11. Zhang Y, Farrugia N, Bellec P. Deep learning models of cognitive processes constrained by human brain connectomes. Med Image Anal. 2022;80:102507.

12. Lei D, Qin K, Pinaya WHL, Young J, Van Amelsvoort T, Marcelis M, et al. Graph Convolutional Networks Reveal Network-Level Functional Dysconnectivity in Schizophrenia. Schizophr Bull. 2022;48(4):881-92.

13. Wang L, Li K, Hu XP. Graph convolutional network for fMRI analysis based on connectivity neighborhood. Netw Neurosci. 2021;5(1):83-95.

14. Ryali S, Supekar K, Chen T, Kochalka J, Cai W, Nicholas J, et al. Temporal Dynamics and Developmental Maturation of Salience, Default and Central-Executive Network Interactions Revealed by Variational Bayes Hidden Markov Modeling. PLoS Comput Biol. 2016;12(12):e1005138.

15. Lurie DJ, Kessler D, Bassett DS, Betzel RF, Breakspear M, Kheilholz S, et al. Questions and controversies in the study of time-varying functional connectivity in resting fMRI. Network Neuroscience. 2020;4(1):30-69.

16. Calhoun VD, Miller R, Pearlson G, Adali T. The Chronnectome: Time-Varying Connectivity Networks as the Next Frontier in fMRI Data Discovery. Neuron. 2014;84(2):262-74.

17. Ryali S, Zhang Y, de Los Angeles C, Supekar K, Menon V. Deep learning models reveal replicable, generalizable, and behaviorally relevant sex differences in human functional brain organization. Proc Natl Acad Sci U S A. 2024;121(9):e2310012121.

18. Eitel F, Schulz MA, Seiler M, Walter H, Ritter K. Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research. Exp Neurol. 2021;339.

19. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e50.

20. Sundararajan M, Taly A, Yan QQ. Axiomatic Attribution for Deep Networks. Pr Mach Learn Res. 2017;70.

21. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. Pr Mach Learn Res. 2017;70.

22. Kindermans P-J, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, et al. The (un) reliability of saliency methods. Explainable AI: Interpreting, explaining and visualizing deep learning. 2019:267-80.

23. Springenberg J, Dosovitskiy A, Brox T, Riedmiller M, editors. Striving for Simplicity: The All Convolutional Net. ICLR (workshop track); 2015.

24. Sanz-Leon P, Knock SA, Woodman MM, Domide L, Mersmann J, McIntosh AR, et al. The Virtual Brain: a simulator of primate brain network dynamics. Front Neuroinform. 2013;7:10.

25. Glasser MF, Smith SM, Marcus DS, Andersson JLR, Auerbach EJ, Behrens TEJ, et al. The Human Connectome Project's neuroimaging approach. Nature Neuroscience. 2016;19(9):1175-87.

26. Oh SW, Harris JA, Ng L, Winslow B, Cain N, Mihalas S, et al. A mesoscale connectome of the mouse brain. Nature. 2014;508(7495):207-+.

27. Rubenstein JL, Merzenich MM. Model of autism: increased ratio of excitation/inhibition in key neural systems. Genes Brain Behav. 2003;2(5):255-67.

28. Uhlhaas PJ, Singer W. Neuronal dynamics and neuropsychiatric disorders: toward a translational paradigm for dysfunctional large-scale networks. Neuron. 2012;75(6):963-80.

29. Breakspear M. Dynamic models of large-scale brain activity. Nat Neurosci. 2017;20(3):340-52.

1000 30. Sohal VS, Rubenstein JLR. Excitation-inhibition balance as a framework for
1001 investigating mechanisms in neuropsychiatric disorders. Mol Psychiatr. 2019;24(9):1248-57.
1002 31. Foss-Feig JH, Adkinson BD, Ji JL, Yang G, Srihari VH, McPartland JC, et al.
1003 Searching for Cross-Diagnostic Convergence: Neural Mechanisms Governing Excitation and
1004 Inhibition Balance in Schizophrenia and Autism Spectrum Disorders. Biol Psychiatry.
1005 2017;81(10):848-61.
1006 32. Tang X, Jaenisch R, Sur M. The role of GABAergic signalling in neurodevelopmental
1007 disorders. Nat Rev Neurosci. 2021;22(5):290-307.
1008 33. Nelson SB, Valakh V. Excitatory/Inhibitory Balance and Circuit Homeostasis in
1009 Autism Spectrum Disorders. Neuron. 2015;87(4):684-98.
1010 34. Hollestein V, Poelmans G, Forde NJ, Beckmann CF, Ecker C, Mann C, et al.
1011 Excitatory/inhibitory imbalance in autism: the role of glutamate and GABA gene-sets in
1012 symptoms and cortical brain structure. Transl Psychiatry. 2023;13(1):18.
1013 35. Horder J, Petrinovic MM, Mendez MA, Bruns A, Takumi T, Spooren W, et al.
1014 Glutamate and GABA in autism spectrum disorder-a translational magnetic resonance
1015 spectroscopy study in man and rodent models. Transl Psychiatry. 2018;8(1):106.
1016 36. Padmanabhan A, Lynch CJ, Schaer M, Menon V. The Default Mode Network in
1017 Autism. Biol Psychiatry Cogn Neurosci Neuroimaging. 2017;2(6):476-86.
1018 37. Hu ML, Zong XF, Mann JJ, Zheng JJ, Liao YH, Li ZC, et al. A Review of the
1019 Functional and Anatomical Default Mode Network in Schizophrenia. Neurosci Bull.
1020 2017;33(1):73-84.
1021 38. Menon V. 20 years of the default mode network: A review and synthesis. Neuron.
1022 2023;111(16):2469-87.

1023 39. Lam NH, Borduqui T, Hallak J, Roque A, Anticevic A, Krystal JH, et al. Effects of
1024 Altered Excitation-Inhibition Balance on Decision Making in a Cortical Circuit Model. J
1025 Neurosci. 2022;42(6):1035-53.
1026 40. Pagani M, Barsotti N, Bertero A, Trakoshis S, Ulysse L, Locarno A, et al. mTOR-
1027 related synaptic pathology causes autism spectrum disorder-associated functional
1028 hyperconnectivity. Nat Commun. 2021;12(1).
1029 41. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An
1030 automated labeling system for subdividing the human cerebral cortex on MRI scans into
1031 gyral based regions of interest. Neuroimage. 2006;31(3):968-80.
1032 42. Supekar K, Uddin LQ, Khouzam A, Phillips J, Gaillard WD, Kenworthy LE, et al.
1033 Brain Hyperconnectivity in Children with Autism and its Links to Social Deficits. Cell Rep.
1034 2013;5(3):738-47.
1035 43. Antoine MW, Langberg T, Schnepel P, Feldman DE. Increased Excitation-Inhibition
1036 Ratio Stabilizes Synapse and Circuit Excitability in Four Autism Mouse Models. Neuron.
1037 2019;101(4):648-+.
1038 44. Alexander AS, Place R, Starrett MJ, Chrastil ER, Nitz DA. Rethinking retrosplenial
1039 cortex: Perspectives and predictions. Neuron. 2023;111(2):150-75.
1040 45. Chao TH, Lee B, Hsu LM, Cerri DH, Zhang WT, Wang TW, et al. Neuronal
1041 dynamics of the default mode network and anterior insular cortex: Intrinsic properties and
1042 modulation by salient stimuli. Sci Adv. 2023;9(7):eade5732.
1043 46. Coletta L, Pagani M, Whitesell JD, Harris JA, Bernhardt B, Gozzi A. Network
1044 structure of the mouse brain connectome with voxel resolution. Sci Adv. 2020;6(51).
1045 47. Gutierrez-Barragan D, Singh NA, Alvino FG, Coletta L, Rocchi F, De Guzman E, et
1046 al. Unique spatiotemporal fMRI dynamics in the awake mouse brain. Curr Biol.
1047 2022;32(3):631-44 e6.

48. Menon V, Cerri D, Lee B, Yuan R, Lee SH, Shih YI. Optogenetic stimulation of anterior insular cortex neurons in male rats reveals causal mechanisms underlying suppression of the default mode network by the salience network. Nat Commun. 2023;14(1):866.

49. Nghiem TE, Lee B, Chao TH, Branigan NK, Mistry PK, Shih YI, et al. Space wandering in the rodent default mode network. Proc Natl Acad Sci U S A. 2024;121(15):e2315167121.

50. Gevaert A, Rousseau AJ, Becker T, Valkenborg D, De Bie T, Saeys Y. Evaluating feature attribution methods in the image domain. Mach Learn. 2024;113(9):6019-64.

51. Lynch CJ, Uddin LQ, Supekar K, Khouzam A, Phillips J, Menon V. Default Mode Network in Childhood Autism: Posteromedial Cortex Heterogeneity and Relationship with Social Deficits. Biol Psychiat. 2013;74(3):212-9.

52. Oblak AL, Gibbs TT, Blatt GJ. Reduced GABAA receptors and benzodiazepine binding sites in the posterior cingulate cortex and fusiform gyrus in autism. Brain Res. 2011;1380:218-28.

53. Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’networks. Nature. 1998;393(6684):440-2.

54. Ryali S, Supekar K, Chen T, Menon V. Multivariate dynamical systems models for estimating causal interactions in fMRI. Neuroimage. 2011;54(2):807-23.

55. Sanz-Leon P, Knock SA, Spiegler A, Jirsa VK. Mathematical framework for large-scale brain network modeling in The Virtual Brain. Neuroimage. 2015;111:385-430.

56. Goldman JS, Kusch L, Aquilue D, Yalcinkaya BH, Depannemaecker D, Ancourt K, et al. A comprehensive neural simulation of slow-wave sleep and highly responsive wakefulness dynamics. Front Comput Neurosc. 2023;16.

57. Volo MD, Romagnoni A, Capone C, Destexhe A. Biologically Realistic Mean-Field Models of Conductance-Based Networks of Spiking Neurons with Adaptation. Neural Comput. 2019;31(4):653-80.

58. Klein A, Tourville J. 101 labeled brain images and a consistent human cortical labeling protocol. Front Neurosci-Switz. 2012;6.

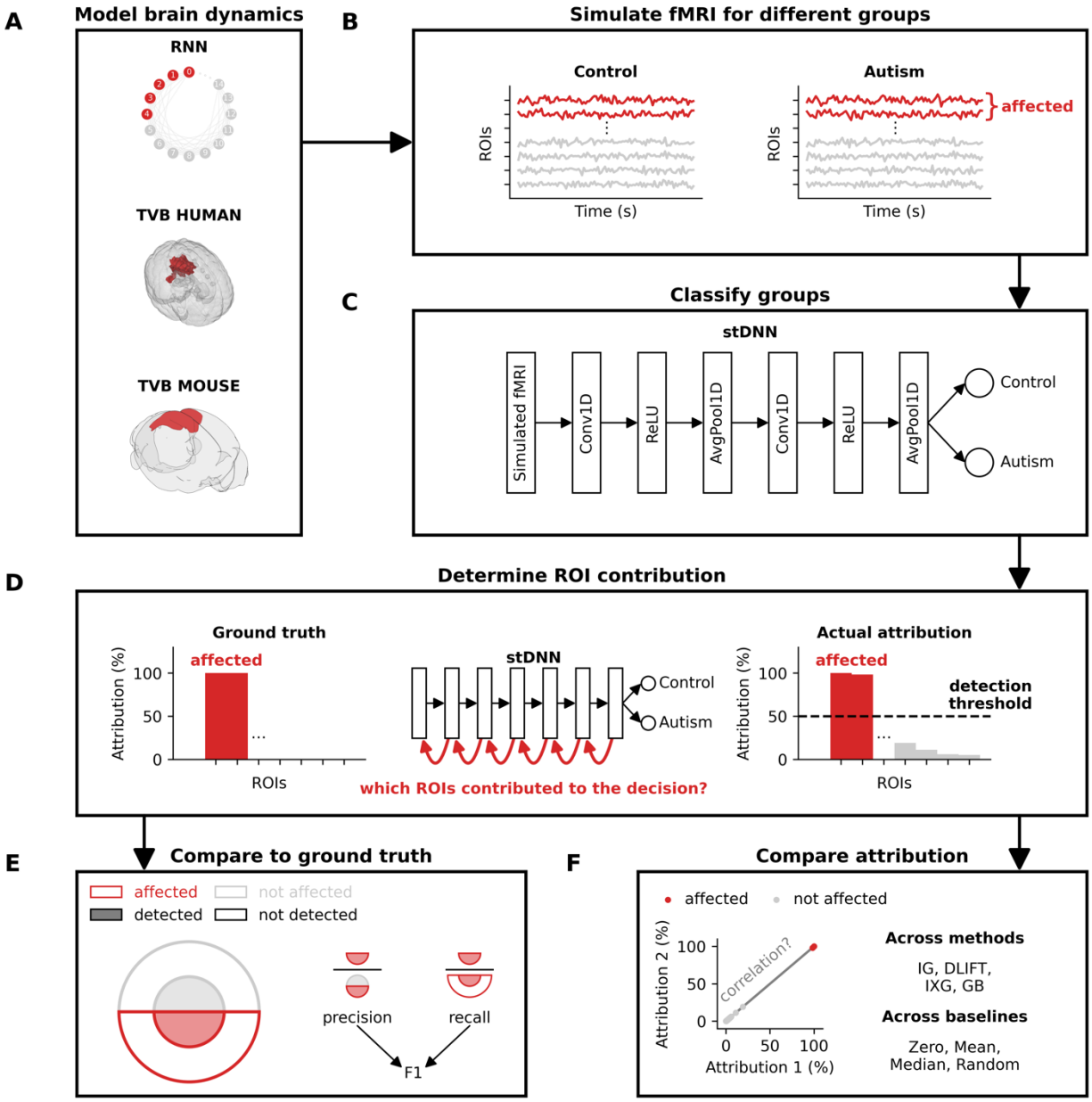

Figure 1. Overview of validation framework for xAI methods. Simulated multivariate fMRI datasets consisting of two classes are generated using two simulation models: RNNs and whole-brain simulations using The Virtual Brain. The two classes share the same structural connectome (i.e. adjacency matrix), but some nodes (highlighted in red) are hyper-excitable in the second population. A spatiotemporal deep neural network (stDNN) model consisting of 1D-CNN layers are trained to classify the two classes. The feature attribution methods are then applied on the trained model to discover the relative importance of each input for a given class.

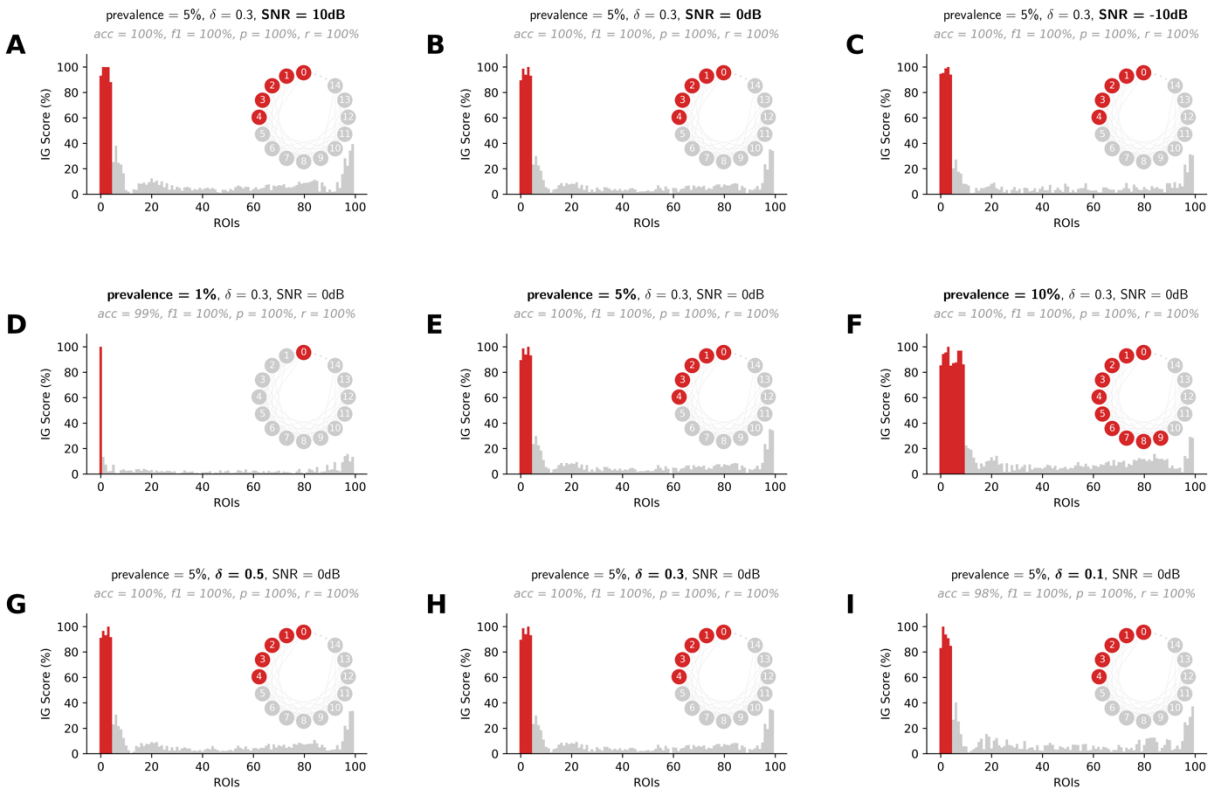

Figure 2. Performance of Integrated Gradients (IG) in detecting hyper-excitable ROIs in small-world RNNs across varying Signal-to-Noise Ratios (SNR), prevalence rates, and hyper-excitability severity. ROIs affected by hyper-excitability, highlighted in red, were retrieved robustly at all prevalence rates with a 100% F1-score. (**A-C**) IG scores for SNRs of 10, 0 and -10 dB respectively for prevalence rate of 5% and hyper-excitability severity δ=0.3. (**D-F**) IG scores for prevalent rates of 1, 5 and 10% respectively for SNR of 0 dB and hyper-excitability severity δ=0.3. (**G-I**) IG scores for hyper-excitability severity δ=0.5, δ=0.3, and δ=0.2 respectively for the given SNR of 0 dB and prevalence rate of 5%.

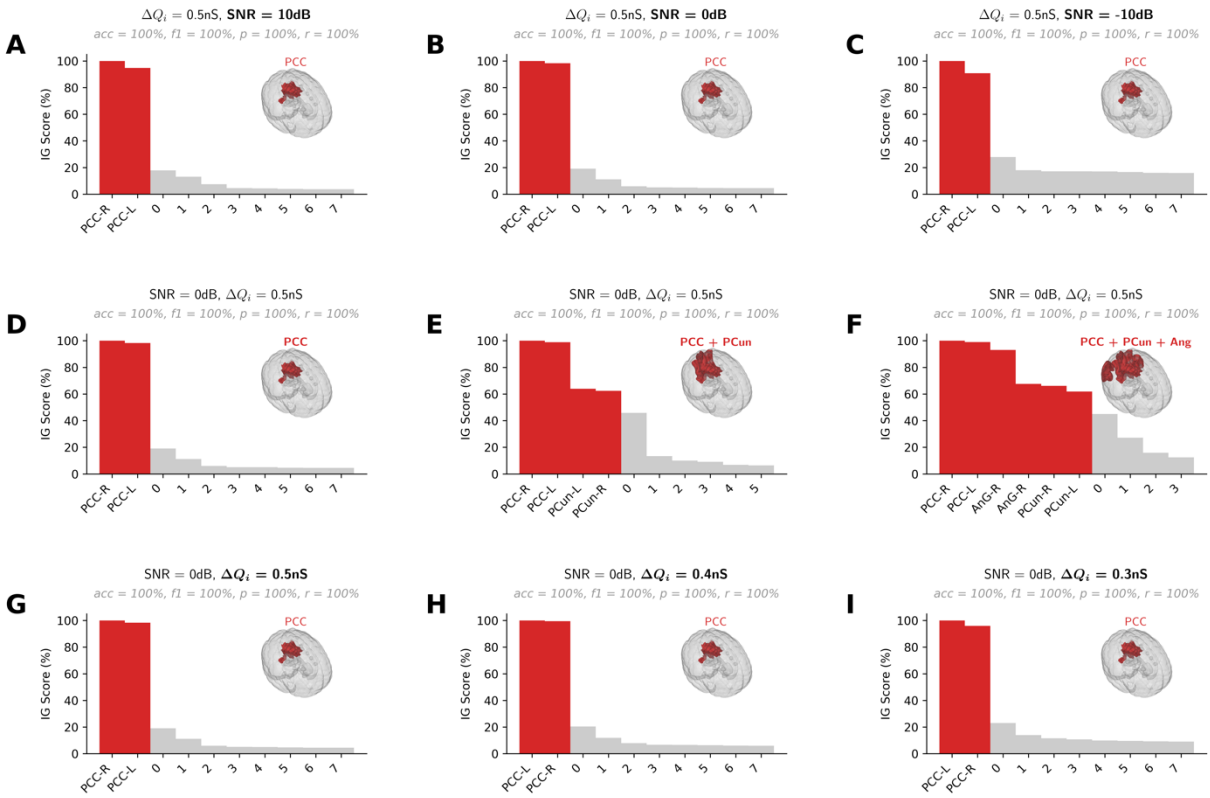

Figure 3. Performance of Integrated Gradients (IG) in detecting hyper-excitable ROIs in whole-brain TVB simulations with a human connectome across varying Signal-to-Noise Ratios (SNR), prevalence rates, and E/I imbalance severity. ROIs affected by E/I imbalance, highlighted in red, were retrieved robustly at all prevalence rates with a 100% F1-score. (**A-C**) IG scores for SNRs of 10, 0 and -10 dB respectively when only the PCC is affected by E/I imbalance with a severity ΔQi=0.5nS. (**D-F**) IG scores for E/I imbalance in the PCC, in the PCC and PCun, and in the PCC, PCun, and AnG respectively with ΔQi=0.5nS and SNR of 0 dB. (**G-I**) IG scores E/I imbalance affecting the PCC with severity ΔQi=0.5nS, ΔQi=0.4nS, and ΔQi=0.3nS respectively for the given SNR of 0 dB.

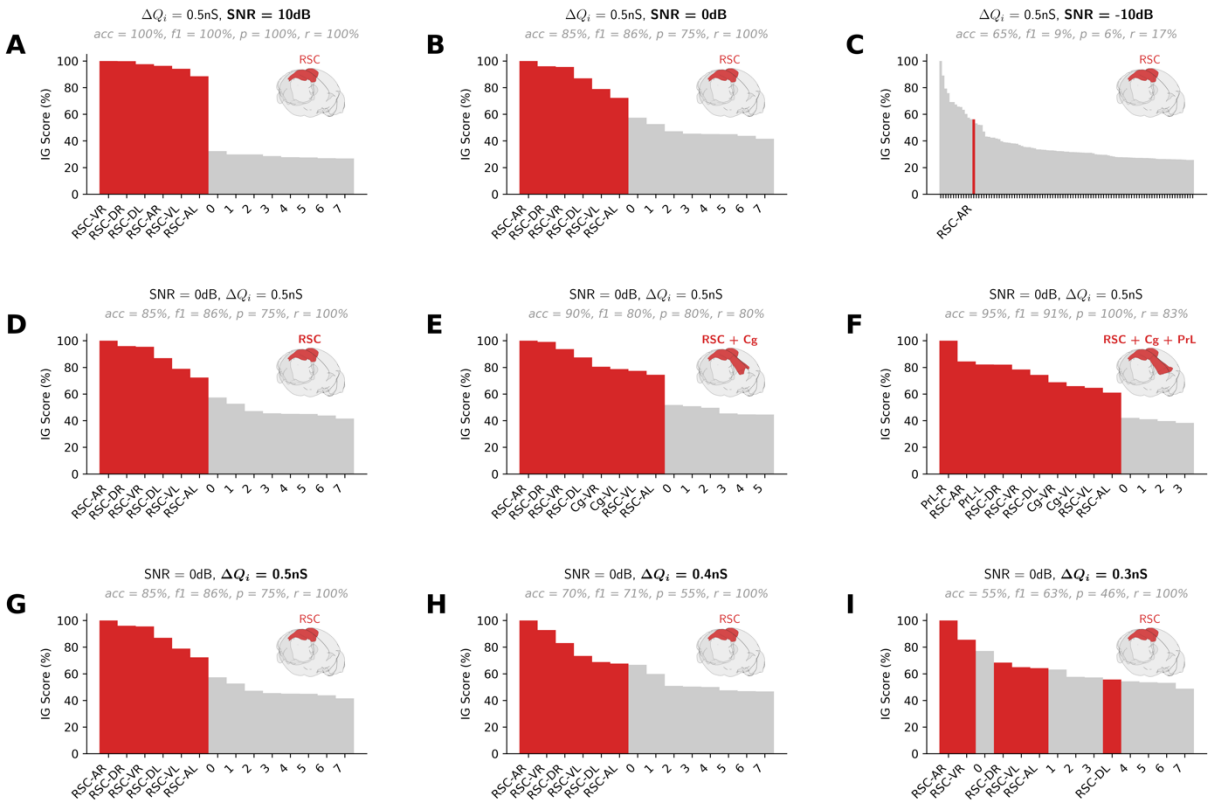

Figure 4. Performance of Integrated Gradients (IG) in detecting hyper-excitable ROIs in whole-brain TVB simulations with a mouse connectome across varying Signal-to-Noise Ratios (SNR), prevalence rates, and E/I imbalance severity. ROIs affected by E/I imbalance, highlighted in red, were retrieved robustly at all prevalence rates with a 100% F1-score. (**A-C**) IG scores for SNRs of 10, 0 and -10 dB respectively when only the RSC is affected by E/I imbalance with a severity ΔQi=0.5nS. (**D-F**) IG scores for E/I imbalance in the RSC, in the RSC and Cg, and in the RSC, Cg, and PrL respectively with ΔQi=0.5nS and SNR of 0 dB. (**G-I**) IG scores E/I imbalance affecting the RSC with severity ΔQi=0.5nS, ΔQi=0.4nS, and ΔQi=0.3nS respectively for the given SNR of 0 dB.

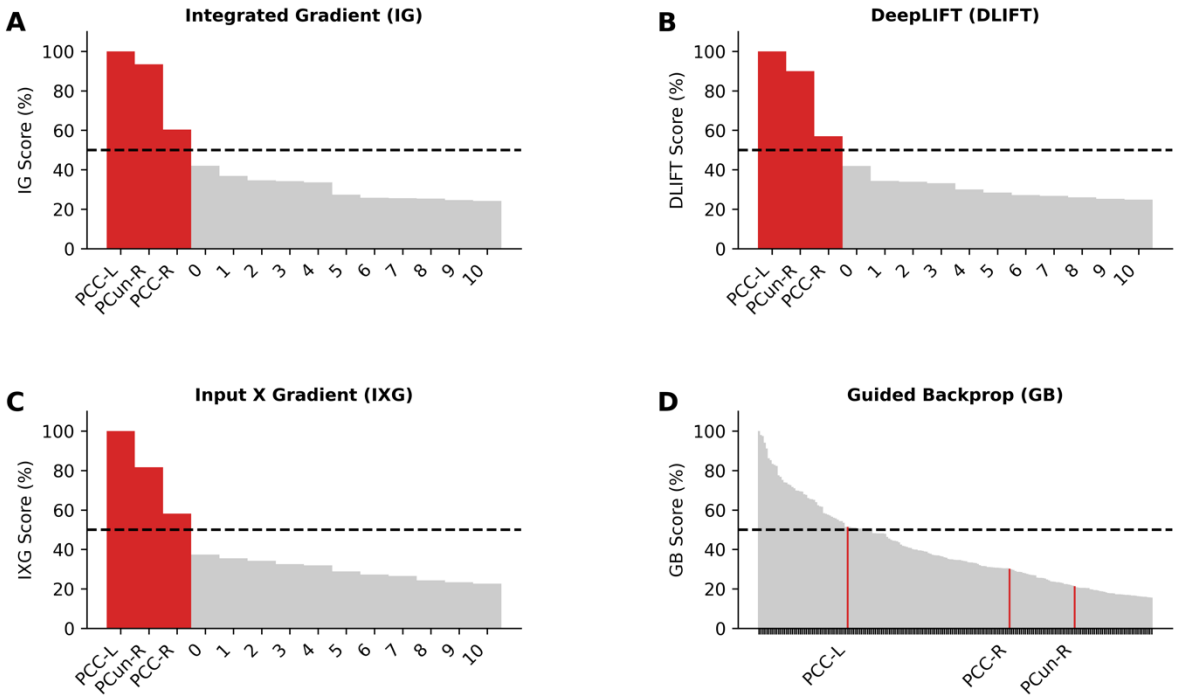

Figure 5. Feature Attribution Scores in ABIDE fMRI dataset. (**A**) IG, (**B**) DLIFT and (**C**) IXG show similar attributions. Bidirectional PCC and right PCun are the only regions with attribution above 0.5 across these attribution methods. (**D**) GB provides less differentiated attributions, i.e. more ROIs have high attribution. Bidirectional PCC and right PCun are not attributed as high by GB.

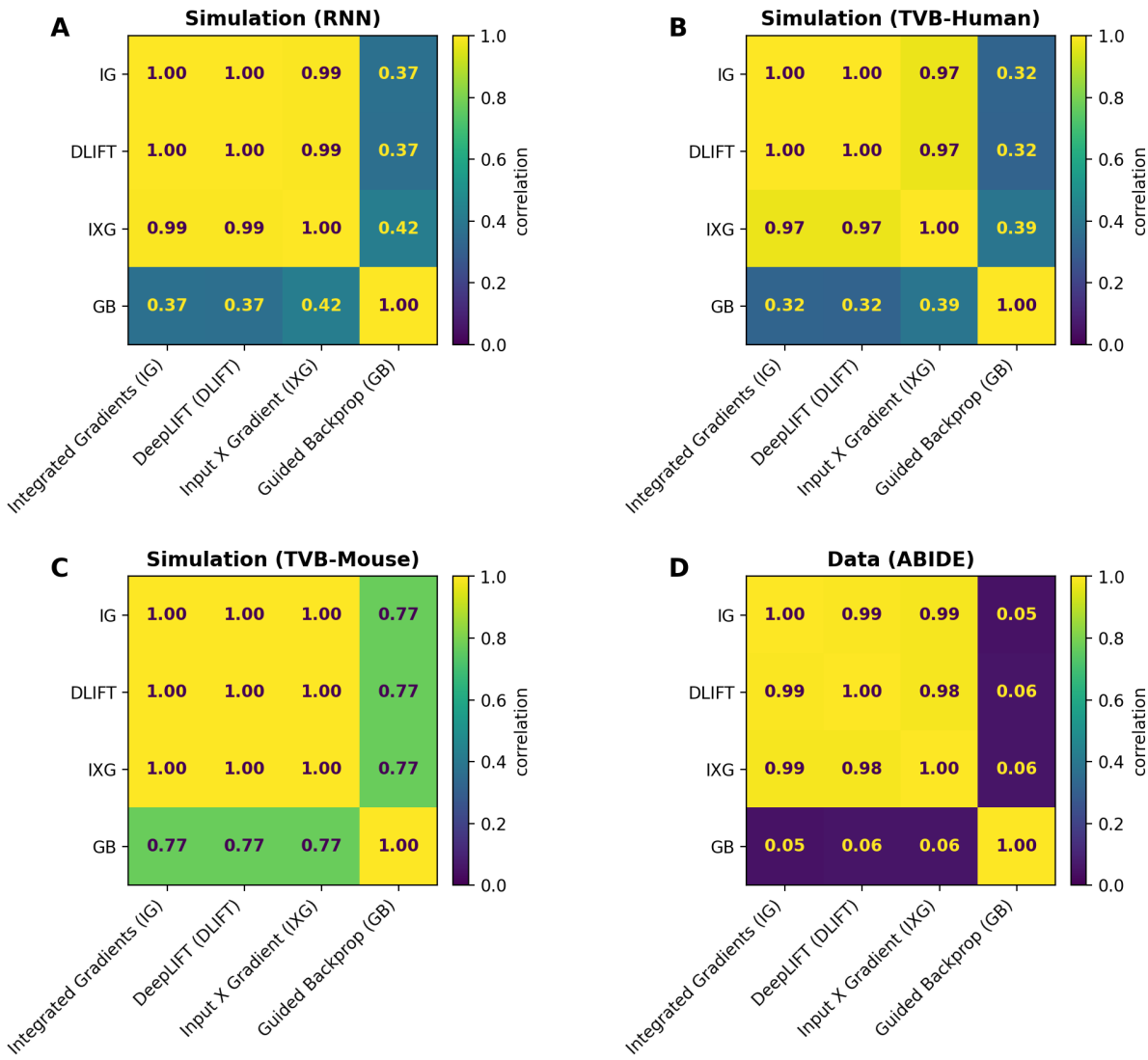

Figure 6. Comparison of feature attribution methods. (**A-D**) Correlations between the relevant input features discovered by each method in small world simulations (**A**), in TVB simulations with human (**B**) and mouse (**C**) connectomes, and in human data (**D**). The correlations between the input discovered by each pair of methods are high except for the Guided Backprop (GB) method.

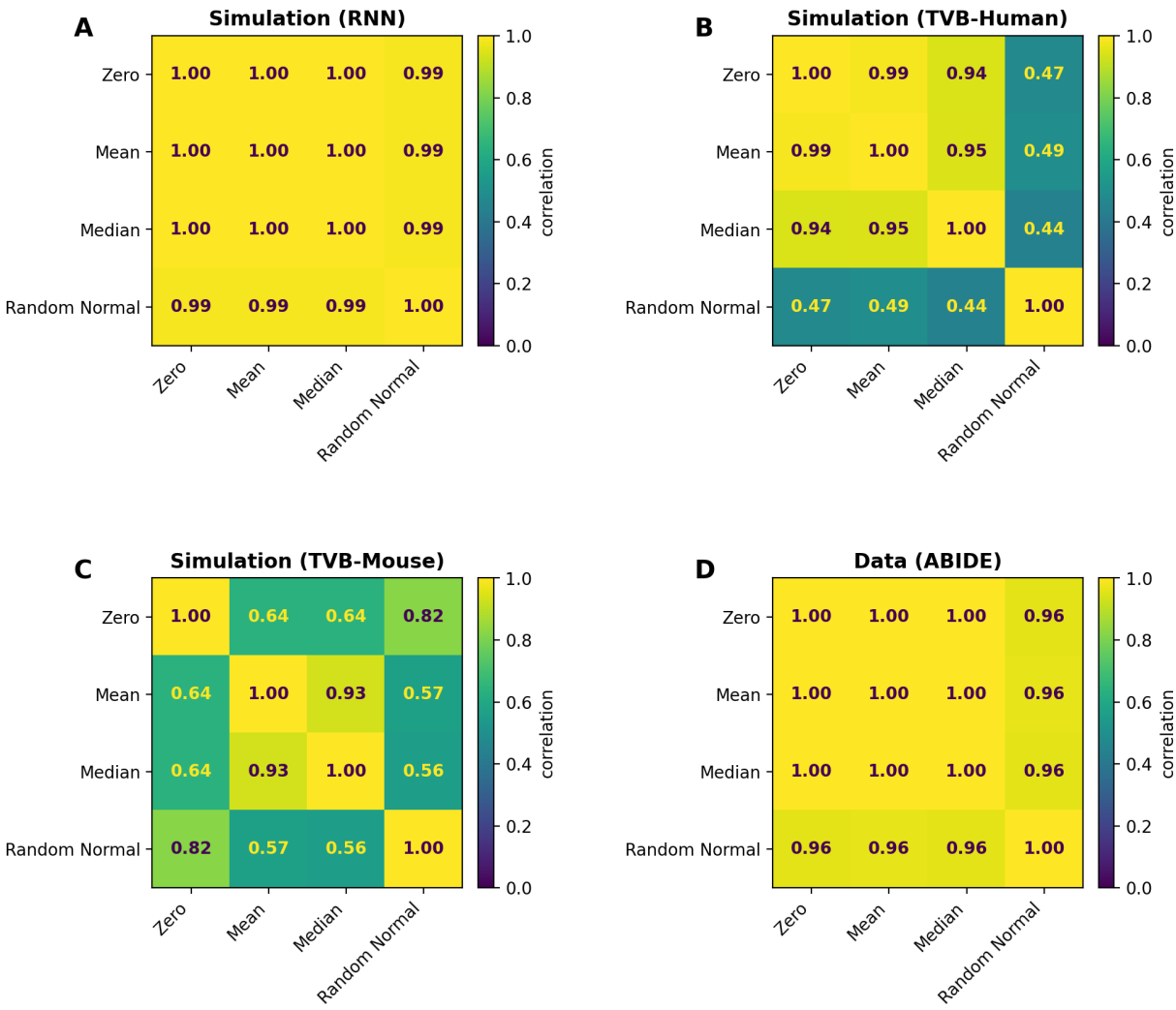

Figure 7. Comparison of the effects of baseline in IG-based feature attribution. (**A-D**) Correlations between the relevant input ROIs discovered by IG using the median, mean, zero and random noise as the baseline, in small world simulations (**A**), in TVB simulations with human (**B**) and mouse (**C**) connectomes, and in human data (**D**). The correlations between methods are high, suggesting that the performance of IG is robust to the baseline used.

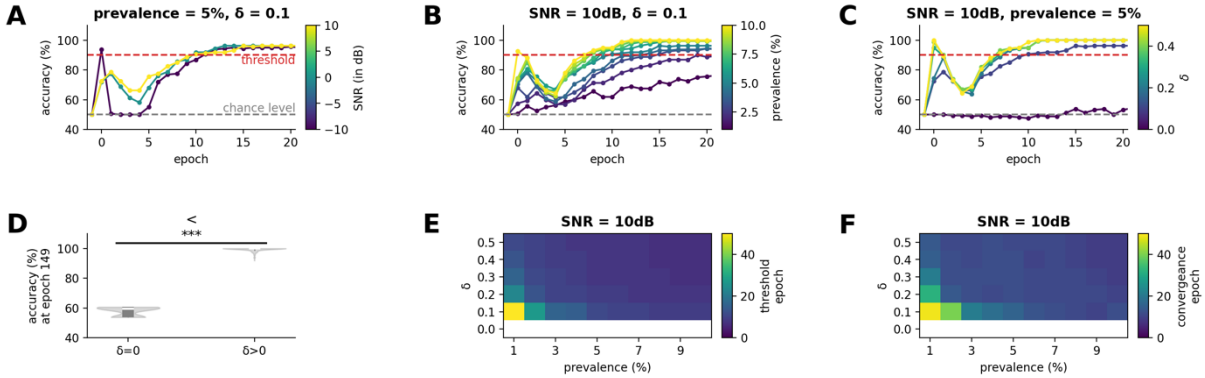

Figure S1. Training and classification performance on simulated data generated using RNNs. (**A-C**) Classification accuracy as a function of training epochs (**A**) for SNRs ranging from -10 dB to 10 dB with a fixed prevalence rate of 5% and hyper-excitability severity of δ=0.1, (**B**) for prevalence rates ranging from 1 to 10% with fixed SNR of -10 dB hyper-excitability severity of δ=0.1, and (**C**) for hyper-excitability severity varying from δ=0.0 to δ=0.5 with fixed SNR of -10 dB and prevalence rate of 5%. (**D**) Classification accuracy at epoch 149 for hyper-excitability severity of δ=0.0 to δ>0.0. (**E-F**) Training epoch during which a 90% accuracy threshold is passed (**E**) and convergence epoch (**F**) as a function of prevalence rate and hyper-excitability severity for an SNR of 10 dB.

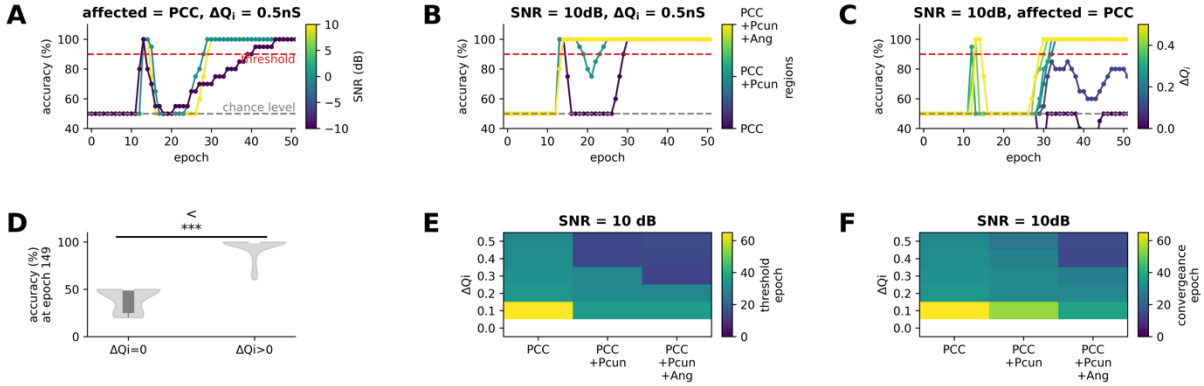

Figure S2. Training and classification performance on whole-brain TVB simulations with a human connectome. (**A-C**) Classification accuracy as a function of training epochs (**A**) for SNRs ranging from -10 dB to 10 dB when only the PCC is affected by E/I imbalance of severity ΔQi=0.5nS (**B**) for only the PCC, the PCC and PCun, and the PCC, PCun, and AnG affected by E/I imbalance of severity ΔQi=0.5nS with fixed SNR of -10 dB, and (**C**) for E/I imbalance of severity varying from ΔQi=0nS to ΔQi=0.5nS in the PCC with fixed SNR of -10. (**D**) Classification accuracy at epoch 149 for E/I imbalance severity of ΔQi=0nS and ΔQi>0nS. (**E-F**) Training epoch during which a 90% accuracy threshold is passed (**E**) and convergence epoch (**F**) as a function of regions affected and E/I imbalance severity for an SNR of 10 dB.

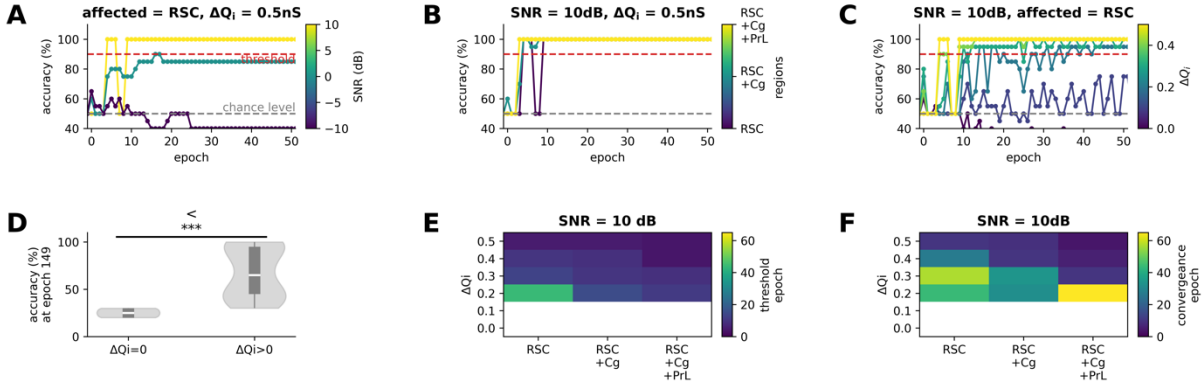

Figure S3. Training and classification performance on whole-brain TVB simulations with a mouse connectome. (**A-C**) Classification accuracy as a function of training epochs (**A**) for SNRs ranging from -10 dB to 10 dB when only the RSC is affected by E/I imbalance of severity ΔQi=0.5nS (**B**) for only the RSC, the RSC and Cg, and the RSC, Cg, and PrL affected by E/I imbalance of severity ΔQi=0.5nS with fixed SNR of -10 dB, and (**C**) for E/I imbalance of severity varying from ΔQi=0nS to ΔQi=0.5nS in the PCC with fixed SNR of -10. (**D**) Classification accuracy at epoch 149 for E/I imbalance severity of ΔQi=0nS and ΔQi>0nS. (**E-F**) Training epoch during which a 90% accuracy threshold is passed (**E**) and convergence epoch (**F**) as a function of regions affected and E/I imbalance severity for an SNR of 10 dB.

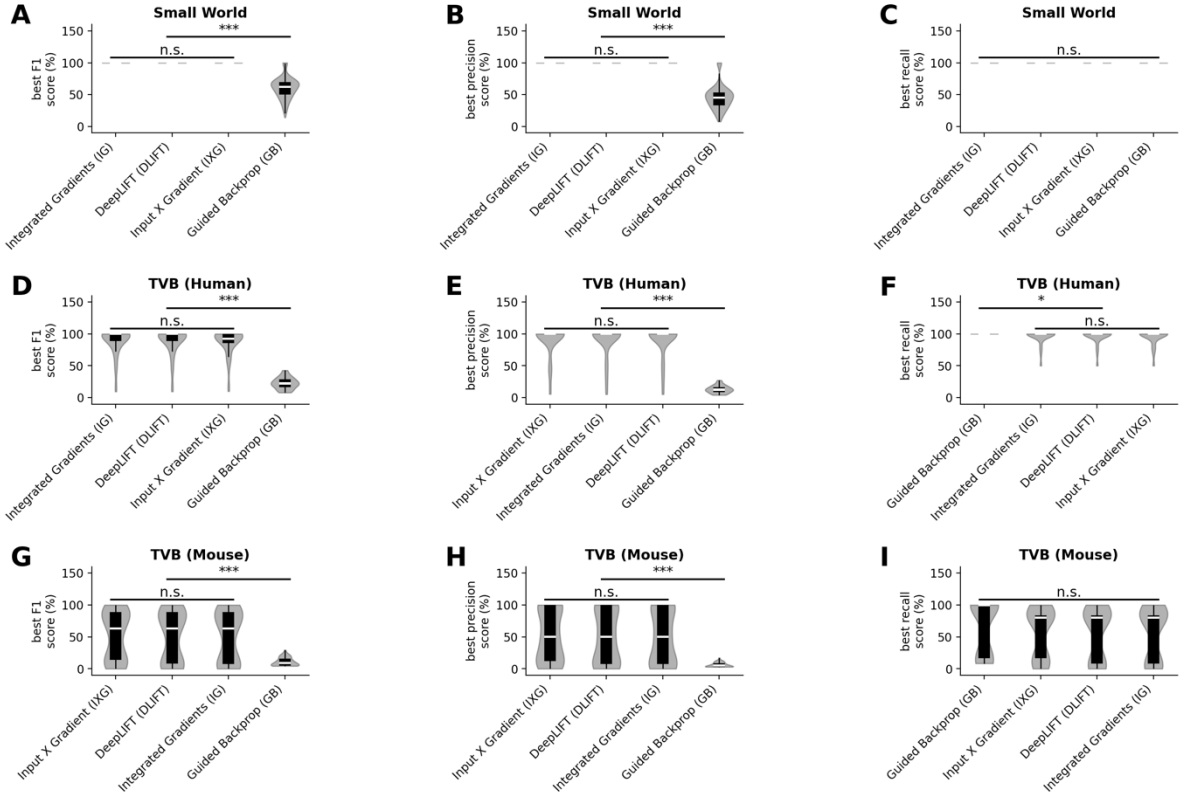

Figure S4. Performance of four different feature attribution methods in detecting affected ROIs. Classifier accuracy quantified by F1score (**A**, **D**, **G**), precision (**B**, **E**, **H**), and recall (**C**, **F**, **I**) in RNN simulations (**A-C**) as well as TVB simulations with human (**D-F**) and mouse (**G-I**) connectomes across feature attribution methods IXG, DLIFT, IG, and GB.

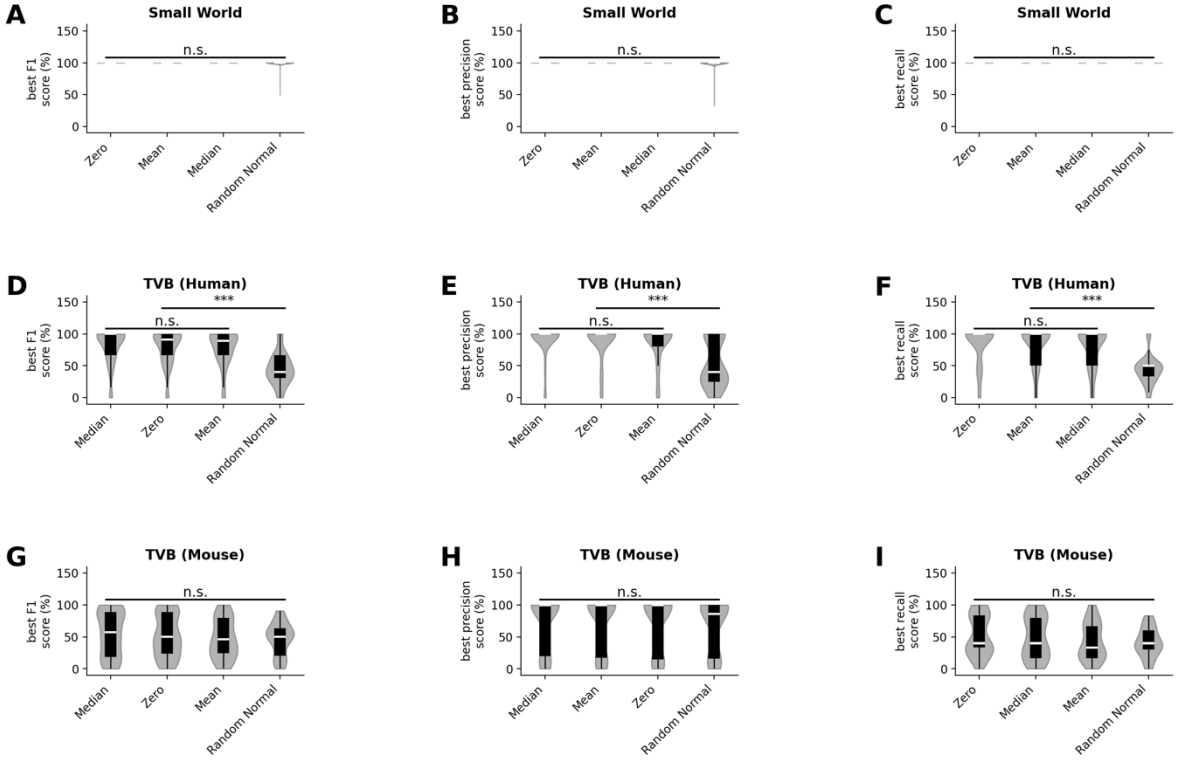

Figure S5. Performance of IG methods with four different baselines in detecting affected ROIs. Classifier accuracy quantified by F1score (**A**, **D**, **G**), precision (**B**, **E**, **H**), and recall (**C**, **F**, **I**) in RNN simulations (**A-C**) as well as TVB simulations with human (**D-F**) and mouse (**G-I**) connectomes across median, mean, zero, and random normal baselines.