Bias, Measurement Error, and Double-Dipping: When Can GNN Convolutions Help Brain Connectome Prediction?

Tommaso Castellani
Columbia University, USA
tc3527@columbia.edu

Jiaqi Li
Rice University, USA
jiaqi.li@rice.edu

Muriah D. Wheelock
Washington University in St. Louis, USA
mdwheelock@wustl.edu

Rezwana Rosen Razzaque
Washington University in St. Louis, USA
r.razzaque@wustl.edu

Helmut Laufs
Christian-Albrechts University Kiel, Germany
h.laufs@neurologie.uni-kiel.de

Hong Chen
Washington University in St. Louis, USA
chen.h1@wustl.edu

Claire Donnat
University of Chicago, USA
cdonnat@uchicago.edu

Abstract

Graph Neural Networks (GNNs) are an important tool for fMRI prediction, where empirical functional connectivity matrices serve both as features and for graph construction. However, recent studies and our own experiments show that these methods often fail to outperform simpler graph-agnostic baselines. We propose a statistical explanation for this failure by formulating connectome prediction as an errors-in-variables (EIV) problem: the true connectome is latent, while the observed covariance matrix is a noisy finite-time-series estimate. In a linear asymptotic setting, we decompose the effect of graph convolution into three mechanisms: smoothing-induced bias and variance reduction, graph estimation error, and a double-dipping effect arising when the same noisy matrix determines both graph and features. Under a latent community model adapted to neuroscience, we identify a narrow favorable regime, set by the brain community structure, the predictive signal, the graph construction and the recording length, in which message passing can improve prediction. Simulations validate this decomposition, while real-data experiments illustrate that connectome GNN pipelines often fall outside this regime. Our results provide a first step towards explaining when graph convolutions can help brain connectome prediction, and why they often do not.[1]

1 Introduction

Functional brain network analysis has emerged as an important approach both for diagnosing neurological disorders and for predicting cognitive characteristics. In this setting, the data typically consist of functional magnetic resonance imaging (fMRI) recordings, captured as time-series matrices $Z_i \in \mathbb{R}^{T\times p}$; each matrix tracks how brain activity unfolds over T time points across p predefined

[1] Code is available at https://github.com/tommicaste/B-ME-DD.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

regions of interest (ROIs) for a given patient. In this context, the primary research question typically concerns how interactions among brain regions, often measured as covariances or correlations between ROIs and often referred to as "the connectome", relate to a subject-specific outcome, represented by y_i. This framework corresponds to positing an underlying model $f^\star$ that links the connectome to the outcome: $f^\star : \Sigma^{(i)} \to y_i$.

Since brain functions are hypothesized to stem from a complex network of interconnected regions [1], graph-based methods are increasingly adopted for modeling the brain connectome. Following the success of GNNs in other applications, algorithms based on GNN message-passing schemes have become prevalent in the field [2, 3], frequently serving as the backbone of state-of-the-art pipelines with an increasing range of variants created and customized for neuroscience applications.

Despite the popularity of graph-based methods, a growing body of evidence suggests that message passing does not actually help in analyzing the human brain connectome. In Figure 1, we show that the performance of standard GNNs (GraphSAGE [4], GCN [5] and GAT [6]) and a popular brain connectomics graph methodology (Residual GNN [3]) does not improve upon simpler (graph-agnostic) methods on three real-world datasets. This echoes the findings of [7], who benchmarked graph deep learning models on large-scale functional connectome datasets and found that their message aggregation brings no gain over graph-agnostic baselines; we refer to their work for an exhaustive empirical demonstration of the problem. Similarly, [8] showed that basic classifiers using functional connectivity matrices consistently outperform message passing architectures. Finally, [9] showed that the message aggregation mechanisms characteristic of GNNs often have a detrimental effect on prediction performance from functional brain networks. From the broader GNNs landscape, the literature has focused on understanding the pitfalls of message passing in node-level tasks such as semi-supervised classification, with particular emphasis on the oversmoothing phenomenon [10–13]. However, these critiques are either predominantly empirical or in a non-applicable setting; hence our main question of interest:

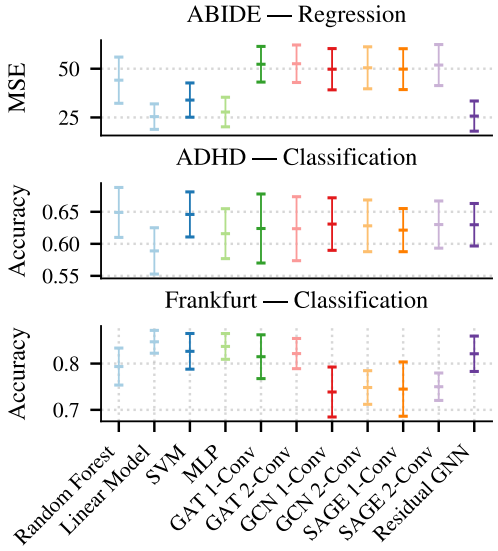


Figure 1: Performance across regression and classification tasks on 3 datasets. Graph-based methods fail to outperform graph-agnostic baselines, both linear (Linear Model) and non-linear (Random Forest, SVM, MLP). Mean $\pm$ 1 SD over 10 seeds. See Appendix D for details.

Why do graph message passing algorithms fail in brain connectome analysis?

To answer this question, we study the traditional brain connectomics pipeline (Figure 2) as a statistical estimation problem. The true connectome of each subject is latent, and we only observe an empirical connectivity matrix estimated from a finite recording, so the analysis is naturally framed as an errors-in-variables (EIV) problem. The empirical connectivity matrix, either the sample covariance $\hat{\Sigma}$ or the sample correlation $\hat{C}$, is used twice: as the node features and as the basis for the graph $\hat{A} = f(\hat{\Sigma})$ or $\hat{A} = f(\hat{C})$.

Main contributions. Our contributions are summarized as follows:

- **An EIV formulation that allows untangling different effects.** Under an errors-in-variables formulation, we show in Section 2 that the effect of graph convolution on the predictive performance of the method splits into three parts: **(i) a graph smoothing effect** that induces a bias-variance tradeoff even with a correctly specified (oracle) graph; **(ii) the graph estimation error**, which is the cost of smoothing over a wrongly estimated graph; and **(iii) a double-dipping effect** that couples the graph operator with the feature noise when both come from the same sample.
- **When convolution helps.** The balance of these three effects, and hence whether graph convolution improves on the raw features, depends on three ingredients: the **community structure** of the

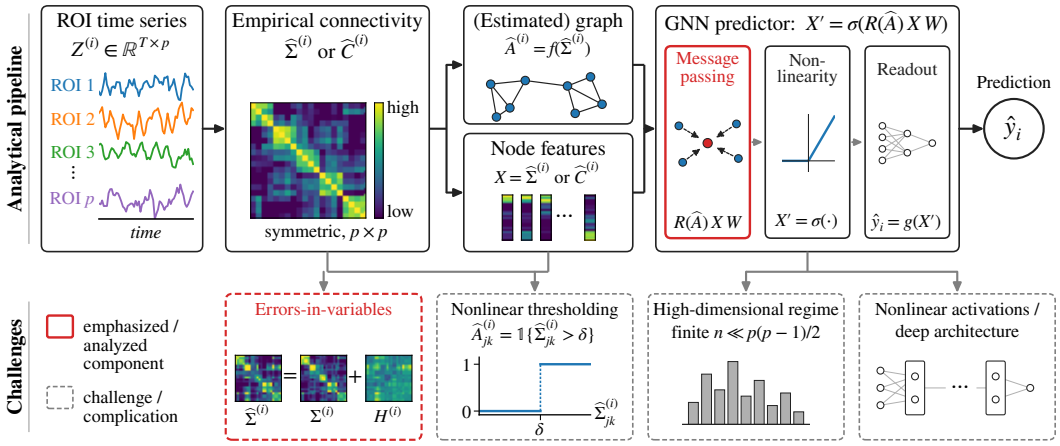


Figure 2: **Brain connectome GNN pipeline and statistical challenges.** *Top:* the standard analytical pipeline, from ROI time series to prediction $\widehat{y}_i$. *Bottom:* statistical challenges induced by the pipeline. The error-in-variables component (red) is analyzed in this paper; the remaining challenges (gray, dashed) are left to future work.

connectome, which describes how sharply ROIs group into communities; the **predictive signal**, which describes how the target aligns with these communities; and the **threshold**, which determines which connections become edges of the graph. We make all three explicit through a latent community model, building on the standard neuroscience assumption that ROIs belong to a small number of communities (Section 3). Under this model, convolution improves on the raw features only when its noise reduction outweighs the signal it distorts, which restricts the gains to short recordings. Simulations, where the true connectome is known, validate the theory (Section 4).

- **Practical relevance for real data.** On fMRI data, we build proxies for the oracle and independent graphs to separate the three effects in practice, and we show how the theory can guide the choice of graph, threshold and window length (Section 5, Appendix D.4).

2 Statistical framework

For each subject $i = 1, \dots, n$, let $Z^{(i)} \in \mathbb{R}^{T\times p}$ denote the centered fMRI time series measured at p regions of interest (ROIs) over T time points. To keep the analysis simple, we assume independent time-points $Z_t^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma^{(i)})$, $t = 1, \dots, T$, where $\Sigma^{(i)} \in \mathbb{R}^{p\times p}$ is the subject's latent functional connectome. The connectome is not observed directly. Instead, we observe the empirical covariance

$$\widehat{\Sigma}^{(i)} = \frac{1}{T}\sum_{t=1}^{T} Z_t^{(i)} Z_t^{(i)\top} = \Sigma^{(i)} + H^{(i)}, \tag{1}$$

where $H^{(i)}$ is the finite-sample estimation error. For simplicity, we focus on the sample covariance $\widehat{\Sigma}^{(i)}$ throughout the theoretical analysis; extensions to the sample correlation matrix $\widehat{C}^{(i)}$, which is the more common choice in fMRI pipelines, are analyzed computationally in Appendix C and behave analogously. We study a simple supervised regression problem in which the patient-level target y_i depends on the latent connectome through some (potentially non-linear) function $f^\star$:

$$y_i = f^\star(\Sigma^{(i)}) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{2}$$

The purpose of brain connectomics analysis is thus to find the best approximation of $f^\star$ within some function class $\mathcal{F}$, typically parametrized by a vector of parameters θ:

$$\hat{\theta} \in \operatorname{argmin}_{f_\theta \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \|y_i - f_\theta(\Sigma^{(i)})\|^2. \tag{3}$$

An errors-in-variables regression problem. Since the true connectivity $\Sigma^{(i)}$ is not observed, a simple solution is to use the estimated covariance matrix (also known as the plug-in estimator) $\widehat{\Sigma}^{(i)}$ and solve Eq. 3. However, predicting y_i from $\hat{\Sigma}^{(i)}$ rather than $\Sigma^{(i)}$ places us in an *errors-in-variables*

regression setting. For simplicity, we restrict the function class to linear maps, $\mathcal{F} = \{f_\theta : x \mapsto x^\top \theta\}$, so that $y_i = \mathrm{vec}(\Sigma^{(i)})^\top \beta^\star + \varepsilon_i$. In this setting, the plug-in OLS estimator has a closed form. Write $x_i = \mathrm{vec}(\Sigma^{(i)})$ for the true features, $\eta_i = \mathrm{vec}(H^{(i)})$ for the noise, and $\hat{x}_i = x_i + \eta_i$ for the observed features. Under the Gaussian model above, the noise η_i has conditional mean zero given $\Sigma^{(i)}$ and its conditional covariance scales as T^{-1}:

$$\mathbb{E}_{\eta_i}[\eta_i \mid \Sigma^{(i)}] = 0, \qquad \mathrm{Cov}_{\eta_i}(\eta_i \mid \Sigma^{(i)}) = \frac{1}{T}\Psi^{(i)}, \qquad \Psi^{(i)} = (I_{p^2} + K_{p,p})(\Sigma^{(i)} \otimes \Sigma^{(i)}), \quad (4)$$

where $K_{p,p}$ is the commutation matrix (Definition 1). Let $S = \mathbb{E}_\Sigma[xx^\top]$ and $\Gamma = \mathbb{E}_\Sigma[\Psi]$, and write $\dagger$ for the Moore–Penrose pseudo-inverse. A classical errors-in-variables computation (Remark 1 in Appendix B.2) gives the population OLS coefficient $\beta^*_{\mathrm{raw}}(T) = \Lambda(T)\beta^\star$, with $\Lambda(T) := (S + \frac{1}{T}\Gamma)^\dagger S$. Its test error is

$$\mathrm{MSE}_{\mathrm{raw}}(T) = \beta^{\star\top}(I - \Lambda(T))^\top S(I - \Lambda(T))\beta^\star + \frac{1}{T}\beta^{\star\top}\Lambda(T)^\top \Gamma \Lambda(T)\beta^\star + \sigma_\varepsilon^2. \quad (5)$$

The first term is a bias: the noise pulls the coefficient away from $\beta^\star$. The second is tied to the variance: the noise in the features is carried into the prediction. Both terms depend on T, not on n, so it is important to note that more subjects do not remove the error; only longer recordings do. $\mathrm{MSE}_{\mathrm{raw}}$ is the baseline we compare every graph against in the rest of the paper. A convolution is appealing because averaging neighboring entries of $\widehat{\Sigma}$ can reduce the noise term, but it can also distort the signal. The next sections quantify this trade-off.

A large literature has studied corrections for covariate measurement error, spanning corrected high-dimensional Lasso objectives [14, 15], score, likelihood, and deconvolution-based approaches for generalized and nonlinear models [16–18], and recent nonparametric or neural approaches [19]. None of these are directly applicable here: the measurement error in $\mathrm{vec}(\widehat{\Sigma})$ is not exogenous with fixed covariance but dense, entry-correlated, and dependent on the unobserved latent connectome (Eq. 4).

The appeal of the graph. An alternative is to exploit the inherent problem structure: in neuroscience, each ROI is typically assumed to participate in $K \leq p$ latent neural communities [1] that can be represented by a graph. For this reason, GNN-based algorithms have been extensively used in this field as standalone standard convolutions [5, 4, 6] or as part of brain-connectome-tailored architectures [3, 20]. Their goal is twofold: (a) obtaining a richer feature map and (b) denoising the empirical covariance matrix by averaging over appropriate neighborhoods. We postulate that the main appeal of these methods in our setting stems from the denoising effect: the usually reported strengths of GNNs as feature maps (such as learning permutation-invariant representations) do not apply here, since ROIs are aligned across patients and therefore non-exchangeable.

A graph message passing layer is defined as $X' = \sigma(R(A)\, X\, W)$, where $A \in \mathbb{R}^{p\times p}$ is an adjacency matrix, $X \in \mathbb{R}^{p\times m}$ a node feature matrix, $W \in \mathbb{R}^{m\times m'}$ learnable weights, $R(A)$ the graph operator induced by A (e.g. $\widetilde{D}^{-1}\widetilde{A}$ for GraphSAGE [4], $\widetilde{D}^{-1/2}\widetilde{A}\,\widetilde{D}^{-1/2}$ for GCN [5], with $\tilde{A} = A + I$ and $\tilde{D} = \mathrm{diag}(\tilde{A}\mathbf{1})$), and σ an activation function.

The brain connectomics pipeline (Figure 2) proceeds as follows. First, each subject's time series is split into windows of length T, so that shorter windows yield more samples per subject. Second, a connectivity matrix is estimated from each window (usually the correlation $\widehat{C}^{(i)}$ rather than the covariance $\widehat{\Sigma}^{(i)}$). Third, this matrix is used twice: as the node features, $X = \widehat{\Sigma}^{(i)}$ or $X = \widehat{C}^{(i)}$, and to build the graph $\widehat{A}$ through a non-linear function f (for instance, hard or absolute-value thresholding [3], either at a fixed level, $\widehat{A}_{jk} = \mathbb{1}\{\widehat{\Sigma}_{jk} > \delta\}$ with $\delta \in \mathbb{R}$, or at a fixed edge density, $\widehat{A}_{jk} = \mathbb{1}\{\widehat{\Sigma}_{jk} > q_{1-\tau}(\widehat{\Sigma})\}$ with $\tau \in (0,1)$, where $q_{1-\tau}$ is the $(1-\tau)$-quantile of the off-diagonal entries). Finally, one or more message-passing layers, usually with non-linear activations and followed by a readout, are trained on the n windows pooled across subjects, and n is often smaller than the number $p(p-1)/2$ of pairwise interactions between the p ROIs.

We focus here on understanding the denoising effect and the bias–variance tradeoff induced by the convolution, which requires disentangling it from the other factors summarized at the bottom of Figure 2. For the purposes of this paper, we therefore simplify the pipeline. We use the covariance and independent time points, as introduced above. We work in an infinite-n (asymptotic) regime, where T is a free parameter rather than tied to n through the windowing. We also restrict our

analysis to linear models applied to the vectorized smoothed covariance, with the graph obtained by thresholding at a fixed level δ rather than a fixed density τ. The linear restriction aligns with much of the empirical work that uses only the message-passing component and does not incorporate additional non-linearities [21–23]. Under these simplifications, the population risk has a closed form, and the analysis still captures how the message-passing layer interacts with the measurement noise. Appendix D.3.1 relaxes these simplifications one at a time on real data.

We compare prediction based on the raw feature map $\widehat{x} = \mathrm{vec}(\widehat{\Sigma}^{(i)})$ with that on the smoothed feature map $\widetilde{x}_A = \mathrm{vec}(R(A)\,\widehat{\Sigma}^{(i)})$, where $R(A)\widehat{\Sigma}^{(i)}$ is a linear message-passing convolution (e.g. $R(A) = \tilde{D}^{-1}\tilde{A}$ for GraphSAGE). We consider four choices of A:

1. A *population graph* $(A^{\mathrm{pop}})_{jk}(\delta) = \mathbb{1}\{(\Sigma_0)_{jk} > \delta\}$, where $\Sigma_0 = \mathbb{E}_\Sigma[\Sigma^{(i)}]$, is the average brain connectivity, independent of each subject's feature noise.
2. The *oracle graph* $A^{\mathrm{oracle}(i)}_{jk}(\delta) = \mathbb{1}\{\Sigma^{(i)}_{jk} > \delta\}$, constructed from the true $\Sigma^{(i)}$, which captures the "best-case" impact of smoothing under correct graph specification.
3. The *same-sample graph* $\widehat{A}^{(i)}_{jk}(\delta) = \mathbb{1}\{\widehat{\Sigma}^{(i)}_{jk} > \delta\}$, built from the same empirical covariance as the node features; the most common choice in practice [20, 24–26].
4. An *independent graph* $A^{\perp(i)}_{jk}(\delta) = \mathbb{1}\{\widehat{\Sigma}^{\perp(i)}_{jk} > \delta\}$, where $\widehat{\Sigma}^{\perp(i)} = \Sigma^{(i)} + H^{\perp(i)}$ with $H^{\perp(i)} \perp\!\!\!\perp H^{(i)} \mid \Sigma^{(i)}$. It has graph-estimation error but the errors in the graph and features are independent.

Replacing $\widehat{x}$ with the smoothed features $\widetilde{x}_A = \mathcal{R}_A\widehat{x}$, where $\mathcal{R}_A = I_p \otimes R(A)$ is the vectorized convolution, keeps us in an errors-in-variables regression, now with the convolution sitting between the noise and the coefficient. For any graph construction A, whether fixed or random, the population OLS test error is

$$\mathrm{MSE}_A(T) = \mathbb{E}_\Sigma[h(\Sigma)^2] - c_A^\top M_A^\dagger c_A + \sigma_\varepsilon^2,$$

where $h(\Sigma) = \mathrm{vec}(\Sigma)^\top\beta^\star$ is the noiseless target, $M_A = \mathbb{E}_{\Sigma,\eta,A}[\widetilde{x}_A\widetilde{x}_A^\top]$ and $c_A = \mathbb{E}_{\Sigma,\eta,A}[\widetilde{x}_A\,h(\Sigma)]$. Writing $\widehat{x} = x + \eta$, both split into a signal part, a noise part, and cross terms between the two:

$$M_A = \mathbb{E}_{\Sigma,\eta,A}[\mathcal{R}_A x x^\top \mathcal{R}_A^\top] + \mathbb{E}_{\Sigma,\eta,A}[\mathcal{R}_A x \eta^\top \mathcal{R}_A^\top] + \mathbb{E}_{\Sigma,\eta,A}[\mathcal{R}_A \eta x^\top \mathcal{R}_A^\top] + \mathbb{E}_{\Sigma,\eta,A}[\mathcal{R}_A \eta\eta^\top \mathcal{R}_A^\top],$$
$$c_A = \mathbb{E}_{\Sigma,\eta,A}[\mathcal{R}_A x x^\top]\beta^\star + \mathbb{E}_{\Sigma,\eta,A}[\mathcal{R}_A \eta x^\top]\beta^\star.$$

How these terms behave depends on how the graph relates to the feature noise, which gives three cases (Proposition 4, Appendix B.3):

(i) For the same-sample graph $\widehat{A}$, the graph is built from the same $\widehat{\Sigma}$ as the features, so it depends on the noise it is meant to average out. All terms are present, and the cross terms do not vanish due to this double-dipping effect.

(ii) For an independent graph that varies across subjects, such as $A^{\mathrm{oracle}}$and $A^\perp$, the graph is independent of the feature noise given Σ, so the cross terms vanish.

(iii) For a fixed graph shared across samples, as for A^{pop}, the same convolution is applied to every subject, so $\mathcal{R}_A$ is a fixed linear map. A linear model on $\widetilde{x}_A$ is then still a linear model on $\widehat{x}$, and message passing can never improve on the raw features. It only ties with the raw features when $R(A)$ is invertible and can only do worse otherwise (Proposition 5, Appendix B.3).

In the two cases where the graph varies across subjects, the sign of each effect is not determined in general. It depends on how the connectome Σ is structured, on how the target relates to that structure, and on how the graph is built. In the next section we introduce a latent community model that makes these three ingredients explicit.

3 Brain connectome inspired latent communities model

We now specialize the previous results to a simple connectome model, showing how the favorable regime changes as we move from the oracle graph A^{oracle}, to the independent graph $A^\perp$, and to the same-sample graph $\widehat{A}$, with the goal of isolating respectively the **graph smoothing effect**, the **graph estimation error**, and the **double-dipping effect**.

The model represents brain activity through a small number of interacting latent components [27–30]. Let $F^{(i)} \in \mathbb{R}^{p\times d}$ collect the ROIs' loadings onto d latent dimensions. Each ROI j belongs to one community $z_j \in \{1,\dots,K\}$, $K \le \min\{p,d\}$; the communities are balanced, with $|C_k| = q =$

Table 1: Graph constructions and prediction conclusions. Comparisons concern unregularized population linear prediction under the stated assumptions.

Graph	Effects isolated	Main prediction conclusion
A^{pop} (population)	Fixed smoothing	Cannot improve optimized raw prediction.
A^{oracle} (oracle)	Subject-specific smoothing	Can help when variance reduction outweighs squared bias and raw refitting.
$A^{\perp}$ (independent)	+ Graph estimation	Same tradeoff, with an additional graph-error bound.
$\widehat{A}$ (same-sample)	+ Noise reuse	Close to independent prediction when graph estimation is reliable and regression is stable.

p/K and $C(j) := C_{z_j}$. Following Appendix B.4.1, we take fixed, orthogonal centroids μ_k with $\|\mu_k\|^2 = \phi^2 d$. For each subject, let $f_j^{(i)} = \mu_{z_j} + \nu_j^{(i)}$, where $\nu_j^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$ across ROIs j and subjects i, and define $\Sigma^{(i)} = F^{(i)} F^{(i)\top} + \psi^2 I_p$, with $f_j^{(i)\top}$ the j-th row of $F^{(i)}$ and the perturbations are independent across ROIs and subjects. The ROI partition, centroids, latent dimension d, and parameters $\phi, \psi > 0$ are fixed across subjects. Under this model, we characterize three ingredients.

Community structure, σ. The parameter σ controls subject variability around the shared community pattern. For $j \neq k$, the within- and between-community covariance means and common variance are $\mu_W = \phi^2 d$, $\mu_B = 0$, and $s^2(\sigma) = 2\sigma^2\phi^2 d + \sigma^4 d$, respectively (Lemma 3). Increasing σ increases the spread around these fixed levels as shown in the heatmaps in Figure 3.

Predictive signal, $B^\star$. The symmetric coefficient $B^\star \in \mathbb{R}^{p\times p}$ specifies $h(\Sigma) = \langle B^\star, \Sigma\rangle$, with $\beta^\star = \mathrm{vec}(B^\star)$. We characterize its alignment with the same communities through $\max_j \max_{\ell\in C(j)} \sum_{k=1}^{p} (B^\star_{jk} - B^\star_{\ell k})^2 \le \kappa_\beta^2$. Small κ_β describes community-smooth predictive profiles; the ideal case $\kappa_\beta = 0$ gives coefficients depending only on (z_j, z_k). Community-level, single-edge, and contrast targets are defined in Appendix B.4.3.

Threshold, δ. The threshold selects averaging neighborhoods from these connectomes. For the oracle, $A^{\mathrm{oracle}(i)}_{jk}(\delta) = \mathbb{1}\{\Sigma^{(i)}_{jk} > \delta\}$ for $j \neq k$, with mean aggregation $R(A) = \widetilde{D}^{-1}(A + I_p)$. At the boundaries, $\delta = +\infty$ gives the empty graph and $R = I_p$, the raw features, while $\delta = -\infty$ gives the complete graph and $R = \mathbf{1}\mathbf{1}^\top/p$, the global average. The interesting case is an intermediate threshold, $\mu_B < \delta < \mu_W$, at distance $\gamma_\delta := \min\{\mu_W - \delta, \delta - \mu_B\}$ from the two means. The graph then mostly links ROIs of the same community, and the operator splits as

$$R(A^{\mathrm{oracle}}) = P_C + \Delta R_\delta, \qquad (P_C)_{j\ell} = q^{-1}\mathbb{1}\{\ell \in C(j)\}.$$

Here P_C averages each ROI over its community, and ΔR_δ collects the subject-specific deviations, which depend on σ and δ.

The following results examine how these three ingredients jointly affect prediction as the measurement length T varies.

Result 1: Oracle graph smoothing effect. *Smoothing over the oracle graph reduces measurement noise but also distorts the signal, and the balance between the two depends on the predictive signal $B^\star$, the community structure σ and the threshold δ. If smoothing introduces prediction error that persists even after refitting the linear predictor on noiseless features, it cannot outperform raw features for sufficiently long recordings.*

The following proposition is a more formal version of Result 1.

Proposition 1 (Oracle smoothing)**.** *Under the EIV assumptions of Remark 1 (Appendix B.2), fix any threshold δ; the oracle operator $R\left(A^{oracle}(\delta)\right)$ is a function of Σ. Assume all displayed moments are finite and define the squared bias and the variance reduction:*

$$\mathcal{B}_{A^{oracle}}(\delta) := \mathbb{E}_\Sigma\left[\langle B^\star, R\left(A^{oracle}(\delta)\right)\Sigma - \Sigma\rangle^2\right],$$
$$\Delta\mathcal{V}_{A^{oracle}}(\delta) := T\,\mathbb{E}_\Sigma\left[\mathrm{Var}_\eta\{\langle B^\star, H\rangle \mid \Sigma\} - \mathrm{Var}_\eta\{\langle B^\star, R\left(A^{oracle}(\delta)\right)H\rangle \mid \Sigma\}\right].$$

Sufficient condition. *The population MSEs satisfy, for every $T > 0$,*

$$\mathrm{MSE}_{\mathrm{oracle}}(T,\delta) - \mathrm{MSE}_{\mathrm{raw}}(T) \le \mathcal{B}_{A^{oracle}}(\delta) - \frac{\Delta\mathcal{V}_{A^{oracle}}(\delta)}{T} + r_{\mathrm{raw}}(T),$$

where

$$r_{\mathrm{raw}}(T) := T^{-2}\beta^{\star\top}\Gamma(S + T^{-1}\Gamma)^{\dagger}\Gamma\beta^{\star} \geq 0.$$

If the support of Σ spans the symmetric matrices (Remark 1), as in the latent community model with $\sigma > 0$, then $r_{\mathrm{raw}}(T) = O(T^{-2})$. Thus oracle smoothing improves whenever $\mathcal{B}_{A^{oracle}}(\delta) + r_{\mathrm{raw}}(T) < \Delta\mathcal{V}_{A^{oracle}}(\delta)/T$.

Necessary condition. *Let $b_\delta := \min_{\gamma\in\mathbb{R}^{p^2}} \mathbb{E}_\Sigma\left[\left\{h(\Sigma) - \gamma^\top \mathrm{vec}(R\left(A^{oracle}(\delta)\right)\Sigma)\right\}^2\right]$ be the minimum noiseless prediction error, with γ shared across subjects. If $b_\delta > 0$, then*

$$\mathrm{MSE}_{\mathrm{oracle}}(T,\delta) < \mathrm{MSE}_{\mathrm{raw}}(T) \quad \Longrightarrow \quad T < \frac{\beta^{\star\top}\Gamma\beta^{\star}}{b_\delta}.$$

The sufficient condition guarantees improvement when the variance reduction exceeds the squared bias plus $r_{\mathrm{raw}}(T)$, which accounts for refitting the raw regression. If a positive error b_δ remains even after fitting on noiseless smoothed features, sufficiently long recordings make raw prediction more accurate, giving the necessary upper limit on T.

For the intermediate threshold, $\mu_B < \delta < \mu_W$, the decomposition introduced above gives

$$\mathcal{B}_{A^{\mathrm{oracle}}}(\delta) = \mathbb{E}_\Sigma\left[\left\{\langle B^\star, (P_C - I_p)\Sigma\rangle + \langle B^\star, \Delta R_\delta \Sigma\rangle\right\}^2\right].$$

Exact community smoothness, $P_C B^\star = B^\star$, removes the first term, but not necessarily the second. Thus $B^\star$ determines which information averaging must preserve, while σ and δ jointly determine the subject-specific neighborhoods and their noise propagation. We also note that strong concentration relative to the threshold margins makes the oracle nearly fixed, unable to improve optimized prediction (Proposition 5). Clear communities and an aligned target therefore do not alone guarantee a gain. Concrete examples for community-strength and single-edge targets are given in Appendix B.5.1.

We now ask whether a graph estimated from an independent sample of the same subject keeps the gain of the oracle.

Result 2: Graph estimation error. *Replacing the oracle with a graph estimated from an independent sample adds a graph-estimation error bound $\mathcal{E}(T,\delta)$ that vanishes with the probability of an edge error. The independent graph still improves on the raw features when this bound is smaller than the gain of the oracle. Recovering the community partition shared by all subjects is not enough, because the gain comes from the subject-specific neighborhoods.*

The independent graph thresholds a second estimate of the same connectome, $A^{\perp}_{jk}(\delta) = \mathbb{1}\{\widehat{\Sigma}^{\perp}_{jk} > \delta\}$ with $\widehat{\Sigma}^{\perp} = \Sigma + H^{\perp}$ and $H^{\perp} \perp\!\!\!\perp H \mid \Sigma$. Its edges can be wrong, which adds graph-estimation error, but its noise is independent of the feature noise, so there is no double dipping.

Proposition 2 (Graph-estimation error for the independent graph)**.** *Under the assumptions of Proposition 1, let $A^{\perp}$ be built from a conditionally independent copy $H^{\perp}$ of H given Σ, and assume $\mathbb{E}_\Sigma\|\Sigma\|_F^4 < \infty$. Define the edge-error probability and the graph-estimation error:*

$$\rho(T,\delta) := \mathbb{P}(A^{\perp} \neq A^{oracle}), \qquad \mathcal{E}(T,\delta) := C(1 + T^{-1})\sqrt{\rho(T,\delta)},$$

where C depends only on p, $\|B^\star\|_F$ and $\mathbb{E}_\Sigma\|\Sigma\|_F^4$, not on T or δ. Then, for every $T > 0$,

$$\mathrm{MSE}_{\mathrm{ind}}(T,\delta) - \mathrm{MSE}_{\mathrm{raw}}(T) \leq \mathcal{B}_{A^{oracle}}(\delta) - \frac{\Delta\mathcal{V}_{A^{oracle}}(\delta)}{T} + \mathcal{E}(T,\delta) + r_{\mathrm{raw}}(T).$$

Thus the independent graph improves whenever $\mathcal{B}_{A^{oracle}}(\delta) + \mathcal{E}(T,\delta) + r_{\mathrm{raw}}(T) < \Delta\mathcal{V}_{A^{oracle}}(\delta)/T$.

For $\mu_B < \delta < \mu_W$, graph estimation must recover the subject-specific neighborhoods $P_C + \Delta R_\delta$. The target $B^\star$ determines how much a wrong edge costs, while δ and σ determine how often one occurs. An edge is wrong when the noise in $\widehat{\Sigma}_{jk}$, of size $\sqrt{v/T}$, exceeds the margin γ_δ between δ and the nearest community mean, where v bounds the scaled noise variance $T\,\mathrm{Var}_\eta(H_{jk} \mid \Sigma) = \Sigma_{jj}\Sigma_{kk} + \Sigma_{jk}^2$. Hence, when communities are well separated, ρ decays exponentially in $T/T_{\mathrm{graph}}(\delta)$, with $T_{\mathrm{graph}}(\delta) = v/\gamma_\delta^2$ (Appendix B.5.2). Graph estimation is therefore reliable only once $T \gg T_{\mathrm{graph}}(\delta)$. Edge errors of this kind are visible in data: thresholded connectivity graphs built from short recordings contain spurious between-network edges. Longer recordings improve

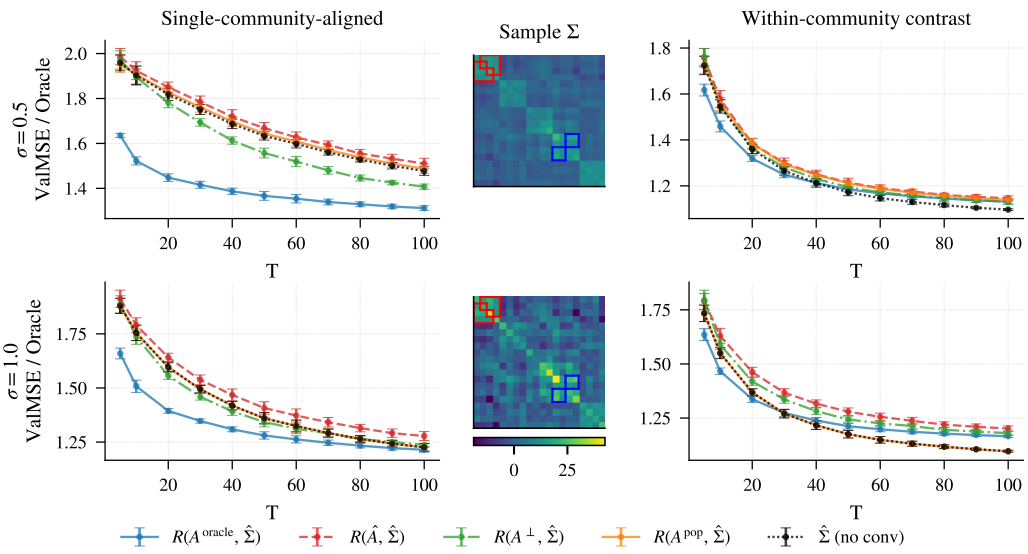


Figure 3: **Decoupling the effects for different values of β and latent structures.** Center: true covariance Σ for two graph regimes (σ); overlays mark the support of two coefficient patterns (β). Left/right: normalized ValMSE vs. T for $\mathrm{SageMean}(A, \widehat{\Sigma})$ and a no-conv baseline. Mean $\pm$ 1 SD over 5 seeds. $p = 20$, $K = 5$, $n = 25{,}000$; thresholds $\delta_{\sigma=0.5} \approx 9.10$, $\delta_{\sigma=1.0} \approx 10.20$.

edge estimation but also reduce the raw-feature noise that smoothing aims to remove, so the gain is confined to an intermediate range of T.

Result 3: Double-dipping effect. *Using the same sample for graph and features adds a signed coupling term $\mathcal{D}_{\mathrm{dd}}(T, \delta)$ to the constraints of Result 2. When measurement noise rarely changes the oracle graph and the fitted regression coefficients remain bounded, the same-sample and independent graphs have similar prediction errors. The sign of $\mathcal{D}_{\mathrm{dd}}$ depends on the target.*

The same-sample graph $\widehat{A}_{jk}(\delta) = \mathbb{1}\{\widehat{\Sigma}_{jk} > \delta\}$ uses the same noise H for edge selection and feature construction. We treat double dipping as a signed contribution because reusing the sample can either improve or worsen prediction compared with using an independent graph.

Proposition 3 (Double-dipping contribution for the same-sample graph)**.** *Under the assumptions of Proposition 2, define the double-dipping contribution $\mathcal{D}_{\mathrm{dd}}(T, \delta) := \mathrm{MSE}_{\mathrm{same}}(T, \delta) - \mathrm{MSE}_{\mathrm{ind}}(T, \delta)$. Then, for every $T > 0$,*

$$\mathrm{MSE}_{\mathrm{same}}(T, \delta) - \mathrm{MSE}_{\mathrm{raw}}(T) \le \mathcal{B}_{A^{\mathrm{oracle}}}(\delta) - \frac{\Delta\mathcal{V}_{A^{\mathrm{oracle}}}(\delta)}{T} + \mathcal{E}(T, \delta) + \mathcal{D}_{\mathrm{dd}}(T, \delta) + r_{\mathrm{raw}}(T).$$

Thus the same-sample graph improves whenever $\mathcal{B}_{A^{\mathrm{oracle}}}(\delta) + \mathcal{E}(T, \delta) + \mathcal{D}_{\mathrm{dd}}(T, \delta) + r_{\mathrm{raw}}(T) < \Delta\mathcal{V}_{A^{\mathrm{oracle}}}(\delta)/T$. The contribution $\mathcal{D}_{\mathrm{dd}}$ is signed: negative values favor the same-sample graph.

The two graphs have the same conditional law given Σ (Lemma 2), so $\mathcal{D}_{\mathrm{dd}}$ isolates noise reuse, not graph quality. For intermediate thresholds both estimate $P_C + \Delta R_\delta$, and genuine subject-specific variation is not double dipping. Under additional moment and spectral assumptions (Appendix B.5.3), $|\mathcal{D}_{\mathrm{dd}}| \le C_{\mathrm{dd}}\sqrt{\rho/T}$. Double dipping therefore vanishes at $\delta = \pm\infty$ and is small whenever edge errors are rare, for instance once $T \gg T_{\mathrm{graph}}(\delta)$ with well-separated communities. In the intermediate regime an independent graph and same-sample graph have similar behavior.

4 Simulations

We validate the theoretical predictions of Section 3 under controlled conditions where the ground-truth covariance $\Sigma^{(i)}$ is known. We generate synthetic data from the latent community model (Appendix B.4) to study how the five feature maps from Section 2 behave as we vary the community structure, the predictive signal, and the threshold. As in the theoretical proofs, we use balanced communities with fixed orthogonal centroids and fixed latent dimension, so that the main latent-structure knob is the subject-level variability σ: small σ gives strongly clustered connectomes, while

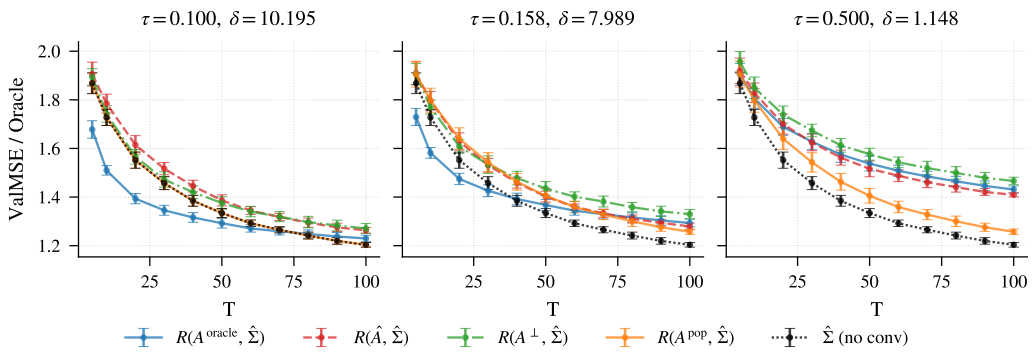


Figure 4: **Decoupling the effect of the threshold δ.** Each panel fixes a different threshold δ on Σ and plots normalized ValMSE vs. T for $\mathrm{SageMean}(A, \widehat{\Sigma})$ and the no-conv baseline. Mean $\pm$ 1 SD over 5 seeds. $\beta =$ `community_aligned`, $\sigma = 1.0$, $p = 20$, $K = 5$, $n = 25{,}000$.

large σ weakens the community structure. For each subject, we sample a time series of length T, compute $\widehat{\Sigma}^{(i)}$ (deferring results for $\widehat{C}^{(i)}$ to Appendix C), and evaluate the various feature maps at threshold level δ: raw, oracle graph, independent graph, same-sample graph, and population graph. Each is fit by OLS with an intercept on standardized features (80/20 train/test split). The dependent variable is generated using regression coefficients that span a range of favorability to graph smoothing.

Results. We disentangle the three effects of Section 3 as gaps between the feature maps $\{\widehat{\Sigma},\ R(A^{\mathrm{oracle}})\widehat{\Sigma},\ R(A^{\perp})\widehat{\Sigma},\ R(\widehat{A})\widehat{\Sigma},\ R(A^{\mathrm{pop}})\widehat{\Sigma}\}$, normalized by the oracle MSE.

We start by isolating the graph smoothing effect. The population graph A^{pop} tracks $\widehat{\Sigma}$ closely in Figure 3, consistent with Proposition 5, so the relevant comparison is A^{oracle} vs. raw. Proposition 1 predicts a bias-variance tradeoff whose benefit is confined to small T whenever the noiseless smoothed features leave residual error, and Figure 3 confirms this: for the community-smooth target (red overlay), $\mathcal{B}_{A^{\mathrm{oracle}}}(\delta)$ is small and $R(A^{\mathrm{oracle}})\widehat{\Sigma}$ improves prediction across all T; for the within-community contrast (blue overlay), smoothing averages together the ROIs whose difference is predictive, so the oracle helps only at small T. Increasing σ from 0.5 to 1.0 spreads the entries around the community levels and narrows the oracle gain. The threshold δ modulates this gain: at sparse, high-margin thresholds the oracle preserves the community structure and the smoothing benefit is maximal, while as δ moves toward the bulk the graph densifies and the oracle gain vanishes (Figure 4, right).

We next turn to the graph-estimation error, comparing $A^{\perp}$ to A^{oracle}. When the oracle does not help, $A^{\perp}$ tracks A^{oracle} trivially. When it does (top-left panel, Figure 3), a large fraction of the oracle gain is lost to mis-estimated edges, matching Proposition 2. Figure 4 traces this dependence in δ: at a sparse, high-margin threshold (left), A^{oracle} delivers a large gain and $A^{\perp}$ recovers part of it; as δ moves toward the bulk (right), the graph approaches complete and the oracle ceases to help (Proposition 5).

Finally, the double-dipping effect is captured by the gap between $\widehat{A}$ and $A^{\perp}$. Across Figure 4, $R(\widehat{A})\widehat{\Sigma}$ tracks $R(A^{\perp})\widehat{\Sigma}$ closely, consistent with the high-margin regime of Proposition 3 where few entries of Σ lie within one Wishart noise scale of δ. Pipelines with thresholds closer to the bulk, or weaker community separation, may incur a larger signed coupling penalty.

Extensions and additional results, including application to correlation matrices as well as tests of robustness under departures from the i.i.d. sampling assumption and at finite sample sizes, are presented in Appendix C.

5 Real-data controlled experiments

We now show that the theoretical findings, already validated in the simulations, extend to real data. The traditional brain connectomics pipeline [3, 2] is employed. From each subject's ROI time series we extract rolling windows of length T and compute the per-window sample correlation matrix $\widehat{C}^{(i)}$, which serves both as the node feature matrix and as the basis for graph construction; the correlation rather than the covariance matrix is used to stay close to practice. Thresholding it yields the four

Table 2: **Decoupling the effects on real data.** Subject-level test accuracy for autism diagnosis on ABIDE (Harvard-Oxford, $p = 106$) with $\mathrm{SageMean}(A, \widehat{C})$, one convolution, identity activation, n fixed across T, $\tau = 0.10$. Mean $\pm$ 1 SD over 20 seeds; paired t-test over seeds against $\widehat{C}$ (left) or zero (right): $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The majority-class predictor has accuracy 0.50.

T	$\widehat{C}$	$A^{\mathrm{oracle}}_{\mathrm{est}}$	$A^{\perp}_{\mathrm{est}}$	$\widehat{A}$	A^{pop}	$A^{\mathrm{oracle}}_{\mathrm{est}} - \widehat{C}$	$A^{\perp}_{\mathrm{est}} - A^{\mathrm{oracle}}_{\mathrm{est}}$	$\widehat{A} - A^{\perp}_{\mathrm{est}}$	$A^{\mathrm{pop}} - \widehat{C}$
5	$0.560_{\pm.033}$	$0.582^{***}_{\pm.034}$	$0.561_{\pm.031}$	$0.563_{\pm.035}$	$0.562_{\pm.039}$	$+0.022^{***}$	-0.021^{***}	$+0.002$	$+0.001$
15	$0.625_{\pm.040}$	$0.630_{\pm.033}$	$0.619_{\pm.040}$	$0.619_{\pm.039}$	$0.618_{\pm.032}$	$+0.004$	-0.010^{*}	$+0.000$	-0.007
25	$0.651_{\pm.037}$	$0.668^{**}_{\pm.030}$	$0.641_{\pm.047}$	$0.650_{\pm.044}$	$0.653_{\pm.037}$	$+0.017^{**}$	-0.027^{***}	$+0.009^{*}$	$+0.002$
40	$0.682_{\pm.034}$	$0.678_{\pm.031}$	$0.669^{**}_{\pm.030}$	$0.663^{**}_{\pm.036}$	$0.672^{*}_{\pm.029}$	-0.004	-0.009^{**}	-0.006	-0.010^{*}

graphs of Section 2: $\widehat{A}$ from the same window (the standard pipeline), $A^{\perp}_{\mathrm{est}}$ from an independent window of the same subject, $A^{\mathrm{oracle}}_{\mathrm{est}}$ from the full-length series (a low-noise proxy for the unobserved oracle), and the population graph A^{pop}. Each graph is paired with the same minimal architecture: a single convolution followed by an activation (identity or ReLU) and a readout (mean or vectorization).

Real data depart from the assumptions of the theory and the simulations. Table 2 retains three of them, which Appendix D relaxes one at a time. First, the theory treats $R(A)$ as linear, whereas practice applies a non-linearity; we compare identity and ReLU activations. Second, n is finite and, with rolling windows, grows as T shrinks; we hold n fixed across T, isolating the window length, or let it vary as in practice. Third, the theory regresses on vectorized features, whereas practice pools node embeddings; we compare both readouts. In every experiment the oracle and independent graphs are estimated and training is regularized.

Table 2 reproduces the simulation ordering on real data. The proxy oracle $A^{\mathrm{oracle}}_{\mathrm{est}}$ improves on the raw features at $T = 5$ and $T = 25$ but not at $T = 40$, so the favorable window closes at finite T (Proposition 1); the population graph A^{pop}, shared by all subjects, never helps. Graph-estimation error is significant at every T and cancels the oracle gain (Proposition 2), whereas the double-dipping term is small, of no consistent sign and comparable to the estimation loss only at $T = 40$ (Proposition 3), so the same-sample graph never beats the raw features.

Table 5 in Appendix D relaxes these three assumptions cumulatively, and Tables 6 to 9 repeat the analysis with GAT, two other thresholds and two more datasets. The ReLU weakens the oracle gain without removing it, while freeing n removes it from $T = 15$ on and strengthens it at $T = 5$, where short windows yield the most samples. Mean pooling then removes the gain left at $T = 5$, so in the deployed pipeline no graph improves significantly on the raw features at any T, as in Figure 1.

6 Conclusions

Contributions. We framed message passing on brain connectomes as an errors-in-variables regression and split its effect on prediction into three parts: graph smoothing, graph estimation error and double dipping. Under a latent community model, we showed how each part depends on the community structure, the predictive signal and the threshold. Even with the oracle graph, convolution helps only when its noise reduction outweighs the signal it distorts, which confines the gain to short recordings, and a graph shared by all subjects can never help. Simulations with the oracle, independent and same-sample graphs isolate each effect and confirm these predictions. On real data, proxies for the oracle and independent graphs reproduce the same picture: the proxy oracle improves on the raw correlation matrix at short windows and the gain disappears as T grows, the population graph never helps, most of the oracle gain is lost to graph-estimation error, and double dipping stays small. This answers the question we started from: the standard pipeline fails to beat graph-agnostic methods because the graph it can estimate from a short window is too noisy to deliver the smoothing gain that the true graph would offer.

Limitations: from simulations to real data. Our theory is stated for a linear model in the population scenario ($n \to \infty$), the most favorable case for graph convolutions. Like the simulations, it relies on strict assumptions chosen to keep the mathematics tractable. ReLU and attention fall outside the linear theory but leave the real-data picture unchanged; a theory for them, for finite samples and for richer connectomes is left to future work. For the same reason, we do not provide an explicit rule for computing the minimum threshold on T. Finally, alternative methods for estimating the graph, beyond simple thresholding, are a relevant direction for future work.

Acknowledgments

Part of this work was done while T.C. and J.L. were at the Department of Statistics, University of Chicago. C.D. would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-2238616). J.L. would like to thank the support for this project from the National Science Foundation (NSF grant DMS-2515926). This work was completed in part with resources provided by the University of Chicago's Research Computing Center.

References

[1] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. doi: 10.1038/nrn2575.

[2] Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. BrainGB: A benchmark for brain network analysis with graph neural networks. *IEEE Transactions on Medical Imaging*, 42(2):493–506, February 2023. ISSN 1558-254X. doi: 10.1109/tmi.2022.3218745. URL `http://dx.doi.org/10.1109/TMI.2022.3218745`.

[3] Anwar Said, Roza G. Bayrak, Tyler Derr, Mudassir Shabbir, Daniel Moyer, Catie Chang, and Xenofon Koutsoukos. NeuroGraph: Benchmarks for graph machine learning in brain connectomics. In *Advances in Neural Information Processing Systems*, volume 36, pages 6509–6531, 2023.

[4] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[5] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[6] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[7] Keqi Han, Yao Su, Lifang He, Liang Zhan, Sergey Plis, Vince Calhoun, and Carl Yang. Rethinking functional brain connectome analysis: Do graph deep learning models help? *npj Artificial Intelligence*, 2(1):19, 2026.

[8] Liang Yang, Yuwei Liu, Jiaming Zhuo, Di Jin, Chuan Wang, Zhen Wang, and Xiaochun Cao. Do we really need message passing in brain network modeling? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 70873–70887. PMLR, 2025. URL `https://openreview.net/forum?id=KRosBwvhDx`.

[9] Jose Lara-Rangel and Clare Heinbaugh. On the limits of applying graph transformers for brain connectome classification. In *Proceedings on "I Can't Believe It's Not Better: Challenges in Applied Deep Learning" at ICLR 2025 Workshops*, volume 296 of *Proceedings of Machine Learning Research*, pages 47–55. PMLR, 2025.

[10] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3538–3545, 2018.

[11] Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over)smoothing. In *Advances in Neural Information Processing Systems*, volume 35, pages 2268–2281, 2022.

[12] Juntong Chen, Johannes Schmidt-Hieber, Claire Donnat, and Olga Klopp. Understanding the effect of GCN convolutions in regression tasks. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4573–4581. PMLR, 2025.

[13] T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks, 2023. URL `https://arxiv.org/abs/2303.10993`.

[14] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, volume 24, pages 2726–2734, 2011.

[15] Abhirup Datta and Hui Zou. CoCoLasso for high-dimensional error-in-variables regression. *Annals of Statistics*, 45(6):2400–2426, 2017.

[16] Leonard A Stefanski and Raymond J Carroll. Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74(4):703–716, 1987.

[17] Xiaohong Chen, Han Hong, and Denis Nekipelov. Nonlinear models of measurement errors. *Journal of Economic Literature*, 49(4):901–937, 2011.

[18] Susanne M Schennach. Measurement error in nonlinear models–a review. In Daron Acemoglu, Manuel Arellano, and Eddie Dekel, editors, *Advances in Economics and Econometrics: Tenth World Congress*, volume 3, pages 296–337. Cambridge University Press, 2013.

[19] Zhirui Hu, Zheng Tracy Ke, and Jun S Liu. Measurement error models: from nonparametric methods to deep neural networks. *Statistical Science*, 37(4):473–493, 2022.

[20] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. BrainGNN: Interpretable brain graph neural network for fMRI analysis. *Medical image analysis*, 74:102233, 2021.

[21] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6861–6871, 2019.

[22] Hoang NT and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*, 2019.

[23] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 639–648, 2020.

[24] Falih Gozi Febrinanto, Mujie Liu, and Feng Xia. Balanced graph structure information for brain disease detection. In *Pacific Rim Knowledge Acquisition Workshop*, pages 134–143. Springer, 2023.

[25] Jiaxing Xu, Kai He, Yue Tang, Wei Li, Mengcheng Lan, Xia Dong, Yiping Ke, and Mengling Feng. BrainPrompt: Multi-level brain prompt enhancement for neurological condition identification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 172–182. Springer, 2025.

[26] Jiaxing Xu, Yongqiang Chen, Xia Dong, Mengcheng Lan, Tiancheng Huang, Qingtian Bian, James Cheng, and Yiping Ke. BrainOOD: Out-of-distribution generalizable brain network analysis. In *International Conference on Learning Representations*, 2025.

[27] Vince D. Calhoun, Tülay Adali, Godfrey D. Pearlson, and James J. Pekar. A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14(3):140–151, 2001. doi: 10.1002/hbm.1048.

[28] Christian F. Beckmann and Stephen M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2): 137–152, 2004. doi: 10.1109/TMI.2003.822821.

[29] Christian F. Beckmann, Mercedes DeLuca, Joseph T. Devlin, and Stephen M. Smith. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1001–1013, 2005. doi: 10.1098/rstb.2005.1634.

[30] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[31] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018. URL `https://arxiv.org/abs/1806.01261`.

[32] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1/2):32–52, 1928. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2331939.

[33] Jan R. Magnus and H. Neudecker. The Commutation Matrix: Some Properties and Applications. *The Annals of Statistics*, 7(2):381 – 394, 1979. doi: 10.1214/aos/1176344621. URL https://doi.org/10.1214/aos/1176344621.

[34] K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. H. Duncan, F. Morales, and A. C. Evans. A general statistical analysis for fMRI data. *NeuroImage*, 15(1):1–15, 2002. doi: 10.1006/nimg.2001.0933.

[35] The ADHD-200 Consortium. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6:62, 2012. doi: 10.3389/fnsys.2012.00062.

[36] Pierre Bellec, Carlton Chu, François Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The Neuro Bureau ADHD-200 preprocessed repository. *NeuroImage*, 144:275–286, 2017. doi: 10.1016/j.neuroimage.2016.06.034.

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

[38] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6): 659–667, 2014. doi: 10.1038/mp.2013.78. URL https://doi.org/10.1038/mp.2013.78.

[39] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al. The Neuro Bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. In *Frontiers in Neuroinformatics Conference Abstract: Neuroinformatics 2013*. Frontiers Media SA, 2013. doi: 10.3389/conf.fninf.2013.09.00041. URL https://www.frontiersin.org/10.3389/conf.fninf.2013.09.00041/event_abstract.

[40] Cameron Craddock, Sharad Sikka, Brian Cheung, Ranjeet Khanuja, Satrajit S Ghosh, Chaogan Yan, Qingyang Li, V Lurko, Joshua Vogelstein, Randal Burns, Stanley Colcombe, Maarten Mennes, Clare Kelly, Adriana Di Martino, Francisco Xavier Castellanos, and Michael Milham. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). In *Frontiers in Neuroinformatics Conference Abstract: Neuroinformatics 2013*. Frontiers Media SA, 2013. doi: 10.3389/conf.fninf.2013.09.00042. URL https://www.frontiersin.org/10.3389/conf.fninf.2013.09.00042/event_abstract.

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

[42] E. Tagliazucchi, F. von Wegner, A. Morzelewski, S. Borisov, K. Jahnke, and H. Laufs. Automatic sleep staging using fMRI functional connectivity data. *NeuroImage*, 63(1):63–72, 2012. doi: 10.1016/j.neuroimage.2012.06.036.

[43] Evan M. Gordon, Timothy O. Laumann, Babatunde Adeyemo, Jeremy F. Huckins, William M. Kelley, and Steven E. Petersen. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral Cortex*, 26(1):288–303, 2016. doi: 10.1093/cercor/bhu239.

A Related literature

The brain connectome is commonly modeled as a network of interconnected regions [1] whose activity is driven by a small number of latent components [27–29]. Message-passing architectures using GCN, GraphSAGE, and/or GAT convolutions [5, 4, 6] have become the backbone of brain-connectome-tailored models [20, 24–26] and of standardized benchmarks [2, 3], with the graph topology typically obtained by thresholding the same empirical matrix that serves as the node feature map.

A growing body of empirical evidence questions the value of message passing on the brain connectome: graph-agnostic classifiers on the connectivity matrix match or outperform GNNs and graph transformers [8, 7], and on the NeuroGraph benchmarks accuracy is unchanged when edges are randomly removed, even all of them [9].

On the theoretical side, the analysis of message passing [31] has focused on node-level tasks: convolutions act, once linearized, as low-pass filters [21, 22], which explains why models without nonlinearities retain most of their accuracy [21, 23], while repeated aggregation leads to oversmoothing [10, 13]. Closest to our setting, Keriven [11] shows on a latent-space random graph that a finite number of mean-aggregation steps improves node-level regression before oversmoothing sets in.

B Full theoretical analysis

B.1 Technical preliminaries

We state important definitions and technical preliminaries that we build on in the results that follow.

Lemma 1 (Wishart noise moments)**.** *Let $\Sigma \in \mathbb{R}^{p\times p}$ be a positive-definite covariance matrix and let $m\hat{\Sigma} \sim \mathcal{W}_p(m, \Sigma)$ [32]. Then $\hat{\Sigma} = \Sigma + H$ with:*

1. $\mathbb{E}_\eta[H_{jk} \mid \Sigma] = 0$,
2. $\mathrm{Cov}_\eta(H_{jk},\, H_{lm} \mid \Sigma) = \dfrac{\Sigma_{jl}\,\Sigma_{km} + \Sigma_{jm}\,\Sigma_{kl}}{m}$.

Next, we show that $\hat{A}$ and $A^{\perp}$ have the same marginal graph distribution. Their difference is the dependence between $\hat{A}$ and the feature noise; see Proposition 3, where we isolate the remaining difference, namely their joint distribution with the feature noise.

Lemma 2 (Marginal equivalence)**.** *Let $\hat{A}^{(i)} = \mathbb{1}\{\hat{\Sigma}^{(i)} > \delta\}$ and $A^{\perp(i)} = \mathbb{1}\{\widehat{\Sigma}^{\perp(i)} > \delta\}$, where $\widehat{\Sigma}^{\perp(i)} = \Sigma^{(i)} + H^{\perp(i)}$ with $H^{\perp(i)} \perp\!\!\!\perp H^{(i)} \mid \Sigma^{(i)}$ and $H^{\perp(i)} \overset{d}{=} H^{(i)} \mid \Sigma^{(i)}$. Then:*

1. *$\hat{A}^{(i)}$ and $A^{\perp(i)}$ have the same conditional distribution given $\Sigma^{(i)}$.*
2. *$A^{\perp(i)} \perp\!\!\!\perp H^{(i)} \mid \Sigma^{(i)}$, whereas $\hat{A}^{(i)}$ depends on $H^{(i)}$ through the shared thresholding.*

Consequently, all marginal quantities (edge probabilities, expected neighbor counts, expected degree) are identical for $\hat{A}$ and $A^{\perp}$. The two differ only in their joint distribution with the feature noise $\eta = \mathrm{vec}(H)$.

With vec stacking columns, $\mathrm{vec}(ABC) = (C^\top \otimes A)\,\mathrm{vec}(B)$; in particular, $\mathrm{vec}(R\widehat{\Sigma}) = (I_p \otimes R)\,\mathrm{vec}(\widehat{\Sigma})$. We also define the commutation matrix as it plays a crucial role when working with vectorized covariance matrices.

Definition 1 (Commutation matrix)**.** The commutation matrix $K_{p,p} \in \mathbb{R}^{p^2\times p^2}$ is the unique permutation matrix satisfying $K_{p,p}\,\mathrm{vec}(A) = \mathrm{vec}(A^\top)$ for every $A \in \mathbb{R}^{p\times p}$ [33].

Writing $\Psi(\Sigma) := (I_{p^2} + K_{p,p})(\Sigma \otimes \Sigma)$, from Lemma 1, we note:

1. $K_{p,p}$ is symmetric and involutory (i.e. $K_{p,p}^2 = I_{p^2}$), and $K_{p,p}(B \otimes C) = (C \otimes B)K_{p,p}$ for all $B, C \in \mathbb{R}^{p\times p}$ [33]; in particular $K_{p,p}$ commutes with $\Sigma \otimes \Sigma$, so $\Psi(\Sigma)$ is symmetric.
2. $\Psi(\Sigma) \succeq 0$, since $T^{-1}\Psi(\Sigma)$ is a conditional covariance matrix (Lemma 1).

3. $\Psi(\Sigma)\,\mathrm{vec}(W) = 0$ for every antisymmetric $W = -W^\top$: indeed $\Sigma W \Sigma$ is again antisymmetric, and $(I_{p^2} + K_{p,p})\,\mathrm{vec}(M) = \mathrm{vec}(M) + \mathrm{vec}(M^\top) = 0$ for antisymmetric M. Hence $\Psi(\Sigma)$ is singular, with $\mathrm{rank}\,\Psi(\Sigma) \le p(p+1)/2$.

B.2 Statistical framework

For each subject $i = 1, \dots, n$, the centered time series $Z^{(i)} \in \mathbb{R}^{T\times p}$ has rows $Z_t^{(i)} \overset{\text{i.i.d.}}{\sim} N(0, \Sigma^{(i)})$, $t = 1, \dots, T$, and we observe the empirical covariance

$$\widehat{\Sigma}^{(i)} = \frac{1}{T}\sum_{t=1}^{T} Z_t^{(i)} Z_t^{(i)\top} = \Sigma^{(i)} + H^{(i)}.$$

The analysis is in the population regime, $n \to \infty$. Let

$$x := \mathrm{vec}(\Sigma), \qquad \widehat{x} := \mathrm{vec}(\widehat{\Sigma}) = x + \eta, \qquad \eta := \mathrm{vec}(H).$$

By Lemma 1,

$$\mathbb{E}_\eta[\eta \mid \Sigma] = 0, \qquad \mathrm{Cov}_\eta(\eta \mid \Sigma) = T^{-1}\Psi(\Sigma), \qquad \Psi(\Sigma) := (I_{p^2} + K_{p,p})(\Sigma \otimes \Sigma).$$

The demeaned estimator has $T-1$ degrees of freedom; every statement below holds with $T-1$ in place of T. The target depends on the latent connectome, $Y = f^\star(\Sigma) + \varepsilon$, with ε independent of (Σ, H) and of any randomness used to build a graph, and $\mathbb{E}_\Sigma\|\Sigma\|_F^4 < \infty$. We restrict $f^\star$ to the linear class $\mathcal{F} = \{x \mapsto x^\top\theta\}$, so that

$$Y = h(\Sigma) + \varepsilon, \qquad h(\Sigma) := x^\top\beta^\star = \langle \Sigma, B^\star\rangle, \qquad \mathrm{vec}(B^\star) = \beta^\star, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2),$$

with $B^\star$ symmetric without loss of generality, since $\langle \Sigma, B^\star\rangle$ only sees the symmetric part of $B^\star$. Since $\widehat{\Sigma}$ is symmetric, the second-moment matrix of $\widehat{x}$ and the Gram matrices below are singular by Definition 1; all inverses are Moore–Penrose pseudo-inverses. For the raw feature map $\widehat{x}$ we write

$$S_0 := \mathbb{E}_\Sigma[xx^\top], \qquad \Gamma_0 := \mathbb{E}_\Sigma[\Psi(\Sigma)], \qquad c_0 := \mathbb{E}_\Sigma[xh(\Sigma)] = S_0\beta^\star.$$

Thus

$$\mathrm{MSE}_{\mathrm{raw}}(T) = \mathbb{E}_\Sigma[h(\Sigma)^2] - c_0^\top(S_0 + T^{-1}\Gamma_0)^\dagger c_0 + \sigma_\varepsilon^2,$$

We now state and prove Remark 1, whose S, Γ are the S_0, Γ_0 above and whose $\mathrm{MSE}_{\mathrm{raw}}(T)$ is the display above written through $\Lambda(T)$.

Remark 1 (Raw Feature Map)**.** Let $y_i = x_i^\top\beta^\star + \varepsilon_i$, where $\Sigma^{(i)} \sim P_\Sigma$ i.i.d. across subjects, $x_i = \mathrm{vec}(\Sigma^{(i)})$, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ independent of $(\Sigma^{(i)}, \eta_i)$, and $\hat{x}_i = x_i + \eta_i$ with $\mathbb{E}_{\eta_i}[\eta_i \mid \Sigma^{(i)}] = 0$. Let

$$\mathrm{Cov}_{\eta_i}(\eta_i \mid \Sigma^{(i)}) = \frac{1}{T}\Psi^{(i)}, \qquad \Psi^{(i)} = (I_{p^2} + K_{p,p})(\Sigma^{(i)} \otimes \Sigma^{(i)}),$$

where $K_{p,p}$ is the commutation matrix (Definition 1). Write $S = \mathbb{E}_\Sigma[xx^\top]$ and $\Gamma = \mathbb{E}_\Sigma[\Psi]$. Since the features are vectorized symmetric matrices, $S + \frac{1}{T}\Gamma$ is singular by Appendix B.1, so inverses below are Moore–Penrose pseudo-inverses. The population raw-feature OLS coefficient is

$$\beta^*_{\mathrm{raw}}(T) = \Lambda(T)\beta^\star, \qquad \Lambda(T) := (S + \tfrac{1}{T}\Gamma)^\dagger S,$$

so, as $T \to \infty$, provided the support of P_Σ spans the symmetric matrices, $\beta^*_{\mathrm{raw}}(T) \to S^\dagger S\,\beta^\star$, the orthogonal projection of $\beta^\star$ onto $\mathrm{range}(S)$; the limiting predictor nevertheless equals $x^\top\beta^\star$ almost surely. Its population test error is

$$\mathrm{MSE}_{\mathrm{raw}}(T) = \beta^{\star\top}(I - \Lambda(T))^\top S(I - \Lambda(T))\beta^\star + \frac{1}{T}\beta^{\star\top}\Lambda(T)^\top\Gamma\Lambda(T)\beta^\star + \sigma_\varepsilon^2.$$

Proof. For any coefficient γ, conditional centering of η and independence of ε make the cross terms vanish:

$$\begin{aligned}\mathbb{E}_{\Sigma,\eta,\varepsilon}(y - \hat{x}^\top\gamma)^2 &= \mathbb{E}_{\Sigma,\eta,\varepsilon}\{x^\top(\beta^\star - \gamma) - \eta^\top\gamma + \varepsilon\}^2 \\ &= (\beta^\star - \gamma)^\top S(\beta^\star - \gamma) + T^{-1}\gamma^\top\Gamma\gamma + \sigma_\varepsilon^2.\end{aligned}$$

Minimizing this quadratic gives $(S + T^{-1}\Gamma)\gamma = S\beta^\star$, whose minimum-norm solution is $\beta^*_{\mathrm{raw}}(T) = \Lambda(T)\beta^\star$. Substituting into the same expansion gives the stated MSE.

Under the spanning assumption, S is invertible on the symmetric subspace, while both S and Γ vanish on its orthogonal complement. Thus $\Lambda(T) \to S^\dagger S$. Since x belongs to that subspace, $x^\top S^\dagger S\beta^\star = x^\top\beta^\star$: the limiting predictor recovers the noiseless target. □

A linear graph convolution replaces $\widehat{\Sigma}$ by $R(A)\widehat{\Sigma}$, where $A \in \{0,1\}^{p\times p}$ is a symmetric adjacency matrix and $R(A)$ its message-passing operator: the layer $X' = \sigma(R(A)XW)$ with the weights and the activation removed. The smoothed feature map is

$$\widetilde{x}_A := \mathrm{vec}(R(A)\widehat{\Sigma}) = \mathcal{R}_A\widehat{x}, \qquad \mathcal{R}_A := I_p \otimes R(A),$$

and, for a graph A, $\mathrm{MSE}_A(T) := \min_\gamma \mathbb{E}_{\Sigma,\eta,\varepsilon,A}[(Y - \gamma^\top \widetilde{x}_A)^2]$ is the optimized population risk, the object compared throughout; $\mathrm{MSE}_{\mathrm{raw}}(T)$ above is its value for the raw features $\widehat{x}$. For a binary adjacency matrix A, the SAGE mean-aggregation operator $R = R(A)$ is

$$\widetilde{A} = A + I_p, \qquad \widetilde{D} = \mathrm{diag}(\widetilde{A}\mathbf{1}), \qquad R = \widetilde{D}^{-1}\widetilde{A}.$$

We compare four graphs, thresholded entrywise off the diagonal at a fixed δ:

1. the population graph $A^{\mathrm{pop}} = \mathbb{1}\{\Sigma_0 > \delta\}$, with $\Sigma_0 = \mathbb{E}_\Sigma[\Sigma]$;
2. the oracle graph $A^{\mathrm{oracle}} = \mathbb{1}\{\Sigma > \delta\}$;
3. the same-sample graph $\widehat{A} = \mathbb{1}\{\widehat{\Sigma} > \delta\}$;
4. the independent graph $A^\perp = \mathbb{1}\{\widehat{\Sigma}^\perp > \delta\}$, $\widehat{\Sigma}^\perp = \Sigma + H^\perp$, with $H^\perp \perp\!\!\!\perp H \mid \Sigma$ and $H^\perp \overset{d}{=} H \mid \Sigma$.

B.3 General errors-in-variables analysis

Predicting from the smoothed features $\widetilde{x}_A = \mathcal{R}_A\widehat{x}$ rather than from $\widehat{x}$ is still an errors-in-variables regression, now with the operator $\mathcal{R}_A$ sitting between the noise and the coefficient. We characterize its population MSE without assumptions on Σ, $B^\star$ or A, distinguishing the case where the graph is independent of the feature noise from the case where graph and features are derived from the same sample. The fixed-operator case, in which the graph does not vary across subjects, is treated separately: there message passing cannot improve on raw features.

Proposition 4 (Population MSE under message passing)**.** *For a graph A, let $R(A) \in \mathbb{R}^{p\times p}$ denote its convolution operator ($R(A) = \widetilde{D}^{-1}\widetilde{A}$ for GraphSAGE), $\mathcal{R}_A := I_p \otimes R(A)$ the corresponding vectorized operator, so that $\mathrm{vec}(R(A)\widehat{\Sigma}) = \mathcal{R}_A\widehat{x}$, and $\widetilde{x}_A := \mathcal{R}_A(x+\eta)$ the smoothed feature map, where $h(\Sigma) := \mathrm{vec}(\Sigma)^\top\beta^\star$ is the noiseless regression target and Ψ is as in Remark 1.*

(a) Conditionally independent graph. *Suppose $A \perp\!\!\!\perp H \mid \Sigma$, as for A^{pop}, A^{oracle}, and $A^\perp$, and $\varepsilon \perp\!\!\!\perp (\Sigma, \eta, A)$. Define*

$$S_A := \mathbb{E}_{\Sigma,A}[\mathcal{R}_A x x^\top \mathcal{R}_A^\top], \qquad \Gamma_A := \mathbb{E}_{\Sigma,A}[\mathcal{R}_A \Psi \mathcal{R}_A^\top], \qquad c_A := \mathbb{E}_{\Sigma,A}[\mathcal{R}_A x\, h(\Sigma)].$$

Then the population OLS coefficient and test MSE are

$$\beta_A^*(T) = (S_A + T^{-1}\Gamma_A)^\dagger c_A, \qquad \mathrm{MSE}_A(T) = \mathbb{E}_\Sigma[h(\Sigma)^2] - c_A^\top(S_A + T^{-1}\Gamma_A)^\dagger c_A + \sigma_\varepsilon^2.$$

(b) Same-sample graph. *Suppose $A = \widehat{A}(\widehat{\Sigma})$ is the same-sample graph, so that $\mathcal{R}_{\widehat{A}}$ is $\widehat{\Sigma}$-measurable and no longer independent of η given Σ, and $\varepsilon \perp\!\!\!\perp (\Sigma, \eta)$. Define $M_{\widehat{A}} := \mathbb{E}_{\Sigma,\eta}[\widetilde{x}_{\widehat{A}}\widetilde{x}_{\widehat{A}}^\top]$, $c_{\widehat{A}} := \mathbb{E}_{\Sigma,\eta}[\widetilde{x}_{\widehat{A}}\, h(\Sigma)]$. Then $\beta_{\widehat{A}}^*(T) = M_{\widehat{A}}^\dagger c_{\widehat{A}}$ and $\mathrm{MSE}_{\mathrm{same}}(T) = \mathbb{E}_\Sigma[h(\Sigma)^2] - c_{\widehat{A}}^\top M_{\widehat{A}}^\dagger c_{\widehat{A}} + \sigma_\varepsilon^2$, where*

$$M_{\widehat{A}} = \mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}} x x^\top \mathcal{R}_{\widehat{A}}^\top] + \underbrace{\mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}} x \eta^\top \mathcal{R}_{\widehat{A}}^\top] + \mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}} \eta x^\top \mathcal{R}_{\widehat{A}}^\top]}_{\textit{cross terms}} + \mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}} \eta\eta^\top \mathcal{R}_{\widehat{A}}^\top],$$

$$c_{\widehat{A}} = \mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}} x x^\top]\beta^\star + \mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}} \eta x^\top]\beta^\star.$$

Proof of (a). Since $A \perp\!\!\!\perp H \mid \Sigma$ and $\varepsilon \perp\!\!\!\perp (\Sigma, \eta, A)$, we have $\mathbb{E}_\eta[\eta \mid \Sigma, A] = 0$ and $\mathrm{Cov}_\eta(\eta \mid \Sigma, A) = T^{-1}\Psi$. By the tower property,

$$\mathbb{E}_{\Sigma,\eta,A}[\widetilde{x}_A\widetilde{x}_A^\top] = S_A + T^{-1}\Gamma_A, \qquad \mathbb{E}_{\Sigma,\eta,\varepsilon,A}[\widetilde{x}_A\, Y] = c_A,$$

and the normal equations $(S_A + T^{-1}\Gamma_A)\gamma = c_A$ are solvable as in Remark 1, and the minimal-norm solution $(S_A + T^{-1}\Gamma_A)^\dagger c_A$ yields $\mathrm{MSE}_A(T)$. □

Proof of (b). The expansions follow by linearity from $\widetilde{x}_{\widehat{A}} = \mathcal{R}_{\widehat{A}}(x+\eta)$, with $\mathbb{E}_{\Sigma,\eta,\varepsilon}[\widetilde{x}_{\widehat{A}}\,\varepsilon] = 0$. No term vanishes in the general case because $\mathcal{R}_{\widehat{A}}$ is not conditionally independent of η given Σ. The coefficient and MSE follow as in Remark 1. If $\mathcal{R}_{\widehat{A}} \perp\!\!\!\perp \eta \mid \Sigma$, then $\mathbb{E}_\eta[\eta \mid \Sigma, \widehat{A}] = 0$, so the cross terms and $\mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}}\,\eta x^\top]\beta^\star$ vanish, while the tower property gives $\mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}}\,\eta\eta^\top\mathcal{R}_{\widehat{A}}^\top] = T^{-1}\mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}}\,\Psi\,\mathcal{R}_{\widehat{A}}^\top]$, recovering case (a). □

For the same-sample graph, the operator depends on the same noise used in the features. The previous cancellation no longer occurs.

Proposition 5 (No linear gain from fixed message passing)**.** *For any fixed $\mathcal{R} \in \mathbb{R}^{p^2\times p^2}$,*

$$\mathcal{F}_{\mathrm{MP}}(\mathcal{R}) := \{x \mapsto \theta^\top\mathcal{R}x : \theta \in \mathbb{R}^{p^2}\} \subseteq \{x \mapsto \beta^\top x : \beta \in \mathbb{R}^{p^2}\} =: \mathcal{F}_{\mathrm{raw}}, \qquad \theta^\top\mathcal{R}x = (\mathcal{R}^\top\theta)^\top x.$$

Consequently, for any loss ℓ, including $\ell(y,\hat{y}) = (y-\hat{y})^2$: $\inf_\beta \mathbb{E}_{Y,x}\,\ell(Y,\beta^\top x) \le \inf_\theta \mathbb{E}_{Y,x}\,\ell(Y,\theta^\top\mathcal{R}x)$.

Proof. As θ ranges over $\mathbb{R}^{p^2}$, $\mathcal{R}^\top\theta$ ranges over $\mathrm{range}(\mathcal{R}^\top)$, so $\mathcal{F}_{\mathrm{MP}}(\mathcal{R}) = \{x \mapsto \beta^\top x : \beta \in \mathrm{range}(\mathcal{R}^\top)\} \subseteq \mathcal{F}_{\mathrm{raw}}$, and the infimum over a subset is at least the infimum over the set. For squared loss the raw minimizers form the solution set of $\mathbb{E}_x[xx^\top]\beta = \mathbb{E}_{Y,x}[xY]$ (consistent, as in Remark 1); the inequality is strict iff this set is disjoint from $\mathrm{range}(\mathcal{R}^\top)$. □

Corollary 1 (Equality when the fixed operator is invertible)**.** *If $\mathcal{R}$ is invertible, then*

$$\mathcal{F}_{\mathrm{MP}}(\mathcal{R}) = \mathcal{F}_{\mathrm{raw}}.$$

Hence the optimal unregularized linear prediction risks are equal:

$$\inf_\beta \mathbb{E}_{Y,x}\left[\left(Y-\beta^\top x\right)^2\right] = \inf_\theta \mathbb{E}_{Y,x}\left[\left(Y-\theta^\top\mathcal{R}x\right)^2\right].$$

Proof. Proposition 5 gives

$$\mathcal{F}_{\mathrm{MP}}(\mathcal{R}) \subseteq \mathcal{F}_{\mathrm{raw}}.$$

If $\mathcal{R}$ is invertible, then for every β we may choose

$$\theta = \mathcal{R}^{-\top}\beta,$$

so that

$$\theta^\top\mathcal{R}x = \beta^\top x.$$

Therefore

$$\mathcal{F}_{\mathrm{raw}} \subseteq \mathcal{F}_{\mathrm{MP}}(\mathcal{R}),$$

and the two classes are identical. □

Beyond Proposition 5, neither case of Proposition 4 yields a conclusion without assumptions: the population risks depend on three ingredients the propositions leave free. We fix each under the latent community model of Appendix B.4 and study their interaction as T varies.

1. **Community structure, Σ.** S, $\Gamma = \mathbb{E}_\Sigma[\Psi(\Sigma)]$ and c, hence every MSE above, are functionals of the law of Σ: its community organization and subject-level variability decide what a neighborhood average can remove (Appendix B.4.1).
2. **Predictive signal, $B^\star$.** The signal enters through $R(A)^\top B^\star$, so smoothing preserves it when $B^\star$ is nearly constant over the averaged neighborhoods and distorts it when the signal lies in differences between entries: community-smooth versus edge or contrast targets (Appendix B.4.3).
3. **Graph and threshold, A and δ.** Which matrix is thresholded (Σ, $\widehat{\Sigma}^\perp$, $\widehat{\Sigma}$, Σ_0) fixes the coupling between graph and feature noise, the split of Proposition 4(a)/(b); the threshold δ fixes the neighborhoods, so for graphs $\mathbb{1}\{\cdot > \delta\}$ the structure reduces to the position of δ relative to the between- and within-community levels of Σ (Appendix B.4.2).

B.4 Latent community model

In order to get more insight into the intractable problem defined in the previous section, we place assumptions on all three ingredients at once: Σ is generated by a latent community model, δ is set relative to its community levels, and $B^\star$ is characterized by its alignment with the communities.

B.4.1 Community structure

We model the p ROIs through $K \leq \min\{p, d\}$ latent communities [1], motivated by low-dimensional representations of brain activity [27–29]. Let $F \in \mathbb{R}^{p\times d}$ collect their latent loadings. The structural parameter is the subject-level variability σ; all other model parameters are held fixed.

For each subject i, the connectome is generated as follows.

1. **Node assignment:** each ROI j belongs to $z_j \in \{1, \dots, K\}$, with balanced communities $|C_k| = q = p/K \in \mathbb{N}$ and $C(j) := C_{z_j}$.
2. **Community centroids:** the vectors $\mu_k \in \mathbb{R}^d$ are fixed across subjects, orthogonal, and satisfy $\|\mu_k\|^2 = \phi^2 d$.
3. **Subject-level variability:** $\nu_j^{(i)} \sim \mathcal{N}(0, \sigma^2 I_d)$ independently across ROIs and subjects.
4. **Latent features:** $f_j^{(i)} = \mu_{z_j} + \nu_j^{(i)}$, with $f_j^{(i)\top}$ the j-th row of $F^{(i)}$.
5. **Connectome:** $\Sigma^{(i)} = F^{(i)}F^{(i)\top} + \psi^2 I_p$.

We assume $\phi, \psi > 0$ and $\sigma \geq 0$. All expectations below keep the partition and centroids fixed.

Lemma 3 (Community levels and fluctuations)**.** *For $j \neq k$,*

$$\mathbb{E}_\Sigma \Sigma_{jk} = \begin{cases} \mu_W := \phi^2 d, & z_j = z_k, \\ \mu_B := 0, & z_j \neq z_k, \end{cases} \qquad \mathrm{Var}_\Sigma(\Sigma_{jk}) = s^2 := 2\sigma^2\phi^2 d + \sigma^4 d.$$

On the diagonal,

$$m_D := \mathbb{E}_\Sigma \Sigma_{jj} = d(\phi^2 + \sigma^2) + \psi^2, \qquad \mathrm{Var}_\Sigma(\Sigma_{jj}) = 2s^2.$$

For any $u > 0$, with probability at least $1 - \{\binom{p}{2} + 2p\}s^2/u^2$, simultaneously

$$\max_{j<k} |\Sigma_{jk} - \mathbb{E}_\Sigma \Sigma_{jk}| \leq u, \qquad \max_j \Sigma_{jj} \leq m_D + u.$$

On this event, the measurement-noise factors satisfy

$$v_{jk} := \Sigma_{jj}\Sigma_{kk} + \Sigma_{jk}^2 \leq 2(m_D + u)^2.$$

Proof. Expanding $f_j^\top f_k$ and $\|f_j\|^2$ gives the moments by independence and Gaussianity. Chebyshev's inequality and union bound give the probability bound. Finally, positive semidefiniteness gives $\Sigma_{jk}^2 \leq \Sigma_{jj}\Sigma_{kk}$. □

At $\sigma = 0$, every subject has the same connectome. Increasing σ introduces subject-level variation and increases the spread s relative to the fixed community separation $\phi^2 d$.

B.4.2 Graph construction and threshold regimes

All four graphs are entrywise thresholdings $\mathbb{1}\{\cdot > \delta\}$ of a matrix (Σ, $\widehat{\Sigma}^\perp$, $\widehat{\Sigma}$ or Σ_0), so their structure is set by where δ sits relative to the between- and within-community levels μ_B and μ_W of Σ. We distinguish:

1. **Disconnected graph, $\delta = +\infty$.** $A = 0$; after adding self-loops $R = I_p$, and the smoothed map is the raw map.
2. **Complete graph, $\delta = -\infty$.** $\widetilde{A} = \mathbf{1}\mathbf{1}^\top$ and $R = P_{\mathbf{1}} := \frac{1}{p}\mathbf{1}\mathbf{1}^\top$, so $P_{\mathbf{1}}\widehat{\Sigma} = \mathbf{1}\widehat{s}^\top$ with $\widehat{s} := \frac{1}{p}\widehat{\Sigma}\mathbf{1}$: the p^2 entries collapse to the p ROI averages.

3. **Intermediate threshold in the regime,** $\mu_B < \delta < \mu_W$**,** with margin $\gamma_\delta := \min\{\mu_W - \delta, \delta - \mu_B\} > 0$. The oracle graph selects mostly within-community neighbors and

$$R(A^{\text{oracle}}) = P_C + \Delta R_\delta, \qquad (P_C)_{j\ell} = \frac{1}{q}\mathbb{1}\{\ell \in C(j)\},$$

where P_C depends only on the community partition and ΔR_δ collects the subject-specific deviations of the oracle neighborhoods.

4. **Finite threshold outside the regime.** If $\delta < \mu_B$, most between-community pairs are also connected: the graph is dense and R approaches $P_{\mathbf{1}}$. If $\delta > \mu_W$, most within-community pairs are dropped: the graph is sparse and R approaches I_p, the surviving edges being the largest entries of $\Sigma^{(i)}$ rather than the partition. In neither case does R split into a community and a subject-specific component.

We focus on case 3 when interpreting the results through community structure. The oracle-smoothing bounds themselves hold for any fixed threshold.

B.4.3 Predictive signal

Let $B^\star \in \mathbb{R}^{p\times p}$ be the matrix form of the regression coefficient, $\beta^\star = \text{vec}(B^\star)$, so that $h(\Sigma) = \langle B^\star, \Sigma\rangle$. We say that $B^\star$ is community-smooth if rows from the same community have similar predictive profiles,

$$\max_j \max_{\ell\in C(j)} \sum_{k=1}^{p} \left(B^\star_{jk} - B^\star_{\ell k}\right)^2 \le \kappa_\beta^2 \quad \text{for some } \kappa_\beta \ge 0;$$

the ideal case $\kappa_\beta = 0$ has $B^\star_{jk}$ depending only on (z_j, z_k). We study two families of targets.

- **Targets constant over communities:** the within-community strength of a fixed ROI a, with e_a the a-th standard basis vector and $\mathbf{1}_{C(a)}$ the indicator of the community containing a,

$$B^\star_{a,C} := \tfrac{1}{2}\left(e_a \mathbf{1}_{C(a)}^\top + \mathbf{1}_{C(a)} e_a^\top\right), \qquad h_{a,C}(\Sigma) = e_a^\top \Sigma \mathbf{1}_{C(a)};$$

and the total within-community connectivity $h_C(\Sigma) = |C|^{-2}\mathbf{1}_C^\top \Sigma \mathbf{1}_C$.

- **Targets supported on single edges or on within-community differences:** the single edge $h_{ab}(\Sigma) = \Sigma_{ab}$, and the within-community contrast

$$B^\star_{\text{contrast}} := \tfrac{1}{2}\{(e_a - e_b)v^\top + v(e_a - e_b)^\top\}, \qquad v = \mathbf{1}_{C(a)}.$$

B.5 Results under the latent community model

B.5.1 Case 1: oracle graph

The oracle graph is constructed from the true connectome. Specifically,

$$A^{\text{oracle}} = \mathbf{1}\{\Sigma > \delta\}, \qquad R(A^{\text{oracle}}) = \widetilde{D}^{-1}(A^{\text{oracle}} + I_p).$$

Proposition 4(a) applies directly to this type of graph construction.

Proposition 1 (Oracle smoothing). Under the EIV assumptions of Remark 1, fix any threshold δ; the oracle operator $R(A^{\text{oracle}})$ is a function of Σ. Assume all displayed moments are finite and define the squared bias and the variance reduction:

$$\mathcal{B}_{A^{\text{oracle}}}(\delta) := \mathbb{E}_\Sigma\left[\langle B^\star, R(A^{\text{oracle}})\Sigma - \Sigma\rangle^2\right],$$
$$\Delta\mathcal{V}_{A^{\text{oracle}}}(\delta) := T\,\mathbb{E}_\Sigma\left[\text{Var}_\eta\{\langle B^\star, H\rangle \mid \Sigma\} - \text{Var}_\eta\{\langle B^\star, R(A^{\text{oracle}})H\rangle \mid \Sigma\}\right].$$

Sufficient condition. The optimized population MSEs satisfy, for every $T > 0$,

$$\text{MSE}_{\text{oracle}}(T,\delta) - \text{MSE}_{\text{raw}}(T) \le \mathcal{B}_{A^{\text{oracle}}}(\delta) - \frac{\Delta\mathcal{V}_{A^{\text{oracle}}}(\delta)}{T} + r_{\text{raw}}(T),$$

where

$$r_{\text{raw}}(T) := T^{-2}\beta^{\star\top}\Gamma(S + T^{-1}\Gamma)^\dagger\Gamma\beta^\star \ge 0.$$

If the support of Σ spans the symmetric matrices (Remark 1), as in the latent community model with $\sigma > 0$, then $r_{\mathrm{raw}}(T) = O(T^{-2})$. Thus oracle smoothing improves whenever $\mathcal{B}_{A^{\mathrm{oracle}}}(\delta) + r_{\mathrm{raw}}(T) < \Delta\mathcal{V}_{A^{\mathrm{oracle}}}(\delta)/T$.

Necessary condition. Let

$$b_\delta := \min_{\gamma \in \mathbb{R}^{p^2}} \mathbb{E}_\Sigma \left[\left\{ h(\Sigma) - \gamma^\top \operatorname{vec}(R(A^{\mathrm{oracle}})\Sigma) \right\}^2 \right]$$

be the minimum noiseless prediction error, with γ shared across subjects. If $b_\delta > 0$, then

$$\mathrm{MSE}_{\mathrm{oracle}}(T, \delta) < \mathrm{MSE}_{\mathrm{raw}}(T) \quad \Longrightarrow \quad T < \frac{\beta^{\star\top} \Gamma \beta^\star}{b_\delta}.$$

Proof. Write $R = R(A^{\mathrm{oracle}})$ and $\widetilde{x}_{A^{\mathrm{oracle}}} = \operatorname{vec}(R\widehat{\Sigma}) = \operatorname{vec}(R\Sigma) + \operatorname{vec}(RH)$. Since R is a function of Σ and $\mathbb{E}_\eta[\eta \mid \Sigma] = 0$, $\mathbb{E}_\eta[\operatorname{vec}(RH) \mid \Sigma] = 0$, and ε is independent of (Σ, H), so for any $\gamma \in \mathbb{R}^{p^2}$ the cross terms vanish and

$$\mathbb{E}_{\Sigma,\eta,\varepsilon}\left[(Y - \gamma^\top \widetilde{x}_{A^{\mathrm{oracle}}})^2\right] = \mathbb{E}_\Sigma\left[\{h(\Sigma) - \gamma^\top \operatorname{vec}(R\Sigma)\}^2\right] + \mathbb{E}_\Sigma\left[\operatorname{Var}_\eta(\gamma^\top \operatorname{vec}(RH) \mid \Sigma)\right] + \sigma_\varepsilon^2. \quad (6)$$

Evaluating at $\gamma = \beta^\star$, rather than at the fitted coefficient, the first term is $\mathcal{B}_{A^{\mathrm{oracle}}}$ and the second is $T^{-1}(\beta^{\star\top}\Gamma\beta^\star - \Delta\mathcal{V}_{A^{\mathrm{oracle}}})$, by definition of $\Delta\mathcal{V}_{A^{\mathrm{oracle}}}$ and $T\,\mathbb{E}_\Sigma[\operatorname{Var}_\eta(\langle B^\star, H\rangle \mid \Sigma)] = \beta^{\star\top}\Gamma\beta^\star$. Hence

$$\mathrm{MSE}_{\mathrm{oracle}} - \sigma_\varepsilon^2 \le \mathcal{B}_{A^{\mathrm{oracle}}} + T^{-1}\left(\beta^{\star\top}\Gamma\beta^\star - \Delta\mathcal{V}_{A^{\mathrm{oracle}}}\right).$$

For raw features, Remark 1, with $M = S + T^{-1}\Gamma$, gives $\mathrm{MSE}_{\mathrm{raw}} - \sigma_\varepsilon^2 = \beta^{\star\top}(S - SM^\dagger S)\beta^\star$. Substituting $S = M - T^{-1}\Gamma$, valid since $M \succeq T^{-1}\Gamma$, gives

$$S - SM^\dagger S = T^{-1}\Gamma - T^{-2}\Gamma M^\dagger \Gamma, \qquad \text{hence} \qquad \mathrm{MSE}_{\mathrm{raw}} - \sigma_\varepsilon^2 = T^{-1}\beta^{\star\top}\Gamma\beta^\star - r_{\mathrm{raw}}(T),$$

with $r_{\mathrm{raw}}(T) \ge 0$ because $M^\dagger \succeq 0$. Subtracting proves the sufficient comparison.

By (6), $\mathrm{MSE}_{\mathrm{oracle}} - \sigma_\varepsilon^2 \ge b_\delta$, while $\mathrm{MSE}_{\mathrm{raw}} - \sigma_\varepsilon^2 \le T^{-1}\beta^{\star\top}\Gamma\beta^\star$. Strict improvement therefore gives $b_\delta < T^{-1}\beta^{\star\top}\Gamma\beta^\star$, the necessary bound on T.

Finally, under the spanning condition S is invertible on the symmetric subspace, which contains $\operatorname{range}(\Gamma)$ (Definition 1). There $M \succeq S$, so $M^\dagger \preceq S^\dagger$ and

$$r_{\mathrm{raw}}(T) \le T^{-2}\beta^{\star\top}\Gamma S^\dagger \Gamma\beta^\star = O(T^{-2}).$$

In the latent community model with $\sigma > 0$, taking $F = ue_1^\top$ shows that the support of Σ contains $uu^\top + \psi^2 I_p$ for every $u \in \mathbb{R}^p$, and these matrices span the symmetric matrices. □

The sufficient condition above can now be interpreted through the three ingredients that determine its bias and variance terms.

Predictive signal, $B^\star$. Since the rows of $R(A^{\mathrm{oracle}})$ sum to one,

$$\langle B^\star, R(A^{\mathrm{oracle}})\Sigma - \Sigma\rangle = \sum_{j,k} B^\star_{jk} \sum_\ell R(A^{\mathrm{oracle}})_{j\ell}\left(\Sigma_{\ell k} - \Sigma_{jk}\right),$$

so the bias is small when the ROIs averaged with j have rows close to $\Sigma_{j\cdot}$ on the support of $B^\star$: community-level targets depend on sums of entries, whereas edge targets depend on specific entries.

Threshold, δ. For a given target, the threshold determines which ROIs enter this average. Lowering δ adds neighbors while raising it removes them. These changes affect both bias and noise propagation, so more averaging is not automatically better.

Disconnected graph. For $\delta = +\infty$, $R(A^{\mathrm{oracle}}) = I_p$ (Appendix B.4.2): the smoothed and raw features coincide, so $\mathrm{MSE}_{\mathrm{oracle}}(T, +\infty) = \mathrm{MSE}_{\mathrm{raw}}(T)$ for every T, and $\mathcal{B}_{A^{\mathrm{oracle}}} = \Delta\mathcal{V}_{A^{\mathrm{oracle}}} = 0$.

Complete graph. For $\delta = -\infty$, $R(A^{\mathrm{oracle}}) = P_{\mathbf{1}}$ (Appendix B.4.2) is the same for every subject and $P_{\mathbf{1}}\widehat{\Sigma} = \mathbf{1}\widehat{s}^\top$, so the smoothed class is the set of linear functions of the degree vector $\widehat{s} = \widehat{\Sigma}\mathbf{1}/p$. By Proposition 5, $\mathrm{MSE}_J(T) \ge \mathrm{MSE}_{\mathrm{raw}}(T)$ for every T, with $J = \mathbf{1}\mathbf{1}^\top$.

Intermediate threshold. For $\mu_B < \delta < \mu_W$, the population operator is P_C. Writing $R(A^{\text{oracle}}) = P_C + \Delta R_\delta$ gives

$$\mathcal{B}_{A^{\text{oracle}}}(\delta) = \mathbb{E}_\Sigma\Big[\{\langle B^\star, (P_C - I_p)\Sigma\rangle + \langle B^\star, \Delta R_\delta \Sigma\rangle\}^2\Big].$$

The first term concerns fixed community averaging; the second captures the effect of subject-specific neighborhoods. For exact community smoothness, $P_C B^\star = B^\star$, the first term vanishes, but the second need not. Since fixed averaging cannot improve optimized prediction (Proposition 5), any oracle gain requires subject-specific variation in the operator and not merely an aligned target.

Community structure, σ. At a given threshold, changing σ changes the connectomes being thresholded. It controls deviations from the shared community loadings, affecting both the selected neighborhoods and the measurement-noise factors. For intermediate thresholds, concentration around the community levels relative to the threshold margins makes the oracle close to P_C. Thus clear communities and an aligned target alone do not guarantee improvement: the prediction effect depends on how σ, δ, and $B^\star$ interact.

Together, these ingredients determine whether the noise reduction outweighs the squared bias and the raw-regression correction. The two targets below illustrate this interaction.

Concrete examples. Consider $h_{a,v}(\Sigma) = e_a^\top \Sigma v$, taking $v = \mathbf{1}_{C(a)}$ for the within-community strength or $v = e_b$ for a single edge. On the oracle-smoothed features, consider the candidate predictor

$$\widehat{h}_\delta = e_a^\top R(A^{\text{oracle}})\widehat{\Sigma} v = w_{a,\delta}^\top \widehat{\Sigma} v, \qquad w_{a,\delta} := R(A^{\text{oracle}})^\top e_a.$$

This is a linear predictor on $\text{vec}(R(A^{\text{oracle}})\widehat{\Sigma})$ with fixed coefficient $\text{vec}(e_a v^\top)$; it need not be the optimal fitted coefficient. Its conditional bias and measurement-noise variance factor are

$$\begin{aligned} b_\delta(a, v; \Sigma) &= (w_{a,\delta} - e_a)^\top \Sigma v, \\ V_\delta(a, v; \Sigma) &= (w_{a,\delta}^\top \Sigma w_{a,\delta})(v^\top \Sigma v) + (w_{a,\delta}^\top \Sigma v)^2. \end{aligned}$$

Write V_0 for the same expression with $w_{a,\delta} = e_a$. Conditional centering and the Wishart moments give

$$\mathbb{E}_{\eta,\varepsilon}\big[(y - \widehat{h}_\delta)^2 \mid \Sigma\big] = b_\delta(a, v; \Sigma)^2 + \frac{V_\delta(a, v; \Sigma)}{T} + \sigma_\varepsilon^2.$$

Optimizing the oracle regression can only decrease this risk. Since the optimized raw risk is $\sigma_\varepsilon^2 + T^{-1}\mathbb{E}_\Sigma[V_0] - r_{\text{raw}}(T)$, the argument of Proposition 1 guarantees improvement whenever

$$\mathbb{E}_\Sigma[b_\delta(a, v; \Sigma)^2] + r_{\text{raw}}(T) < \frac{\mathbb{E}_\Sigma[V_0(a, v; \Sigma) - V_\delta(a, v; \Sigma)]}{T},$$

where $r_{\text{raw}}(T)$ is evaluated for the chosen target.

B.5.2 Case 2: independent-copy graph and graph-estimation error

The independent graph is built from a second estimate $\widehat{\Sigma}^\perp = \Sigma + H^\perp$ with $H^\perp \perp\!\!\!\perp H \mid \Sigma$ so that $A^\perp = \mathbf{1}\{\widehat{\Sigma}^\perp > \delta\}$. By Lemma 2 it has the same conditional law as $\widehat{A}$ given Σ but is independent of the feature noise, so Proposition 4(a) applies with $A = A^\perp$. The graph-estimation contribution is $\text{MSE}_{\text{ind}}(T) - \text{MSE}_{\text{oracle}}(T)$: all differences relative to the oracle come from smoothing over the wrong neighborhoods and not from coupling with H.

Proposition 2 (Graph-estimation error for the independent graph). Under the assumptions of Proposition 1, let $A^\perp$ be built from a conditionally independent copy $H^\perp$ of H given Σ, and assume $\mathbb{E}_\Sigma\|\Sigma\|_F^4 < \infty$. Define the edge-error probability and the graph-estimation error:

$$\rho(T, \delta) := \mathbb{P}(A^\perp \neq A^{\text{oracle}}), \qquad \mathcal{E}(T, \delta) := C(1 + T^{-1})\sqrt{\rho(T, \delta)},$$

where C depends only on p, $\|B^\star\|_F$ and $\mathbb{E}_\Sigma\|\Sigma\|_F^4$, not on T or δ. Then, for every $T > 0$,

$$\text{MSE}_{\text{ind}}(T, \delta) - \text{MSE}_{\text{raw}}(T) \le \mathcal{B}_{A^{\text{oracle}}}(\delta) - \frac{\Delta\mathcal{V}_{A^{\text{oracle}}}(\delta)}{T} + \mathcal{E}(T, \delta) + r_{\text{raw}}(T).$$

Thus the independent graph improves whenever $\mathcal{B}_{A^{\text{oracle}}}(\delta) + \mathcal{E}(T, \delta) + r_{\text{raw}}(T) < \Delta\mathcal{V}_{A^{\text{oracle}}}(\delta)/T$.

Proof. As in the proof of Proposition 1, we evaluate both graphs at the true coefficient $\beta^\star$ instead of the fitted one. Given Σ and the graph A, the resulting error, excluding the response noise, is

$$L_A = \langle B^\star, R(A)\Sigma - \Sigma\rangle^2 + \mathrm{Var}_\eta\{\langle B^\star, R(A)H\rangle \mid \Sigma, A\}.$$

For the oracle, its expectation is $\mathcal{B}_{A^{\text{oracle}}} + T^{-1}(\beta^{\star\top}\Gamma\beta^\star - \Delta\mathcal{V}_{A^{\text{oracle}}})$. Hence

$$\mathrm{MSE}_{\text{ind}} - \sigma_\varepsilon^2 \le \mathbb{E}_{\Sigma,H^\perp}[L_{A^\perp}] = \mathbb{E}_\Sigma[L_{A^{\text{oracle}}}] + \mathbb{E}_{\Sigma,H^\perp}[L_{A^\perp} - L_{A^{\text{oracle}}}].$$

The last term is the graph-estimation error. It is zero when $A^\perp = A^{\text{oracle}}$. Since $\|R(A)\|_{\text{op}} \le \sqrt{p}$ and $\|\Psi\|_{\text{op}} \le 2\|\Sigma\|_F^2$, each loss is at most $C_0(1+T^{-1})\|\Sigma\|_F^2$ with $C_0 = 4p\|B^\star\|_F^2$, and Cauchy–Schwarz bounds the last term by $\mathcal{E}$ with $C = C_0(\mathbb{E}_\Sigma\|\Sigma\|_F^4)^{1/2}$. Subtracting $\mathrm{MSE}_{\text{raw}} - \sigma_\varepsilon^2 = T^{-1}\beta^{\star\top}\Gamma\beta^\star - r_{\text{raw}}(T)$ (proof of Proposition 1) gives the bound. □

Proposition 2 asks whether graph estimation is accurate enough to retain the improvement certified for the oracle.

Predictive signal, $B^\star$. The target determines the oracle bias-variance tradeoff and the predictive consequences of changing its neighborhoods, but not the edge-error probabilities. In the preceding examples, neighborhood errors change averages of community strengths or individual connectivity entries. The same graph-estimation accuracy therefore need not give the same prediction benefit for different targets.

Threshold, δ. The threshold determines both the oracle neighborhoods and how reliably they can be estimated. For the following calculation, adopt the Gaussian surrogate $\widehat{\Sigma}_{jk}^\perp \mid \Sigma \sim \mathcal{N}(\Sigma_{jk}, v_{jk}/T)$, where $v_{jk} = \Sigma_{jj}\Sigma_{kk} + \Sigma_{jk}^2 > 0$. Standardizing the threshold-crossing event gives

$$\pi_{jk} := \mathbb{P}(A_{jk}^\perp \neq A_{jk}^{\text{oracle}} \mid \Sigma) = \Phi\left(-\frac{|\Sigma_{jk} - \delta|\sqrt{T}}{\sqrt{v_{jk}}}\right),$$

where Φ is the standard normal CDF. Reliability therefore depends on actual threshold margins relative to measurement noise, not on graph density alone.

Disconnected and complete graphs. At $\delta = +\infty$ both operators equal I_p; at $\delta = -\infty$ both equal $P_{\mathbf{1}}$. In either case, $\rho = \mathcal{E} = 0$. These graphs are perfectly estimated, but cannot improve optimized prediction over raw features.

Intermediate thresholds. For $\mu_B < \delta < \mu_W$, the oracle decomposition gives

$$R(A^\perp) = P_C + \Delta R_\delta + \big\{R(A^\perp) - R(A^{\text{oracle}})\big\}.$$

The last term is the graph-estimation change; ΔR_δ is genuine subject-specific variation, not estimation error. The aim is therefore to recover the oracle neighborhoods, not merely the population partition.

Set $\gamma_\delta = \min\{\mu_W - \delta, \delta - \mu_B\}$. Suppose an event G, with $\mathbb{P}(G^c) \le \epsilon_0$, ensures that all off-diagonal entries lie within $\gamma_\delta/2$ of their community levels and $v_{jk} \le v_{\max} := \max\{v_W, v_B\}$, with deterministic bounds independent of T. On G, the actual margins are at least $\gamma_\delta/2$, so the Gaussian tail bound gives

$$\pi_{jk} \le e^{-T/(8T_{\text{graph}}(\delta))}, \qquad T_{\text{graph}}(\delta) := \frac{v_{\max}}{\gamma_\delta^2}.$$

A union bound over edges, including the exceptional event, yields

$$\rho(T,\delta) \le \min\left\{1,\ \epsilon_0 + \binom{p}{2} e^{-T/(8T_{\text{graph}}(\delta))}\right\}.$$

Community structure, σ. With fixed centroids, changing σ leaves μ_W, μ_B and γ_δ unchanged, but changes concentration around those levels and the measurement-noise scales. Lemma 3 makes this dependence explicit: the preceding event can be obtained with

$$\epsilon_0 = \min\left\{1,\ \frac{4\{\binom{p}{2} + 2p\}s^2}{\gamma_\delta^2}\right\}, \qquad v_W = v_B = 2(m_D + \gamma_\delta/2)^2.$$

The community-margin guarantee is useful only when its exceptional probability is small; an intermediate threshold alone does not ensure reliable graph estimation.

Together, these ingredients determine whether the graph-estimation bound fits within the oracle's sufficient improvement margin. Increasing T improves edge estimation under the surrogate, but also reduces the raw-feature noise available to denoise. An intermediate favorable range may therefore arise, but is not guaranteed. Failure of the sufficient condition remains inconclusive.

B.5.3 Case 3: same-sample graph and double dipping

The standard pipeline thresholds the same empirical covariance that serves as node features, $\widehat{A} = \mathbf{1}\{\widehat{\Sigma} > \delta\}$, so $\mathcal{R}_{\widehat{A}}$ depends on H. By Lemma 2 the independent graph has the same conditional law given Σ but none of the couplings, so $\mathcal{D}_{\mathrm{dd}}(T,\delta) = \mathrm{MSE}_{\mathrm{same}}(T) - \mathrm{MSE}_{\mathrm{ind}}(T)$ isolates the double-dipping contribution.

Proposition 3 (Double-dipping contribution for the same-sample graph). Under the assumptions of Proposition 2, define the double-dipping contribution $\mathcal{D}_{\mathrm{dd}}(T,\delta) := \mathrm{MSE}_{\mathrm{same}}(T,\delta) - \mathrm{MSE}_{\mathrm{ind}}(T,\delta)$. Then, for every $T > 0$,

$$\mathrm{MSE}_{\mathrm{same}}(T,\delta) - \mathrm{MSE}_{\mathrm{raw}}(T) \le \mathcal{B}_{A^{\mathrm{oracle}}}(\delta) - \frac{\Delta\mathcal{V}_{A^{\mathrm{oracle}}}(\delta)}{T} + \mathcal{E}(T,\delta) + \mathcal{D}_{\mathrm{dd}}(T,\delta) + r_{\mathrm{raw}}(T).$$

Thus the same-sample graph improves whenever $\mathcal{B}_{A^{\mathrm{oracle}}}(\delta) + \mathcal{E}(T,\delta) + \mathcal{D}_{\mathrm{dd}}(T,\delta) + r_{\mathrm{raw}}(T) < \Delta\mathcal{V}_{A^{\mathrm{oracle}}}(\delta)/T$. The contribution $\mathcal{D}_{\mathrm{dd}}$ is signed: negative values favor the same-sample graph.

Proof. Since $\mathrm{MSE}_{\mathrm{same}} = \mathrm{MSE}_{\mathrm{ind}} + \mathcal{D}_{\mathrm{dd}}$ by definition, the bound follows from Proposition 2. □

Size of the double-dipping contribution. Suppose in addition that $\mathbb{E}_{\Sigma,\eta}\|H\|_F^4 \le C_H^4 T^{-2}$ and that the positive eigenvalues of $M_{\widehat{A}}$ and $M_{A^\perp}$ are at least $\lambda > 0$ for all $T \ge 1$. Then, for $T \ge 1$,

$$|\mathcal{D}_{\mathrm{dd}}(T,\delta)| \le C_{\mathrm{dd}}\sqrt{\frac{\rho(T,\delta)}{T}},$$

where C_{dd} depends only on p, $\|B^\star\|_F$, C_H, λ and $\mathbb{E}_\Sigma\|\Sigma\|_F^4$, not on T.

Proof. Let $F_A(\gamma) := \mathbb{E}_\Sigma[h^2] - 2\gamma^\top c_A + \gamma^\top M_A\gamma$, with c_A and M_A as in Proposition 4 and $M_{A^\perp} = S_{A^\perp} + T^{-1}\Gamma_{A^\perp}$, so that $\mathrm{MSE}_A = \sigma_\varepsilon^2 + \min_\gamma F_A(\gamma)$. Evaluating each risk at the other graph's minimizer gives

$$q(\beta^*_{\widehat{A}}) \le \mathcal{D}_{\mathrm{dd}} \le q(\beta^*_{A^\perp}), \qquad q(\gamma) := F_{\widehat{A}}(\gamma) - F_{A^\perp}(\gamma) = \gamma^\top\Delta_M\gamma - 2\gamma^\top\Delta_c,$$

with $\Delta_M := M_{\widehat{A}} - M_{A^\perp}$ and $\Delta_c := c_{\widehat{A}} - c_{A^\perp}$. The graphs $\widehat{A}$ and $A^\perp$ have the same law given Σ (Lemma 2), so every term without η cancels and only the noise reuse remains:

$$\Delta_c = \mathbb{E}_{\Sigma,\eta}[(\mathcal{R}_{\widehat{A}} - \mathcal{R}_{A^{\mathrm{oracle}}})\eta h], \qquad \Delta_M = \mathbb{E}_{\Sigma,\eta}[\mathcal{R}_{\widehat{A}}Q_T\mathcal{R}_{\widehat{A}}^\top - \mathcal{R}_{A^{\mathrm{oracle}}}Q_T\mathcal{R}_{A^{\mathrm{oracle}}}^\top],$$

where $Q_T := \widehat{x}\widehat{x}^\top - xx^\top - T^{-1}\Psi$. The oracle terms have mean zero because $\mathbb{E}_\eta[\eta \mid \Sigma] = 0$ and $\mathbb{E}_\eta[Q_T \mid \Sigma] = 0$. Both integrands vanish unless $\widehat{A} \ne A^{\mathrm{oracle}}$, an event of probability ρ. Cauchy–Schwarz with $\|\mathcal{R}_A\|_{\mathrm{op}} \le \sqrt{p}$ and the fourth moments gives $\|\Delta_c\| + \|\Delta_M\|_{\mathrm{op}} \le C_1\sqrt{\rho/T}$. Finally, $\lambda\|\beta^*_A\|^2 \le \beta^{*\top}_A M_A\beta^*_A \le \mathbb{E}_\Sigma[h^2]$ for both graphs, hence

$$|q(\beta^*_A)| \le \frac{\mathbb{E}_\Sigma[h^2]}{\lambda}\|\Delta_M\|_{\mathrm{op}} + 2\sqrt{\frac{\mathbb{E}_\Sigma[h^2]}{\lambda}}\|\Delta_c\|,$$

which gives the bound. □

Since $\widehat{A}$ and $A^\perp$ have the same law given Σ, $\mathcal{D}_{\mathrm{dd}}$ isolates noise reuse, not a difference in graph quality. For $\mu_B < \delta < \mu_W$, both graphs estimate $R(A^{\mathrm{oracle}}) = P_C + \Delta R_\delta$. The subject-specific part ΔR_δ is not double dipping. What differs is that $R(\widehat{A}) - R(A^{\mathrm{oracle}})$ depends on the feature noise H, while $R(A^\perp) - R(A^{\mathrm{oracle}})$ does not. The bound shows that double dipping lives on edge flips: it is zero at $\delta = \pm\infty$, where both operators equal I_p or $P_{\mathbf{1}}$, and small whenever ρ is small. Its sign depends on $B^\star$, and community smoothness alone fixes neither its sign nor its size. When the bound is small, the two optimized errors are close: removing noise reuse changes little, and the shared graph-estimation error remains.

C Additional simulation results

In this section we report additional simulation experiments that further validate our theoretical findings and extend them to settings beyond the scope of the main analysis.

We begin by showing that the same conclusions hold when the feature map is the sample correlation $\widehat{C}$ rather than the sample covariance $\widehat{\Sigma}$, the more common choice in fMRI pipelines. We then provide ablations along the latent structure σ, the threshold δ, and the regression target β, isolating how each axis modulates the three effects identified in Section 3. Departing from the population regime ($n \to \infty$), we show that at small n even smoothing with the correct graph topology may fail to improve over raw features. We further demonstrate that the same statistical framework applies to GCN [5] without changing the qualitative ordering between feature maps. Finally, we extend the analysis from i.i.d. samples to autoregressive time series, which more faithfully reflect the temporal structure of real fMRI signals [34]; this departure preserves the framework and amounts to a shift along the T axis driven by the reduction in effective sample size.

C.1 Extension to correlation matrices: C

The theoretical analysis in Section 2 is stated for the sample covariance $\widehat{\Sigma}$, but fMRI pipelines almost always operate on the sample correlation $\widehat{C} = D^{-1/2}\widehat{\Sigma}D^{-1/2}$. Figure 5 reproduces the experiment of Figure 3 with $\widehat{C}$ as the feature map and confirms that the ordering between feature maps is preserved.

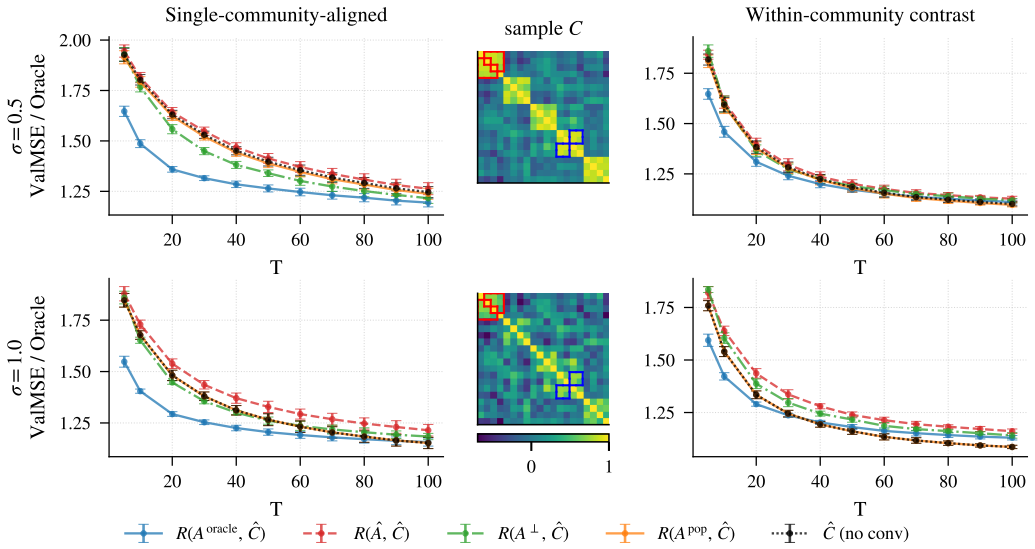


Figure 5: **Decoupling the effects for different values of β and latent structures, with C in place of Σ.** Center: true correlation C for two graph regimes ($\sigma \in \{0.5, 1.0\}$); overlays mark the support of $\beta =$ `single_community_aligned` (red, full community block) and $\beta =$ `within_community_contrast` (blue, single in-community edge). Left/right: normalized ValMSE vs. T for $\mathrm{SageMean}(A, \widehat{C})$ under $A \in \{A^{\mathrm{oracle}}, \widehat{A}, A^{\perp}, A^{\mathrm{pop}}\}$ and a no-conv baseline. Mean $\pm$ 1 SD over 5 seeds. $p = 20$, $K = 5$, $n = 25{,}000$; thresholds $\delta_{\sigma=0.5} \approx 0.79$, $\delta_{\sigma=1.0} \approx 0.53$.

C.2 The role of the coefficient: β

Fixing the latent structure ($\sigma = 1.0$) and a sparse threshold ($\delta \approx 10.20$), we vary the regression target β. Figure 6 shows that smoothing helps when β is community-smooth (`community_aligned`): the within-community averaging operator induces little bias and reduces variance, so A^{oracle} improves over raw features across all T. For the `within_community_contrast` and `edge_specific` targets, smoothing averages together precisely the entries whose differences carry the predictive signal, and the bias dominates the variance reduction except at very small T.

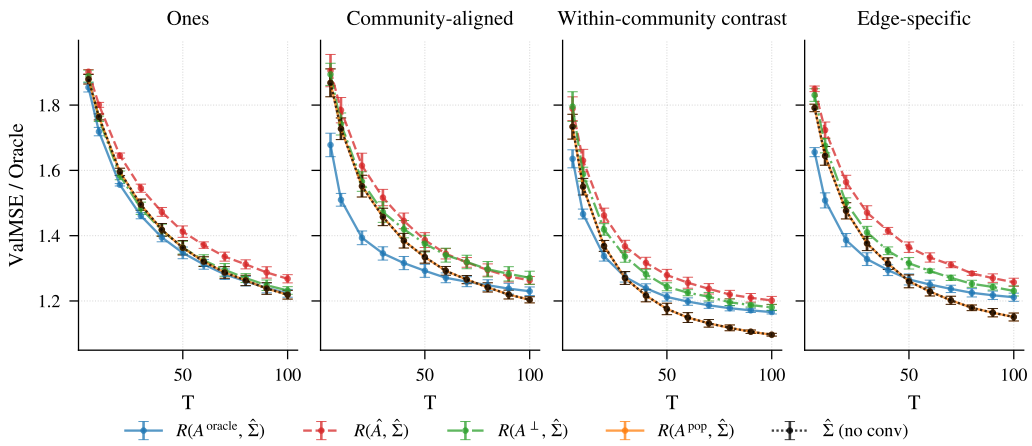


Figure 6: **Decoupling the effect of the regression target β.** Each panel fixes a different β pattern (`ones`, `community_aligned`, `within_community_contrast`, `edge_specific`) and plots normalized ValMSE vs. T for $\mathrm{SageMean}(A, \widehat{\Sigma})$ under $A \in \{A^{\text{oracle}}, \widehat{A}, A^{\perp}, A^{\text{pop}}\}$ and the no-conv baseline. Mean $\pm$ 1 SD over 5 seeds. $\sigma = 1.0$, $p = 20$, $K = 5$, $n = 25{,}000$, threshold $\delta \approx 10.20$.

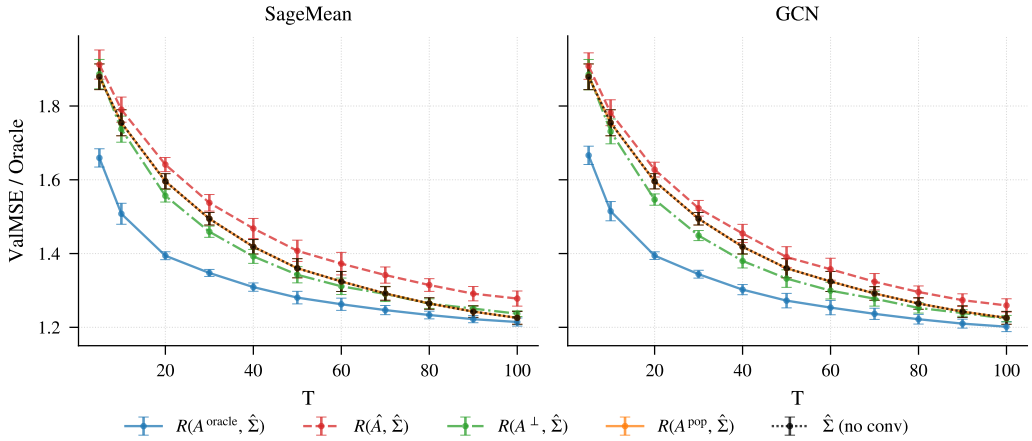


Figure 7: **Decoupling the effect of the convolution operator.** Each panel fixes a different graph convolution (SageMean, GCN) and plots normalized ValMSE vs. T for $\mathrm{Conv}(A, \widehat{\Sigma})$ under $A \in \{A^{\text{oracle}}, \widehat{A}, A^{\perp}, A^{\text{pop}}\}$ and the no-conv baseline. Mean $\pm$ 1 SD over 5 seeds. $\beta =$ `single_community_aligned`, $\sigma = 1.0$, $p = 20$, $K = 5$, $n = 25{,}000$, $k = 1$ convolution layer, threshold $\delta \approx 10.20$.

C.3 Different types of convolutions

The analysis in Section 3 uses the SageMean operator $R_{\text{SAGE}} = \widetilde{D}^{-1}\widetilde{A}$ [4]. We consider its linearized GCN counterpart $R_{\text{GCN}} = \widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}$ [5]. Figure 7 shows that SageMean and GCN behave near-identically: with balanced communities and a moderate δ, neighborhoods have roughly equal cardinality, so row- and symmetric-normalization coincide up to a constant. The qualitative ordering between A^{oracle}, $A^{\perp}$, $\widehat{A}$, and A^{pop} is preserved across both operators.

C.4 Temporally correlated samples

The analysis in Section 3 assumes i.i.d. samples $Z_t \sim \mathcal{N}(0, \Sigma)$, but real fMRI time series are temporally autocorrelated [34]. We replace the i.i.d. samples with AR(1) processes at $\rho \in \{0.4, 0.7\}$. For an AR(1) process the effective sample size is $T_{\text{eff}} \approx T(1-\rho^2)/(1+\rho^2)$, so $\widehat{\Sigma}$ is noisier at fixed T. Figure 8 confirms this: the curves for $\rho = 0.4$ and $\rho = 0.7$ are shifted to the right relative to the i.i.d. panel, while the relative ordering between A^{oracle}, $A^{\perp}$, $\widehat{A}$, and A^{pop} is preserved. Temporal correlation therefore rescales rather than reshapes the bias-variance tradeoff identified in the main analysis.

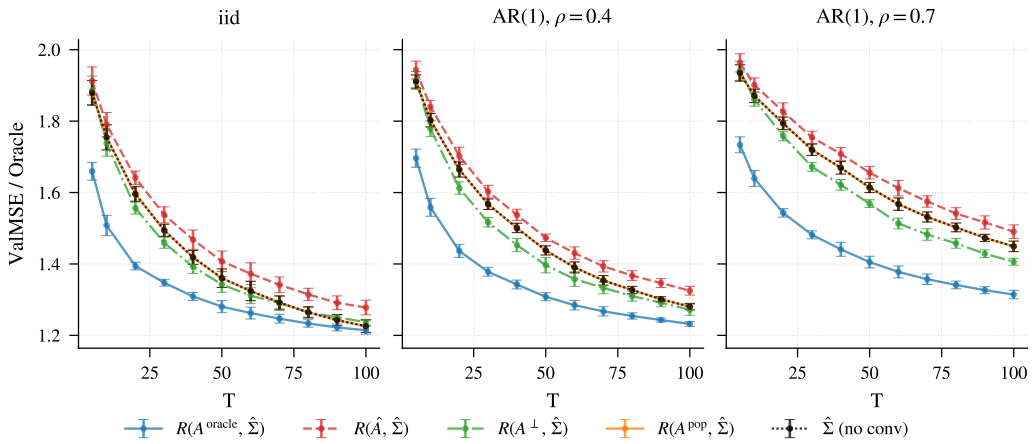


Figure 8: **Decoupling the effect of temporal correlation.** Each panel fixes a different temporal-noise regime for the time series used to estimate $\widehat{\Sigma}$, iid samples (left), AR(1) with $\rho = 0.4$ (center), AR(1) with $\rho = 0.7$ (right), and plots normalized ValMSE vs. T for $\mathrm{SageMean}(A, \widehat{\Sigma})$ under $A \in \{A^{\mathrm{oracle}}, \widehat{A}, A^{\perp}, A^{\mathrm{pop}}\}$ and the no-conv baseline. Mean $\pm$ 1 SD over 5 seeds. $\beta =$ `single_community_aligned`, $\sigma = 1.0$, $p = 20$, $K = 5$, $n = 25{,}000$, $k = 1$ convolution layer, threshold $\delta \approx 10.20$.

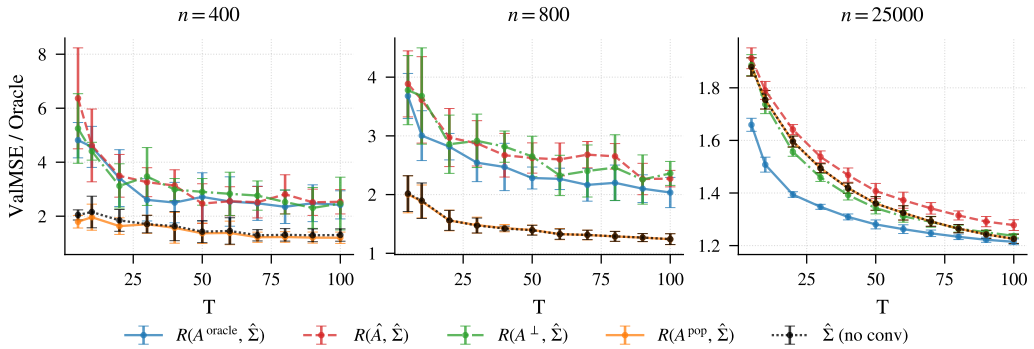


Figure 9: **Decoupling the effect of the sample size n.** Each panel fixes a different number of training samples $n \in \{400, 800, 25000\}$ and plots normalized ValMSE vs. T for $\mathrm{SageMean}(A, \widehat{\Sigma})$ under $A \in \{A^{\mathrm{oracle}}, \widehat{A}, A^{\perp}, A^{\mathrm{pop}}\}$ and the no-conv baseline. Mean $\pm$ 1 SD over 5 seeds. $\beta =$ `single_community_aligned`, $\sigma = 1.0$, $p = 20$, $K = 5$, threshold $\delta \approx 10.20$.

C.5 Small sample size

Our analysis restricts to the population regime $n \to \infty$. Figure 9 departs from this assumption and varies $n \in \{400, 800, 25{,}000\}$. At $n = 25{,}000$ the population predictions of the main analysis are recovered cleanly. At smaller n, finite-sample estimation variance becomes the dominant source of error and the ordering between feature maps is no longer well predicted by the EIV decomposition: smoothing with the correct topology no longer guarantees an improvement over raw features.

D Real data

D.1 Datasets

D.1.1 ADHD dataset

The ADHD-200 dataset is a multi-site collection of resting-state functional magnetic resonance imaging data targeting attention deficit hyperactivity disorder identification and cognitive assessment [35]. We utilized the preprocessed data provided by the Neuro Bureau ADHD-200 Preprocessed repository [36]. Specifically, the imaging data were processed using the Athena pipeline. Regional mean time series were extracted using the Craddock 200 (CC200) atlas [37], with 190 retained

regions and scan lengths ranging from 55 to 257 time points across 1,198 scans from 768 subjects. Acquisition parameters vary by site, with repetition times (TR) ranging from 1.96 to 2.5 seconds [36]. For our analysis, we focus on the following target variables:

- `Diagnosis`: The primary clinical label used for classification, distinguishing between individuals with attention deficit hyperactivity disorder and typical controls.
- `Inattentive`: The inattention symptom score reported in the ADHD-200 phenotypic file, used for regression.

D.1.2 ABIDE dataset

The Autism Brain Imaging Data Exchange (ABIDE) aggregates resting-state functional magnetic resonance imaging data from 17 international sites to investigate Autism Spectrum Disorder (ASD) [38]. We utilized the preprocessed data provided by the Preprocessed Connectomes Project (PCP) [39]. Specifically, imaging data were processed using the Configurable Pipeline for the Analysis of Connectomes (C-PAC) [40]. Based on our preprocessing selection, the signals underwent band-pass filtering and global signal regression. Regional mean time series were obtained via the Nilearn library [41]. The benchmark uses the Craddock 200 atlas [37] with 200 regions, while the controlled experiments use Harvard-Oxford with 106 retained regions. Acquisition parameters vary across the 17 sites, with the most common repetition time (TR) being 2.0 seconds and a typical scan duration of approximately 5–8 minutes per subject [38]. For our analysis, we focus on the following target variables:

- `DX_GROUP`: The primary diagnostic label used for binary classification, distinguishing between individuals with ASD and Typical Controls (TC).
- `AGE_AT_SCAN`: The age of the subject at the time of the scan, in years, used for regression.

D.1.3 Frankfurt dataset

The dataset consists of concurrent EEG-fMRI recordings from 72 healthy adults, comprising 45 females and 27 males with an age range of 19 to 48 years. Participants underwent 52-minute resting-state scans where they were instructed to remain still with their eyes closed. Despite these instructions, many participants transitioned through various stages of NREM sleep, with 64 subjects entering stage N1, 43 entering N2, and 17 reaching stage N3. The imaging data were acquired with a repetition time (TR) of 2080 ms [42]. The supplied regional time series contain 1,505 frames. Our loader removes the first five frames, leaving 1,500 frames aligned with the supplied hypnograms. For spatial definition, we utilized a parcellation consisting of 294 regions of interest, which included 286 cortical nodes and 8 subcortical nodes [43]. The benchmark uses all 294 regions. The controlled experiments average regional time series within the 15 groups specified by the supplied atlas mapping. For each multidimensional brain time series, an associated time series indicates the sleep stage at each time point, categorized as Wake (N0), N1, N2, and N3. A distinct characteristic of this dataset is that the target variable (y) is also time-dependent. Consequently, when applying the rolling window approach with a specific window length and stride, the target variable must be aggregated within the same temporal window. We construct the window-level targets separately for the two tasks:

- `sleep_frac`: The fraction of time points in the window spent asleep (N1 to N3), used for controlled regression.
- `sleep_wake`: The benchmark classification label indicates whether the most frequent original sleep stage in the window is sleep (N1 to N3) or wake (N0).

D.2 Benchmark experiments

D.2.1 Experiment set-up

For each scan, we use the supplied regional mean time series from the chosen parcellation and apply a rolling-window approach: ABIDE and Frankfurt use overlapping windows of length T with stride s, while ADHD uses the full series as one window. Within each window we compute the sample correlation matrix $\widehat{C} \in \mathbb{R}^{p \times p}$, where p is the number of regions. These correlation matrices serve as input features for all downstream models, used both as node features and as the basis for graph

construction in the GNN baselines. We report window-level test metrics, with means and standard deviations over 10 random seeds.

Dataset-specific settings. Table 3 reports the parcellation, number of ROIs, window length T, stride, and number of subjects for each dataset.

Table 3: Dataset-specific experimental settings. Subjects are those with the target available, split 60/20/20 by subject. For ABIDE and Frankfurt, the stride is set to $\lfloor 0.10 \cdot T \rfloor$; ABIDE uses 4 windows per subject and Frankfurt all 66. For ADHD, a single window is used, corresponding to the maximum available time-series length. Targets: age at scan (ABIDE), diagnosis (ADHD), sleep versus wake per window (Frankfurt).

Dataset	Task	Parcellation	# ROIs	T	Stride	Subjects
ABIDE	Regression	`rois_cc200`	200	150	15	785
ADHD	Classification	`cc200`	190	full TS	N/A	768
Frankfurt	Classification	Gordon286 + ASEG8	294	200	20	72

Model hyperparameters. Table 4 reports the most frequently selected hyperparameters for each model across benchmark runs. Hyperparameters are tuned separately for each model, dataset–task panel, and seed.

Table 4: Most frequently selected hyperparameters across benchmark runs. All neural models were trained with 10 seeds per dataset–task panel; GAT uses a single attention head. Graph density is $\tau = 0.10$ for all graph-based benchmark models.

Model	Hyperparameters
Random Forest	`n_estimators=400`, `max_depth=None`, `min_samples_leaf=5`, `max_features=sqrt`
Ridge	`alpha=100`
SVM	`C=1`, `gamma=scale`, `kernel=rbf`
SVR	`C=1`, `gamma=scale`, `kernel=rbf`
MLP	`lr=1e-2`, `weight_decay=1e-3`, `l1_lambda=1e-4`, `hidden_dim=64`
GAT (1-conv)	`lr=1e-2`, `weight_decay=0`, `l1_lambda=1e-4`, `hidden_dim=64`
GAT (2-conv)	`lr=1e-2`, `weight_decay=1e-4`, `l1_lambda=1e-5`, `hidden_dim=64`
GCN (1-conv)	`lr=1e-2`, `weight_decay=0`, `l1_lambda=1e-4`, `hidden_dim=128`
GCN (2-conv)	`lr=1e-2`, `weight_decay=1e-4`, `l1_lambda=1e-5`, `hidden_dim=256`
SAGE (1-conv)	`lr=1e-2`, `weight_decay=0`, `l1_lambda=0`, `hidden_dim=256`
SAGE (2-conv)	`lr=1e-2`, `weight_decay=0`, `l1_lambda=1e-5`, `hidden_dim=256`
ResidualGNN	`lr=1e-2`, `weight_decay=1e-3`, `l1_lambda=1e-4`, `hidden_channels=128`, `hidden=128`, `num_layers=3`, `gnn_type=SAGEConv`

D.3 Controlled experiments

We provide further details of the controlled experiments introduced in Section 5, followed by the full results for the progressive relaxation of assumptions, the comparisons of convolution operators and graph densities on ABIDE, and the extensions to the ADHD and Frankfurt datasets. Throughout, the independent graph $A^{\perp}_{\text{est}}$ is built from the furthest non-overlapping window of the same subject, to limit the effect of autocorrelation.

D.3.1 Progressive relaxation of assumptions

Starting from the configuration in the main text, we first introduce ReLU while retaining the vectorized readout and a fixed number of windows across T. The first block of Table 5 shows that the oracle proxy remains the strongest graph construction at every T. Its mean accuracy exceeds that of the no-convolution baseline at $T = 5, 15, 25$, but none of these gains reaches significance at the 5% level. At $T = 40$, even the oracle proxy falls below the baseline. Thus, the broad ordering persists after introducing the nonlinearity, although the evidence for a smoothing benefit is weaker. Neither estimated graph significantly improves on the baseline.

We next retain all available windows, allowing n to vary with T. The no-convolution baseline improves at shorter window lengths: at $T = 5$, its accuracy rises from 0.558 to 0.593. The oracle advantage is now confined to this shortest window, where it reaches $+0.037$ accuracy ($p < 0.01$); at $T \geq 15$, the oracle proxy no longer improves on the baseline. The remaining advantage is substantially reduced when the graph must be estimated from a separate window: at $T = 5$, this replacement lowers accuracy by 0.030 ($p < 0.05$). Neither the separate-window nor the same-window graph significantly outperforms the no-convolution baseline.

Finally, replacing vectorization with mean pooling removes the remaining oracle advantage at $T = 5$: the oracle-minus-baseline difference changes from $+0.037$ to -0.003. In this final configuration, no graph construction significantly improves on the corresponding no-convolution baseline at any tested T. For ABIDE, the benefit observed under the initial restrictions therefore does not survive the full progression toward the deployed pipeline, connecting the controlled experiments to the benchmark findings.

Table 5: **Progressive relaxation of the assumptions of Table 2.** Same cell and conventions, SageMean, $\tau = 0.10$; each block adds one relaxation to the previous one, starting from Table 2. Mean $\pm$ 1 SD over 20 seeds; paired t-test over seeds against $\widehat{C}$ (left) or zero (right): $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

T	$\widehat{C}$	$A^{\text{oracle}}_{\text{est}}$	$A^{\perp}_{\text{est}}$	$\widehat{A}$	A^{pop}	$A^{\text{oracle}}_{\text{est}} - \widehat{C}$	$A^{\perp}_{\text{est}} - A^{\text{oracle}}_{\text{est}}$	$\widehat{A} - A^{\perp}_{\text{est}}$	$A^{\text{pop}} - \widehat{C}$
ReLU activation (vectorized readout, n fixed)									
5	$0.558_{\pm.037}$	$0.567_{\pm.044}$	$0.537^{**}_{\pm.037}$	$0.546_{\pm.038}$	$0.550_{\pm.031}$	$+0.009$	-0.029^{***}	$+0.009$	-0.008
15	$0.617_{\pm.051}$	$0.632_{\pm.044}$	$0.615_{\pm.051}$	$0.620_{\pm.045}$	$0.609_{\pm.051}$	$+0.015$	-0.017^{*}	$+0.005$	-0.009
25	$0.643_{\pm.045}$	$0.660^{*}_{\pm.032}$	$0.641_{\pm.048}$	$0.653_{\pm.030}$	$0.640_{\pm.046}$	$+0.017^{*}$	-0.019^{*}	$+0.012$	-0.003
40	$0.680_{\pm.035}$	$0.676_{\pm.031}$	$0.666^{**}_{\pm.033}$	$0.662^{***}_{\pm.031}$	$0.656^{**}_{\pm.046}$	-0.005	-0.009^{**}	-0.004	-0.024^{**}
ReLU activation and n free (vectorized readout)									
5	$0.593_{\pm.057}$	$0.630^{***}_{\pm.044}$	$0.600_{\pm.064}$	$0.601_{\pm.056}$	$0.601_{\pm.059}$	$+0.037^{***}$	-0.030^{**}	$+0.002$	$+0.008$
15	$0.669_{\pm.036}$	$0.667_{\pm.034}$	$0.663_{\pm.034}$	$0.651^{**}_{\pm.039}$	$0.653^{***}_{\pm.033}$	-0.002	-0.004	-0.013	-0.016^{***}
25	$0.681_{\pm.036}$	$0.680_{\pm.040}$	$0.673_{\pm.039}$	$0.676_{\pm.041}$	$0.682_{\pm.037}$	-0.001	-0.007	$+0.003$	$+0.001$
40	$0.676_{\pm.031}$	$0.672_{\pm.036}$	$0.670_{\pm.036}$	$0.671_{\pm.040}$	$0.667_{\pm.035}$	-0.004	-0.003	$+0.002$	-0.009
ReLU activation, n free and mean pooling (deployed pipeline)									
5	$0.623_{\pm.038}$	$0.620_{\pm.049}$	$0.627_{\pm.050}$	$0.626_{\pm.051}$	$0.627_{\pm.045}$	-0.003	$+0.007$	-0.001	$+0.004$
15	$0.672_{\pm.038}$	$0.657^{**}_{\pm.042}$	$0.652^{***}_{\pm.042}$	$0.659^{**}_{\pm.042}$	$0.657^{**}_{\pm.042}$	-0.015^{**}	-0.005	$+0.007^{**}$	-0.015^{**}
25	$0.666_{\pm.032}$	$0.663_{\pm.032}$	$0.664_{\pm.033}$	$0.667_{\pm.033}$	$0.666_{\pm.030}$	-0.003	$+0.000$	$+0.003$	$+0.000$
40	$0.668_{\pm.041}$	$0.664_{\pm.035}$	$0.658^{*}_{\pm.033}$	$0.660_{\pm.027}$	$0.661^{**}_{\pm.032}$	-0.004	-0.007	$+0.003$	-0.007^{**}

D.3.2 Extensions to other convolutions and datasets

GAT lies outside our linear theory, because its attention weights are computed from the same noisy matrix as the features and therefore inherit its noise. GAT follows the same qualitative pattern as SageMean: in the simulation-like setting, the oracle proxy helps at short windows, but neither estimated graph significantly outperforms its matched no-convolution baseline. In the deployed setting, neither operator benefits significantly from any graph construction.

Table 6: **GAT in place of** SageMean**.** Same cell and conventions as Table 2, $\tau = 0.10$, with one GAT convolution (8 heads of 8); the $\widehat{C}$ column is the same architecture with message passing removed. Mean $\pm$ 1 SD over 20 seeds; paired t-test over seeds against $\widehat{C}$ (left) or zero (right): $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

T	$\widehat{C}$	$A^{\text{oracle}}_{\text{est}}$	$A^{\perp}_{\text{est}}$	$\widehat{A}$	A^{pop}	$A^{\text{oracle}}_{\text{est}} - \widehat{C}$	$A^{\perp}_{\text{est}} - A^{\text{oracle}}_{\text{est}}$	$\widehat{A} - A^{\perp}_{\text{est}}$	$A^{\text{pop}} - \widehat{C}$
Simulation-like scenario (identity activation, vectorized readout, n fixed)									
5	$0.558_{\pm.033}$	$0.585^{***}_{\pm.031}$	$0.547_{\pm.037}$	$0.557_{\pm.031}$	$0.554_{\pm.034}$	$+0.027^{***}$	-0.038^{***}	$+0.010$	-0.004
15	$0.621_{\pm.040}$	$0.643^{***}_{\pm.035}$	$0.608^{*}_{\pm.032}$	$0.618_{\pm.038}$	$0.605^{***}_{\pm.032}$	$+0.022^{***}$	-0.035^{***}	$+0.010$	-0.016^{***}
25	$0.649_{\pm.037}$	$0.654_{\pm.033}$	$0.637^{*}_{\pm.035}$	$0.643_{\pm.036}$	$0.638^{**}_{\pm.032}$	$+0.005$	-0.017^{**}	$+0.006$	-0.011^{**}
40	$0.680_{\pm.047}$	$0.663^{**}_{\pm.037}$	$0.662^{***}_{\pm.035}$	$0.665^{***}_{\pm.038}$	$0.654^{***}_{\pm.035}$	-0.018^{**}	-0.001	$+0.003$	-0.027^{***}
Deployed pipeline (ReLU activation, mean pooling, n free)									
5	$0.633_{\pm.049}$	$0.616^{**}_{\pm.036}$	$0.620^{*}_{\pm.037}$	$0.631_{\pm.048}$	$0.615^{**}_{\pm.042}$	-0.017^{**}	$+0.004$	$+0.011$	-0.018^{**}
15	$0.677_{\pm.036}$	$0.665^{**}_{\pm.033}$	$0.662^{**}_{\pm.031}$	$0.664^{**}_{\pm.036}$	$0.667^{*}_{\pm.035}$	-0.012^{**}	-0.003	$+0.003$	-0.010^{*}
25	$0.670_{\pm.024}$	$0.658^{**}_{\pm.026}$	$0.662^{*}_{\pm.023}$	$0.663^{**}_{\pm.024}$	$0.668_{\pm.026}$	-0.012^{**}	$+0.004$	$+0.001$	-0.002
40	$0.670_{\pm.031}$	$0.659^{**}_{\pm.031}$	$0.659^{**}_{\pm.028}$	$0.660^{**}_{\pm.035}$	$0.661^{*}_{\pm.031}$	-0.011^{**}	$+0.000$	$+0.000$	-0.009^{*}

On ADHD, the oracle proxy provides a modest benefit at $T = 7$, which is lost when the graph is estimated from a separate window; at $T = 40$, even the oracle proxy falls below the baseline. In the deployed setting, no subject-specific graph significantly improves prediction, although the population graph provides a small gain at $T = 40$.

Table 7: **ADHD.** Subject-level test R^2 for the inattention score on ADHD-200 (CC200, $p = 190$) with $\mathrm{SageMean}(A, \widehat{C})$, one convolution, $\tau = 0.10$, in the simulation-like scenario and the deployed pipeline. Mean $\pm$ 1 SD over 20 seeds; paired t-test over seeds against $\widehat{C}$ (left) or zero (right): $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The train-mean predictor has $R^2 = -0.03$.

T	$\widehat{C}$	$A^{\text{oracle}}_{\text{est}}$	$A^{\perp}_{\text{est}}$	$\widehat{A}$	A^{pop}	$A^{\text{oracle}}_{\text{est}} - \widehat{C}$	$A^{\perp}_{\text{est}} - A^{\text{oracle}}_{\text{est}}$	$\widehat{A} - A^{\perp}_{\text{est}}$	$A^{\text{pop}} - \widehat{C}$
Simulation-like scenario (identity activation, vectorized readout, n fixed)									
5	$0.243_{\pm.060}$	$0.261_{\pm.063}$	$0.219^{*}_{\pm.069}$	$0.225_{\pm.063}$	$0.251_{\pm.057}$	$+0.018$	-0.042^{***}	$+0.006$	$+0.008$
7	$0.318_{\pm.047}$	$0.333^{**}_{\pm.057}$	$0.267^{***}_{\pm.061}$	$0.312_{\pm.058}$	$0.320_{\pm.055}$	$+0.015^{**}$	-0.066^{***}	$+0.045^{***}$	$+0.002$
15	$0.463_{\pm.043}$	$0.469_{\pm.057}$	$0.443^{**}_{\pm.051}$	$0.448^{*}_{\pm.053}$	$0.466_{\pm.059}$	$+0.005$	-0.026^{***}	$+0.006$	$+0.002$
25	$0.531_{\pm.035}$	$0.526_{\pm.043}$	$0.511^{***}_{\pm.043}$	$0.516^{**}_{\pm.040}$	$0.532_{\pm.046}$	-0.006	-0.015	$+0.005$	$+0.001$
40	$0.543_{\pm.063}$	$0.523^{***}_{\pm.058}$	$0.526^{*}_{\pm.051}$	$0.525^{**}_{\pm.054}$	$0.529^{*}_{\pm.050}$	-0.020^{***}	$+0.003$	-0.002	-0.014^{*}
Deployed pipeline (ReLU activation, mean pooling, n free)									
5	$0.289_{\pm.027}$	$0.276^{**}_{\pm.027}$	$0.280^{***}_{\pm.025}$	$0.273^{***}_{\pm.027}$	$0.280^{*}_{\pm.026}$	-0.013^{**}	$+0.003$	-0.007	-0.009^{*}
7	$0.329_{\pm.032}$	$0.319^{*}_{\pm.033}$	$0.328_{\pm.033}$	$0.314^{***}_{\pm.031}$	$0.331_{\pm.032}$	-0.009^{*}	$+0.008^{*}$	-0.014^{***}	$+0.002$
15	$0.411_{\pm.036}$	$0.398^{**}_{\pm.044}$	$0.401^{*}_{\pm.043}$	$0.382^{***}_{\pm.044}$	$0.416_{\pm.040}$	-0.014^{**}	$+0.004$	-0.020^{***}	$+0.005$
25	$0.450_{\pm.055}$	$0.425^{***}_{\pm.054}$	$0.424^{**}_{\pm.061}$	$0.423^{***}_{\pm.054}$	$0.444_{\pm.055}$	-0.025^{***}	-0.001	-0.001	-0.006
40	$0.432_{\pm.062}$	$0.428_{\pm.050}$	$0.411_{\pm.074}$	$0.421_{\pm.046}$	$0.452^{**}_{\pm.050}$	-0.004	-0.018	$+0.011$	$+0.019^{**}$

Frankfurt differs because sleep state changes within a scan: changing T changes both connectivity-estimation noise and the window-level sleep fraction being predicted. The full-series graph also averages across sleep states and may therefore poorly represent an individual window. In the simulation-like setting, the oracle proxy is strongest at the shortest windows but weakest at $T = 45$, while the separate-window graph performs best from $T = 14$ onward. Graph benefits persist in the deployed setting, with all four constructions significantly improving prediction at $T = 25$ and 45.

Table 8: **Frankfurt.** Window-level test R^2 for the fraction of the window spent asleep on the Frankfurt dataset (15 networks) with $\mathrm{SageMean}(A, \widehat{C})$, one convolution, $\tau = 0.10$, in the simulation-like scenario and the deployed pipeline. The target is defined per window, so $A^{\text{oracle}}_{\text{est}}$ is built from a series whose label changes within the scan. Mean $\pm$ 1 SD over 20 seeds; paired t-test over seeds against $\widehat{C}$ (left) or zero (right): $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The train-mean predictor has $R^2 = -0.03$.

T	$\widehat{C}$	$A^{\text{oracle}}_{\text{est}}$	$A^{\perp}_{\text{est}}$	$\widehat{A}$	A^{pop}	$A^{\text{oracle}}_{\text{est}} - \widehat{C}$	$A^{\perp}_{\text{est}} - A^{\text{oracle}}_{\text{est}}$	$\widehat{A} - A^{\perp}_{\text{est}}$	$A^{\text{pop}} - \widehat{C}$
Simulation-like scenario (identity activation, vectorized readout, n fixed)									
5	$0.069_{\pm.067}$	$0.126^{***}_{\pm.090}$	$0.066_{\pm.072}$	$0.060^{***}_{\pm.068}$	$0.066_{\pm.071}$	$+0.057^{***}$	-0.060^{***}	-0.007	-0.003
8	$0.141_{\pm.052}$	$0.193^{***}_{\pm.070}$	$0.145_{\pm.059}$	$0.144_{\pm.056}$	$0.145_{\pm.047}$	$+0.052^{***}$	-0.048^{***}	-0.002	$+0.005$
14	$0.246_{\pm.073}$	$0.262_{\pm.082}$	$0.271^{***}_{\pm.070}$	$0.252^{**}_{\pm.073}$	$0.249_{\pm.073}$	$+0.016$	$+0.009$	-0.018^{***}	$+0.004$
25	$0.374_{\pm.063}$	$0.369_{\pm.085}$	$0.398^{***}_{\pm.066}$	$0.381^{*}_{\pm.064}$	$0.379^{**}_{\pm.064}$	-0.004	$+0.029^{**}$	-0.017^{***}	$+0.005^{**}$
45	$0.461_{\pm.065}$	$0.445_{\pm.083}$	$0.488^{***}_{\pm.076}$	$0.461_{\pm.071}$	$0.477^{***}_{\pm.066}$	-0.016	$+0.044^{***}$	-0.028^{***}	$+0.016^{***}$
Deployed pipeline (ReLU activation, mean pooling, n free)									
5	$0.051_{\pm.049}$	$0.063^{***}_{\pm.054}$	$0.054_{\pm.058}$	$0.058^{**}_{\pm.051}$	$0.058^{**}_{\pm.049}$	$+0.012^{***}$	-0.009^{***}	$+0.004$	$+0.007^{**}$
8	$0.096_{\pm.053}$	$0.104^{***}_{\pm.054}$	$0.098_{\pm.057}$	$0.097_{\pm.059}$	$0.097_{\pm.058}$	$+0.009^{***}$	-0.006^{**}	-0.001	$+0.002$
14	$0.165_{\pm.047}$	$0.181^{***}_{\pm.061}$	$0.178^{**}_{\pm.068}$	$0.174^{*}_{\pm.061}$	$0.171_{\pm.064}$	$+0.016^{***}$	-0.003	-0.004	$+0.006$
25	$0.246_{\pm.063}$	$0.276^{***}_{\pm.062}$	$0.283^{***}_{\pm.067}$	$0.272^{***}_{\pm.059}$	$0.270^{***}_{\pm.059}$	$+0.029^{***}$	$+0.007^{**}$	-0.011^{***}	$+0.024^{***}$
45	$0.316_{\pm.063}$	$0.344^{***}_{\pm.064}$	$0.356^{***}_{\pm.066}$	$0.347^{***}_{\pm.060}$	$0.340^{***}_{\pm.063}$	$+0.029^{***}$	$+0.012^{***}$	-0.009^{**}	$+0.025^{***}$

D.3.3 Effect of graph density

We repeat the ABIDE simulation-like experiment at graph densities $\tau = 0.05$ and 0.20, comparing them with $\tau = 0.10$ in the main text. At $T = 5$, the oracle proxy significantly improves accuracy at all three densities, with gains of 0.019, 0.022, and 0.024 for $\tau = 0.05$, 0.10, and 0.20, respectively. However, neither the separate-window nor the same-window graph significantly outperforms the no-convolution baseline at any tested density or T. The oracle benefit also varies with density: at $T = 25$, it is significant only for $\tau = 0.10$, while at $T = 40$ its mean performance falls below the baseline at every density. Thus, making the graph sparser or denser does not consistently recover a predictive advantage in this setting.

Table 9: **Threshold ablation of Table 2.** Same cell and conventions, $\mathrm{SageMean}$, simulation-like scenario, at $\tau = 0.05$ and $\tau = 0.20$ ($\tau = 0.10$ is Table 2); the $\widehat{C}$ column does not depend on τ. Mean $\pm$ 1 SD over 20 seeds; paired t-test over seeds against $\widehat{C}$ (left) or zero (right): $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

T	$\widehat{C}$	$A_{\mathrm{est}}^{\mathrm{oracle}}$	$A_{\mathrm{est}}^{\perp}$	$\widehat{A}$	A^{pop}	$A_{\mathrm{est}}^{\mathrm{oracle}} - \widehat{C}$	$A_{\mathrm{est}}^{\perp} - A_{\mathrm{est}}^{\mathrm{oracle}}$	$\widehat{A} - A_{\mathrm{est}}^{\perp}$	$A^{\mathrm{pop}} - \widehat{C}$
$\tau = 0.05$									
5	$0.560_{\pm.033}$	$0.580^{**}_{\pm.034}$	$0.563_{\pm.033}$	$0.563_{\pm.033}$	$0.567_{\pm.035}$	$+0.019^{**}$	-0.016^{**}	-0.000	$+0.006$
15	$0.625_{\pm.040}$	$0.628_{\pm.035}$	$0.616_{\pm.039}$	$0.616_{\pm.039}$	$0.623_{\pm.040}$	$+0.003$	-0.013^{**}	$+0.001$	-0.003
25	$0.651_{\pm.037}$	$0.654_{\pm.043}$	$0.649_{\pm.051}$	$0.651_{\pm.045}$	$0.652_{\pm.045}$	$+0.003$	-0.005	$+0.002$	$+0.001$
40	$0.682_{\pm.034}$	$0.661^{**}_{\pm.030}$	$0.668^{**}_{\pm.033}$	$0.660^{**}_{\pm.035}$	$0.668_{\pm.035}$	-0.021^{**}	$+0.007$	-0.008	-0.014
$\tau = 0.20$									
5	$0.560_{\pm.033}$	$0.584^{***}_{\pm.036}$	$0.566_{\pm.033}$	$0.558_{\pm.037}$	$0.567_{\pm.038}$	$+0.024^{***}$	-0.019^{***}	-0.008	$+0.006$
15	$0.625_{\pm.040}$	$0.629_{\pm.033}$	$0.622_{\pm.041}$	$0.623_{\pm.037}$	$0.617_{\pm.034}$	$+0.004$	-0.008^{*}	$+0.001$	-0.008
25	$0.651_{\pm.037}$	$0.661_{\pm.040}$	$0.644_{\pm.049}$	$0.651_{\pm.045}$	$0.653_{\pm.035}$	$+0.009$	-0.017^{**}	$+0.007$	$+0.002$
40	$0.682_{\pm.034}$	$0.676_{\pm.031}$	$0.671^{*}_{\pm.034}$	$0.664^{**}_{\pm.036}$	$0.675_{\pm.032}$	-0.006	-0.005	-0.008	-0.007

D.4 Practical recommendations

Our results suggest a few guidelines for connectome GNN pipelines. First, convolutions help only at short windows: when longer recordings are available, longer windows combined with graph-agnostic models are likely to perform better, so convolutions should be used with care. Second, estimating the graph from a longer series than the window, as the oracle proxy does, extends the range of T over which convolutions help, provided the target does not change within the scan. Third, building the graph from the same window as the features adds a double-dipping term of no consistent sign, which makes results less stable rather than systematically worse; cross-fitting removes this term but not the graph-estimation error, which is the main bottleneck. Fourth, smoothing can only help when the target is community-smooth: if the signal sits in single edges or in contrasts within a community, smoothing removes it. Finally, when these conditions are met, the graph density must be tuned; otherwise, no choice of density recovers a gain.

E Compute resources

All experiments were run on CPU nodes equipped with Intel Broadwell (28 cores at 2.4 GHz, 64 GB memory) and Intel Skylake (40 cores at 2.4 GHz, 96 GB memory) processors, for a total of approximately 3,000 CPU hours.

F Broader impacts

Our work is a theoretical and methodological analysis of when graph neural network methods provide statistical benefits over simpler baselines in brain connectome analysis. By clarifying the regimes in which these methods help versus hurt, we hope to inform more reliable use of machine learning in clinical neuroscience applications such as the diagnosis of psychiatric and neurological disorders. We do not introduce new models, datasets, or deployment ready systems, and we do not foresee direct negative societal impacts of this work.

G LLM usage declaration

We used large language models (LLMs) as coding assistants to help implement and run some of the experimental pipelines, and as writing assistants to copy edit and improve the clarity of the manuscript.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer:[Yes]

   Justification: The abstract and introduction correctly identify our three main contributions: (1) decomposing the effect of graph convolutions into three separate mechanisms, (2) characterizing the scenarios in which each mechanism is beneficial, and (3) validating these claims through rigorous simulations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Conclusion (Section 6) explicitly discusses the limitations of our analysis and outlines directions for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical results are presented in a discursive form in the main body (Sections 2–3), with each proposition stated together with its assumptions. More proofs in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the full code and hyperparameters for both the simulations and the real-data analyses.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All experimental details are provided in the Appendix. The simulations are fully reproducible from the released code. For the real-data analyses, we provide the full pipeline and hyperparameters. Regarding data access: the ADHD-200 dataset is open-source, ABIDE requires registration that is automatically granted, and the Frankfurt EEG-fMRI dataset requires access through an explicit request to the data owners.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and test details, including data splits, hyperparameters, and optimizer settings, are specified in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All results are averaged over random seeds, 5 in the simulations and 10 or 20 in the real-data analyses, as stated in each caption. Error bars in all plots correspond to ± 1 standard deviation.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources used for all experiments are specified in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: A short discussion of broader impacts is provided in the Appendix.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets used in the paper are properly credited to their original creators: ADHD-200 [35, 36], ABIDE [38, 39], and the Frankfurt EEG-fMRI dataset [42]. The license and access terms of each dataset are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: A declaration of LLM usage is provided in the Appendix. LLMs were used to assist with running some of the experiments and with editing the manuscript.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.