Connectopy

Riley Harper Sean Shen Yinyu Yao
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
{riley.harper,sjshen,yyao2}@unc.edu

Abstract

We study how structural and functional brain connectivity jointly relate to cognition and lifetime alcohol abuse, and whether these relationships differ by sex. Using data from the Human Connectome Project, we combine diffusion MRI derived structural connectomes, resting state functional connectomes, and a rich set of cognitive, substance use, and socio-economic traits. Each 68×68 connectome is reduced to low-dimensional features using three approaches: tensor network PCA (TN-PCA), vector PCA, and variational autoencoders (VAEs), yielding 60 structural and 60 functional components per method. We then model a binarized lifetime alcohol abuse diagnosis separately in men ($n = 482$) and women ($n = 562$) with Explainable Boosting Machines and random forests, comparing cognition-only, connectome-only, and combined feature sets. Cognitive-only EBMs already achieve modest discrimination (AUC ≈ 0.63 in men, 0.58 in women). Adding PCA-based connectome features improves test AUCs by only 0.03–0.07, and most connectome-only models perform near chance, with the exception of a male PCA-only EBM (AUC ≈ 0.67). Global importance and shape plots indicate that age and higher-level cognitive performance are the dominant predictors, with connectome components contributing many small, distributed effects. Thus, in this cohort, alcohol abuse is only weakly predictable, and most of the available signal resides in behavioural measures rather than brain connectivity. All analyses are implemented in an open source, containerized pipeline, available at `https://github.com/Sean0418/connectopy`, to facilitate reproducibility.

1 Introduction

Understanding how patterns of brain connectivity relate to individual differences in cognition and behavior is a central goal of modern neuroscience. Large scale initiatives such as the Human Connectome Project (HCP) have made it possible to study this question at scale, by providing high-quality diffusion MRI (dMRI), resting-state fMRI (rs-fMRI), and extensive behavioral phenotyping in the same participants [Van Essen et al., 2013]. However, translating high dimensional connectome data into interpretable predictors of behavior remains an open question.

Each structural or functional connectome is represented as a 68×68 network, when using the Desikan-Killiany cortical atlas, yielding $2{,}278$ unique undirected edges per subject. For the HCP sample considered here, this implies millions of edge-level measurements across more than one thousand participants. A key methodological question is therefore how to obtain low-dimensional yet informative representations of individual connectomes that can be related to behavioral traits. In this project we analyze a subset of HCP data comprising diffusion-based structural connectomes, resting-state functional connectomes, and 175 behavioral and demographic traits for approximately $1{,}065$ subjects. For each subject we have a 68×68 structural connectivity matrix summarizing white-matter streamline counts between cortical regions, and a 68×68 functional connectivity matrix

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

summarizing pairwise correlations between rs-fMRI time series. In addition, we have a collection of 175 traits spanning cognition, substance use, psychopathology, personality, and physical health. Our goal is to build a coherent pipeline that integrates these data, reduces the dimensionality of the connectomes using tensor network principal component analysis (TN-PCA) Zhang et al. [2019], vector PCA Jolliffe and Cadima [2016], and variational autoencoders Kingma and Welling [2022] to investigate how connectivity-derived features relate to cognitive abilities and alcohol abuse.

1.1 Problem Statement

The central scientific question we address is:

> *How do structural and functional brain networks mediate the relationship between cognitive traits and alcohol abuse differently across sexes?*

To answer this question, we must merge structural connectomes, functional connectomes, low dimension representations of the original data, and behavioral traits across subjects. Using the TN-PCA scores and other dimension reduction techniques as connectome-derived features, we study how connectivity patterns relate alcohol abuse using predictive models such as Support Vector Machines (SVM) Cortes and Vapnik [1995], Random Forests (RF) Breiman [2001], Explainable Boosting Machines (EBM) Nori et al. [2019], and logistic regression Cox [1958]. Recognizing the biological differences present in the sample, we stratified our analyses by sex. This allowed us to uncover distinct neuro-cognitive pathways and biomarker efficacy for males versus females, rather than treating gender merely as a covariate.

1.2 Overview of the HCP Dataset

We analyze connectome and behavioral data from a subset of the Human Connectome Project (HCP) S1200 release [Van Essen et al., 2013]. After applying the preprocessing pipeline provided by the original authors of the TN-PCA code, we obtained:

We obtained diffusion MRI-derived structural connectivity matrices for 1,065 subjects. Each 68×68 symmetric matrix ($\mathbf{S}_i$) represents streamline counts between Desikan-Killiany regions. The matrices are sparse (approx. 26.5% non-zero entries) with a mean edge density of 0.537 and highly variable edge weights (mean 67.1, SD 314.3). Resting-state functional connectivity matrices were available for 1,058 subjects. After removing empty entries, we obtained a numeric array of 68×68 correlation matrices. Values range from -0.58 to 0.99 (mean 0.302), reflecting standard resting-state correlations. Pre-computed TN-PCA coefficients were extracted for both modalities. Structural data yielded a $1,065 \times 60$ matrix, and functional data yielded a $1,058 \times 60$ matrix. All components are standardized (mean ≈ 0, SD ≈ 0.031). A $175 \times N$ matrix of behavioral traits covers cognition, substance use, and health. We mapped these to informative variable names using the provided metadata file.

2 Methods

2.1 Data Integration

We integrate structural connectomes, functional connectomes, TN-PCA scores, and behavioral traits at the subject level. We use MATLAB for initial processing and sanity checks, while Python and R are used for downstream statistical analysis.

The structural connectome file contains 1,065 subjects with 68×68 connectivity matrices. We processed the functional connectome file by removing empty entries which resulted in a final array of 1,058 subjects. We confirmed that all subjects in the functional dataset are present within the structural dataset. The processed structural TN-PCA scores ($1,065 \times 60$) were loaded from `TNPCA_Coeff_HCP_Structural_Connectome.mat`. The processed functional TN-PCA scores ($1,058 \times 60$) were loaded from `TNPCA_Coeff_HCP_Functional_Connectome.mat`. Subject IDs were verified. Both matrices were saved as CSV files with component columns (`tn_sc_pc1`–`tn_sc_pc60`, `tn_fc_pc1`–tn_fc_pc60) and a subject-ID column. Processed behavioral and demographic traits were loaded from `HCP_175Traits.mat`. This was transposed to an $N \times 175$ matrix, where rows are subjects and columns are traits. Using the metadata file `Details_175_Traits.xlsx`, the trait columns were renamed from their HCP variable names.

The final dataset was compiled by merging four subject-level feature sets: TN-PCA, traditional (vector) PCA, VAE latent scores, and behavioral traits. These were all derived from the original connectome data via dimension reduction. Using a custom data loader, the sets were joined by subject ID to create a master data file, saved as `full_data.csv`.

2.2 Tensor Network PCA

Regressing traits directly on all connectome edges is impractical, as the 68×68 matrices yield $2{,}278$ unique edges and ignores the data's inherent network structure. Tensor network principal component analysis (TN-PCA) provides a structure-preserving method for reducing such network data [Zhang et al., 2019].

Each subject's structural ($\mathbf{S}_i$) and functional ($\mathbf{F}_i$) 68×68 connectivity matrix is approximated by a low-rank orthogonal decomposition:

$$\mathbf{S}_i \approx \sum_{k=1}^{K} c_{ik}^{(\mathrm{SC})}\, \mathbf{U}_k^{(\mathrm{SC})}, \qquad \mathbf{F}_i \approx \sum_{k=1}^{K} c_{ik}^{(\mathrm{FC})}\, \mathbf{U}_k^{(\mathrm{FC})},$$

where $\mathbf{U}k$ are symmetric 68×68 loading matrices (spatial patterns) and cik are subject-specific scores. TN-PCA estimates the loadings and scores by minimizing reconstruction error under orthogonality constraints. For our data, TN-PCA was run with $K = 60$ components per modality, producing two $N \times 60$ score matrices, one for structural and one for functional connectivity. These scores summarize subject-level variation in a compact, standardized form.

Each connectome is mapped from a 68×68 matrix into a K-dimensional vector (with $K = 60$ in our case), dramatically reducing the number of predictors while retaining most of the variance in the data. The decomposition respects the matrix/tensor nature of networks, yielding interpretable spatial patterns. Lower-dimensional TN-PCA scores are less collinear and more suitable for regression, classification, and correlation analyses with behavioral traits. (mean ≈ 0, standard deviation ≈ 0.031 across subjects)

2.3 Additional Dimension Reduction Approaches

To reduce the dimensionality of the $2{,}278$ unique connectivity features, we implemented both a linear and a non-linear approach.

Utilizing vector PCA, we created a linear baseline by vectorizing each subject's symmetric 68×68 structural connectome into a $2{,}278$-dimensional vector. After centering the data matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$, we performed singular value decomposition (SVD):

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \tag{1}$$

Retaining the top $K = 60$ principal directions $\mathbf{V}K$, we projected the data to obtain a $N \times 60$ score matrix $\mathbf{Z}_{PCA} = \mathbf{X}\mathbf{V}_K$. This linear approach captured 37.95% of the variance in structural and 85.64% in functional connectomes.

To capture non-linear structure, we trained a VAE with a 60-dimensional latent space. The model consists of an encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and a decoder $p_\theta(\mathbf{x}|\mathbf{z})$, implemented as multi-layer perceptrons. Given the power-law distribution of streamline counts, we applied a log-transformation ($\log(x+1)$) followed by MinMax scaling to normalize the input range to $[0, 1]$ Hagmann et al. [2008].

For the encoder, we compressed the 2,278 dimensional input into a probabilistic latent space of dimension $K = 60$. It outputs two vectors: a mean $\boldsymbol{\mu}$ and a log-variance $\log \boldsymbol{\sigma}^2$. To match the dimensionality of the PCA approach, the latent dimension was set to 60. The encoder utilizes a hidden dimension of 256. The decoder then aims to reconstruct the original connectivity vector from a sampled latent point $\mathbf{z}$. The VAE was trained by maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L} = \underbrace{\mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}})}_{\text{Reconstruction}} + \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))}_{\text{Regularization}} \tag{2}$$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{x}))) \tag{3}$$

Post-training, we extracted the deterministic mean vector $\boldsymbol{\mu}_i$ for each subject to serve as the stable non-linear biomarker:

$$\mathbf{Z}_{VAE} = \boldsymbol{\mu}_\phi(\mathbf{x}_i) \in \mathbb{R}^{60} \tag{4}$$

This approach achieved a variance explained (R^2) of 76.87% for the structural connectome, significantly outperforming linear PCA (37.95%), demonstrating the non-linear nature of structural brain topology. In contrast, for the functional connectome, the VAE explained 76.18% of the variance compared to 85.64% by linear PCA, suggesting that functional connectivity patterns are inherently more linear in nature.

3 Results

3.1 Predicting alcohol abuse from cognition and connectome features

Lifetime alcohol abuse was modeled using the binarized SSAGA diagnosis `SSAGA_Alc_D4_Ab_Dx`. After QC and exclusion of missing diagnoses, the analytic sample contained 482 men and 562 women, with a prevalence of lifetime alcohol abuse of approximately 20% in men and 11% in women. All models were fit separately by sex. Class imbalance was handled during training by SMOTE Chawla et al. [2002]. For both Explainable Boosting Machines (EBMs) and random forests (RFs) we compared the same seven feature sets per sex:

- **Cognitive only** (`cog_only`): all behavioural and cognitive traits (fluid intelligence, language, memory, attention, delay discounting, executive function, and socio-economic covariates).
- **Connectome only** (`tnpca_only`, `vae_only`, `pca_only`): 60-dimensional structural and 60-dimensional functional latent features derived from TN-PCA, VAE, or vector PCA.
- **Cognitive + connectome** (`tnpca`, `vae`, `pca`): the union of cognitive traits and the corresponding connectome representation.

Table 1 summarizes the key out of sample performance metrics for selected EBM and RF models. We report the best cross-validated AUC on the training set, together with test AUC, and test accuracy on the held-out set. In addition to EBM and RF, we also trained logistic regression and SVM with RBF kernels to compare. For the full result comparison, see Appendix.

Table 1: CV AUC, test AUC, and test accuracy for selected EBM and RF by sex and feature set.

Family	Sex	Feature set	CV AUC	Test AUC	Test acc.
EBM	M	Cognitive only	0.594	0.631	0.763
EBM	M	Cognitive + PCA	0.643	0.663	0.814
EBM	M	PCA only	0.697	0.673	0.786
EBM	F	Cognitive only	0.491	0.584	0.839
EBM	F	Cognitive + PCA	0.533	0.651	0.812
EBM	F	TN-PCA only	0.598	0.551	0.730
RF	M	Cognitive only	0.602	0.517	0.763
RF	M	Cognitive + VAE	0.539	0.703	0.794
RF	M	Cognitive + PCA	0.549	0.660	0.804
RF	F	Cognitive only	0.563	0.554	0.839
RF	F	Cognitive + PCA	0.498	0.613	0.884
RF	F	VAE only	0.531	0.541	0.861

3.1.1 Explainable Boosting Machines

In men, the cognitive-only EBM achieved a test AUC of 0.63 (cross-validated AUC $= 0.59$). Adding connectome features produced modest but consistent gains. The cognitive + PCA model reached a test AUC of 0.66 (CV AUC $= 0.64$), and the PCA-only model achieved the highest male EBM test AUC at 0.67 (CV AUC $= 0.70$), with a test balanced accuracy of 0.63 and precision/recall of roughly 0.44/0.37. In contrast, TN-PCA-only and VAE-only EBMs were only marginally above chance, with

test AUCs of 0.52 and 0.50 and balanced accuracies around 0.52 and 0.49, despite CV AUCs around 0.61 and 0.71.

In women, the cognitive-only EBM reached a test AUC of 0.58 (CV AUC $= 0.49$). The cognitive + PCA variant improved performance to a test AUC of 0.65 (CV AUC $= 0.53$), with test balanced accuracy around 0.56 and precision/recall of 0.21/0.23. The cognitive + TN-PCA model showed the highest single-split test AUC (0.72) but exhibited degenerate threshold behaviour: at the default 0.5 cutoff it predicted no positives (test precision and recall both 0), and its CV AUC (0.57) was only slightly higher than the cognitive-only model. We therefore treat this as an unstable split-specific artefact rather than evidence for a substantially better model. Among connectome-only EBMs in women, TN-PCA-only was slightly above chance (test AUC $= 0.55$), whereas VAE-only and PCA-only were at or below chance (test AUCs ≈ 0.50 and 0.44).

Overall, EBMs indicate that both cognition and low-dimensional connectome representations carry information about lifetime alcohol abuse. Cognitive-only models reach AUCs of 0.63 (men) and 0.58 (women), while adding PCA-based connectome features provides modest but reproducible improvements in test AUC to 0.66 and 0.65. The only connectome-only EBM with clearly above-chance performance is the male PCA-only model; other connectome-only variants contribute little beyond noise.

3.1.2 Random forests

Random forests showed the expected pattern of excellent in-sample fit and more modest generalization. Cross-validated AUCs typically lay between 0.49 and 0.61, while training AUCs were essentially 1.0 for most variants, indicating substantial overfitting even with regularization and imbalance handling.

In men, the cognitive-only RF was only slightly above chance (test AUC $= 0.52$). Adding connectome features improved performance substantially. The cognitive + VAE model was the best male RF overall (test AUC $= 0.70$, CV AUC $= 0.54$), followed by cognitive + PCA (test AUC $= 0.66$). The TN-PCA-only and PCA-only forests were moderately above chance (test AUCs 0.57 and 0.53), while the VAE-only forest fell below chance (test AUC $= 0.42$). In women, the cognitive-only RF achieved a test AUC of 0.55. The best female RF used cognitive + PCA features (test AUC $= 0.61$), whereas the cognitive + TN-PCA and cognitive + VAE models offered no clear advantage (test AUCs 0.52 and 0.48). Connectome-only RF models remained weak: TN-PCA-only and VAE-only were only slightly above chance (test AUCs 0.51 and 0.54), and PCA-only was clearly below chance (test AUC $= 0.37$).

Notably, the best RF and best EBM models have very similar test AUCs for each sex: around 0.67–0.70 in men and 0.61–0.65 in women. Random forests reach slightly higher peak AUCs in some configurations (e.g., the male cognitive + VAE model), but their cross-validated AUCs are not clearly better than EBMs, and they provide far less interpretability. Additional benchmarks with logistic regression and linear SVMs on the same feature sets show broadly similar patterns and do not materially change these conclusions (Appendix A).

3.1.3 Global feature importance across EBMs and random forests

Global importance plots for the EBMs (Figure 1) show a consistent pattern across sexes: in the cognitive + PCA models, the highest-ranked predictors are overwhelmingly PCA-derived structural and functional connectome components. In the male model, the entire top-ranked set is composed of connectome PCs, with age and cognitive/behavioural traits appearing only further down the list. In the female model, a single processing-speed measure (ProcSpeed_Unadj) enters near the lower end of the top-importance features, but most of the importance mass is still carried by connectome PCs. For the cognitive + TN-PCA and cognitive + VAE EBMs (not shown), the pattern is similar: a small number of connectome components account for most of the global importance, with age and cognition contributing more modest mean absolute effects.

Random forest variable-importance plots (Figure 2) tell a related but distinct story. In the cognitive + PCA forests, top-ranked features are a mixture of behavioural and socio-economic traits (e.g., processing speed, reading/language, delay-discounting measures, education and income) and PCA-derived connectome components, rather than being dominated by a single group of predictors. Thus, while EBMs concentrate most of their global importance on connectome PCs, RFs distribute importance more evenly between cognition/SES and connectivity. In the connectome-only forests,

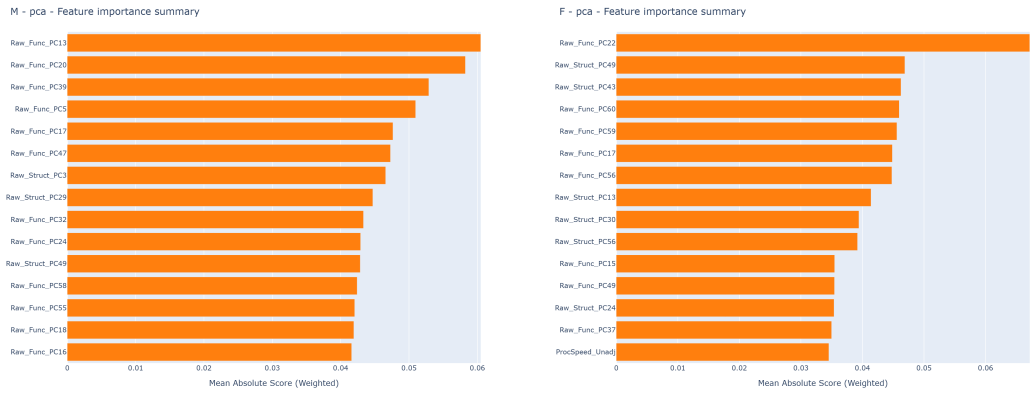

Figure 1: Global importance and summary plots for the cognitive + PCA EBMs in men (left, `summary_M_pca`) and women (right, `summary_F_pca`).

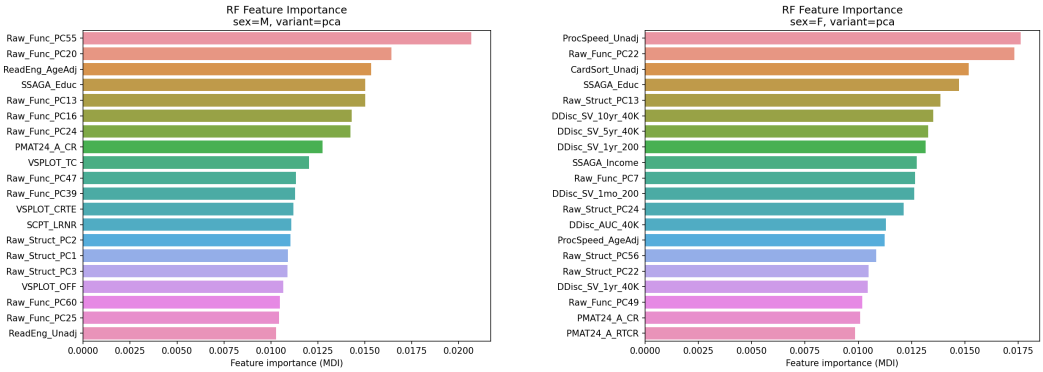

Figure 2: Permutation-based variable importance for the cognitive + PCA RF models in men (left, `varimp_M_pca`) and women (right, `varimp_F_pca`). Bars show the mean decrease in out-of-bag AUC when each feature is permuted, with higher values indicating greater importance.

as expected, PCs dominate the importance rankings, but these models generally achieve lower test AUCs than the best combined models.

3.2 EBM Shape plots

To characterize how individual predictors relate to alcohol abuse risk, we examined EBM shape plots (partial dependence functions) for the most important features. In the main text we focus on age (Figure 3) and a small set of representative cognitive features; additional shape plots are provided in the Appendix.

For both features, the learned shape functions have small amplitudes on the log-odds scale (typically within ± 0.1 to ± 0.2), indicating modest effects. Risk tends to be slightly elevated at mid-range or slightly above-average scores and lower at the very high end, suggesting that very strong cognitive performance is mildly protective, whereas participants in the middle of the distribution carry somewhat higher risk. Age shape functions display a robust non-linear pattern across PCA, TN-PCA, and VAE combined models: predicted risk is lowest at the youngest ages, rises to a peak in the late 20s to early 30s, and then declines again in older participants. This pattern is more pronounced in women, consistent with the strong importance of age in the female models. Shape plots for connectome latent features in the combined models generally show small, often monotonic effects, with contributions rarely exceeding ± 0.3 on the log-odds scale. Together with the near-chance performance of most connectome-only models, this suggests that connectivity does contain signal related to alcohol misuse, but that the signal is weak and distributed across many latent dimensions. EBMs aggregate these many weak effects into a modest overall gain in AUC.

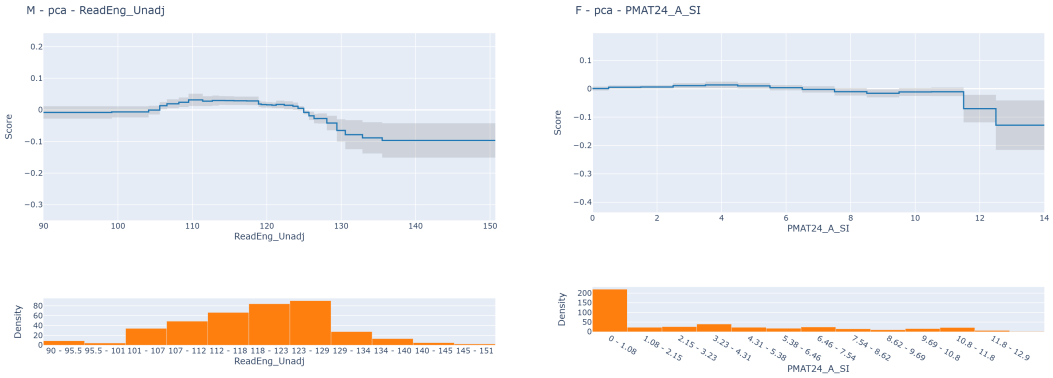

Figure 3: Illustrative EBM shape plots for age in the cognitive and PCA models for men (left, `shape_age_pca_male`) and women (right, `shape_age_pca_female`). Placeholders should be replaced with the actual exported shape-plot figures.

4 Discussion

4.1 Main findings

Our overarching question was how cognitive traits and structural and functional connectome features jointly relate to lifetime alcohol abuse in men and women. From the combined EBM and RF results, three main conclusions emerge. Cognitive and behavioural measures are the primary predictors, as cognitive only EBMs already achieve AUCs of 0.63 in men and 0.58 in women (RFs 0.52 and 0.55), indicating a real but modest signal. Connectome features add small, incremental value when combined with cognition, as the best combined EBMs, especially cognitive and PCA, improve test AUCs from 0.63 to 0.66 in men and from 0.58 to 0.65 in women (gains of roughly 0.03 and 0.07, respectively). Connectome only models are generally weak, with the exception of the male PCA-only EBM, which reaches a test AUC of 0.67. EBMs and RFs perform similarly, but EBMs are more trustworthy. Random forests achieve slightly higher peak test AUCs in some settings (e.g., male cognitive and VAE RF at 0.70), but with nearly perfect training AUCs and only modest cross-validated AUCs. EBMs provide comparable generalization with far greater interpretability and more controlled overfitting.

Thus, in this HCP sample, lifetime alcohol abuse is predominantly associated with age and subtle variations in higher level cognitive function, with connectome structure and function providing only secondary predictive information.

4.2 Interpretation of feature effects

The age-related risk peak in the late 20s to early 30s, seen consistently across models and sexes, aligns with the idea that heavy drinking is concentrated in a relatively narrow young adult window in this cohort. Once age is accounted for, the shape plots suggest that heavy drinkers in this sample are not characterized by gross cognitive impairment. Instead, risk is often highest for participants with mid-range or slightly above average performance on fluid intelligence, language, and executive function measures, and lowest for those at the very top of the distribution. Connectome features appear to carry additional information about alcohol misuse, but the EBMs use them conservativel, such that they rarely enter the top importance ranks in the combined models, and their shape functions have small amplitudes. Combined with the weak performance of most connectome only models, this suggests that, at least in this dataset and with these latent representations, connectivity differences are modest risk modifiers rather than primary drivers.

From an applied perspective, these results imply that age and a relatively small behavioral battery already provide most of the predictive power available in this cohort. The incremental gain from adding MRI-based connectome features may not justify the additional cost and complexity if prediction alone is the goal. For understanding mechanisms, however, the combination of low-dimensional

connectome representations with transparent, shape-plot-based EBMs offers a coherent framework for probing how brain connectivity and cognition jointly relate to substance-use outcomes.

4.3 Limitations and future directions

Several limitations temper our conclusions. First, the outcome is cross-sectional lifetime alcohol abuse in a relatively healthy, highly screened cohort, and the positive class is small within each sex. Even with imbalance-aware training, this reduces power and inflates variability across train-test splits. Second, we treated PCA, TN-PCA, and VAE latent variables as black-box features and did not attempt to map them back to specific networks; any mechanistic claims about particular brain systems would require additional work. Third, although we used cross-validation and a held out test split, we did not have an independent cohort for external validation, so generalization remains uncertain. Finally, we restricted EBMs to additive models. Allowing interactions between key cognitive traits and selected connectome components might reveal more complex patterns at the cost of additional complexity.

Despite these caveats, the a broader picture is clear that within this sample, alcohol abuse is only weakly predictable from the available data, and most of that signal resides in age and cognitive/behavioural measures. Connectome features provide small but consistent incremental value, and interpretable models such as EBMs are well suited to quantifying and visualizing these subtle effects.

Reproducibility

Our analysis was designed from the outset to follow modern best practices for reproducible data science, as were mentioned during the Kitware guest lecture. All code, configuration files, and documentation are publicly available in the `connectopy` repository at `https://github.com/Sean0418/connectopy`. The repository is structured as a Python package with a `src/` layout (`connectopy/`), dedicated `Runners/` scripts for executing end-to-end pipelines, a `tests/` directory with unit tests, Sphinx documentation in `docs/`, and CI/CD workflows in `.github/workflows/`. The `data/` and `output/` directories hold, respectively, user-supplied HCP data and derived analysis outputs.

The core analysis is implemented as reusable Python modules for data loading, dimensionality reduction (PCA, VAE, TN–PCA wrappers), sex-stratified mediation analysis, and machine learning classifiers (RF, XGBoost, EBM, SVM, logistic regression). The same APIs are used by the paper's scripts and by the public Python interface shown in the project documentation. Formal API documentation and a quick start guide are built with Sphinx and shipped in the repository as HTML docs. For users who prefer R, the repository also provides an R interface built on `reticulate`, along with the original R scripts from the initial JASA style template.

To make the computing environment reproducible, we provide both a standard Python package configuration and a containerized setup. The file `pyproject.toml` (and an accompanying `requirements.txt` for users who prefer it) pins all Python dependencies and enables installation via `pip install -e ".[dev,docs]"`. For completely isolated execution, we publish a multi-architecture Docker image `ghcr.io/sean0418/connectopy:latest` that bundles all dependencies; the entire pipeline can then be run with a single `docker run` command after mounting `data/` and `output/`. A Google Colab notebook (`notebooks/colab_demo.ipynb`) offers a "one-click" cloud demo of the pipeline without any local installation.

The main experiments in the paper are reproduced by the script `Runners/run_pipeline.py`, which orchestrates a seven-step workflow: merge and preprocess HCP data, compute vector PCA features, train the VAE and extract latent features, perform sex-dimorphism analyses, fit and evaluate ML classifiers (including EBMs and RFs) for alcohol abuse, run mediation analyses, and generate the figures and tables reported in the paper. Optional flags (e.g., `-quick`, `-skip-vae`, `-skip-plots`) allow users to trade off runtime against completeness, but the default invocation reproduces all results used in the paper. Separate runner scripts (`run_alcohol_analysis.py`, `run_mediation_hcp.py`) expose the alcohol-classification and extended mediation analyses described in section 3.

All stochastic components such as train/validation splits, SMOTE resampling, model initialization, and VAE training are controlled by a single global random seed, set to 42 in the runners, so repeated runs with the same data and environment produce identical metrics and plots. Continuous integration

via GitHub Actions runs linting, type checking, unit tests, and documentation builds on every push, and automatically builds and publishes Docker images. A reproducibility checklist in the repository summarizes these guarantees (packaged code, scripted dependencies, container image, tests, and single-command execution) and is aligned with the NeurIPS reproducibility checklist included in section A.

AI Usage

We used large language models in a limited, post-hoc way to assist with writing and editing documentation. The model did not have access to unpublished subject level information, and it was not used to design experiments, choose outcomes, select models, tune hyperparameters, or generate synthetic data. All statistical analyses, model implementations, evaluation code, and figure generation scripts were executed and debugged by the authors. Any text suggested by an AI system was manually reviewed, edited for technical accuracy, and checked against our own results. The authors take full responsibility for the correctness of the analyses and for all claims, errors, and omissions in the paper.

Author Contributions

The authors are listed alphabetically by last name. All three contributed substantially to every stage of the project.

Conceptualization and study design: Riley Harper (RH), Sean Shen (SS), and Yinyu Yao (YY) and formulated the scientific question about sex differences in the joint relationship between cognition, structural and functional connectomes, and alcohol abuse.

Data curation and preprocessing: SS led the integration of structural and functional connectome data with behavioral and demographic traits, including construction of subject level feature matrices (TN-PCA, PCA, VAE, and cognitive variables), and management of subject IDs across modalities. RH assisted with quality control of derived features and creation of the final analysis ready dataset.

Methodology and software: SS implemented the vector PCA and VAE pipelines and generated the corresponding latent feature sets. YY implemented EBM. RH and YY implemented RF. SS implemented logistic regression and SVMs. RH and YY implemented class-imbalance handling and model comparison scripts. All authors contributed to model specification, evaluation strategy, and visualization.

Reproducibility and infrastructure: RH and SS set up the Git/GitHub repository. RH set up the containerized computing environment and automation scripts for end-to-end reproducibility, including the demo workflow used in the course demonstration. All authors contributed to documentation, testing of the reproduction pipeline, and alignment with the course reproducibility rubric.

Writing: All authors revised the manuscript critically for important intellectual content, approved the final version, and contributed to the slide deck presentation.

References

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, Sydney NSW Australia, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2788613. URL https://dl.acm.org/doi/10.1145/2783258.2788613.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL https://www.jair.org/index.php/jair/article/view/10302.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. ISSN 0885-6125, 1573-0565. doi: 10.1007/BF00994018. URL http://link.springer.com/10.1007/BF00994018.

D. R. Cox. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958. ISSN 00359246. URL http://www.jstor.org/stable/2983890. Publisher: [Royal Statistical Society, Oxford University Press].

Patric Hagmann, Leila Cammoun, Xavier Gigandet, Reto Meuli, Christopher J Honey, Van J Wedeen, and Olaf Sporns. Mapping the Structural Core of Human Cerebral Cortex. *PLoS Biology*, 6(7):e159, 2008. ISSN 1545-7885. doi: 10.1371/journal.pbio.0060159. URL https://dx.plos.org/10.1371/journal.pbio.0060159.

Benjamin J. Heil, Michael M. Hoffman, Florian Markowetz, Su-In Lee, Casey S. Greene, and Stephanie C. Hicks. Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 18(10):1132–1135, 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01256-7. URL https://doi.org/10.1038/s41592-021-01256-7.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374 (2065):20150202, 2016. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2015.0202. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, 2022. URL http://arxiv.org/abs/1312.6114. arXiv:1312.6114 [stat].

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A Unified Framework for Machine Learning Interpretability, 2019. URL http://arxiv.org/abs/1909.09223. arXiv:1909.09223 [cs].

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79, 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.05.041. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811913005351.

Zhengwu Zhang, Genevera I. Allen, Hongtu Zhu, and David Dunson. Tensor network factorizations: Relationships between brain structural connectomes and traits. *NeuroImage*, 197:330–343, 2019. ISSN 10538119. doi: 10.1016/j.neuroimage.2019.04.027. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811919303131.

A Additional model comparison results

Model	Feature set	Sex	Test AUC
EBM	ALL	F	0.562549
EBM	ALL	M	0.699730
Logistic	ALL	F	0.515929
Logistic	ALL	M	0.709177
RF	ALL	F	0.511267
RF	ALL	M	0.660931
SVM	ALL	F	0.541570
SVM	ALL	M	0.681511
EBM	PCA	F	0.641026
EBM	PCA	M	0.641026
Logistic	PCA	F	0.428904
Logistic	PCA	M	0.699055
RF	PCA	F	0.697747
RF	PCA	M	0.642375
SVM	PCA	F	0.465423
SVM	PCA	M	0.608637
EBM	TNPCA	F	0.530692
EBM	TNPCA	M	0.659919
Logistic	TNPCA	F	0.406371
Logistic	TNPCA	M	0.622132
RF	TNPCA	F	0.588190
RF	TNPCA	M	0.542848
SVM	TNPCA	F	0.399378
SVM	TNPCA	M	0.580297
EBM	VAE	F	0.564103
EBM	VAE	M	0.655196
Logistic	VAE	F	0.392385
Logistic	VAE	M	0.621457
RF	VAE	F	0.466977
RF	VAE	M	0.633266
SVM	VAE	F	0.418026
SVM	VAE	M	0.591093

Table 2: Model comparison summary by feature set, sex, and algorithm. Test AUC is computed on the held-out set.

Feature set	Best model (sex = M)	Test AUC
TNPCA	EBM	0.660
PCA	Logistic	0.699
VAE	EBM	0.655
ALL	Logistic	0.709

Table 3: Best-performing model per feature set (by test AUC, males).

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

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 clearly state that our multi-modal approach (VAE, SVM) aims to predict alcohol abuse and identify biomarkers, which is supported by our experimental results and mediation analysis.

   Guidelines:

   - The answer NA 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 NA 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: In the limitations section, we discussed issues such as class imbalance in the subject cohort reducing power in predictions. We also mentioned additional problems such as black box features of dimension reduction and the concern for the generality of the data.

Guidelines:

- The answer NA 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: [NA]

   Justification: This is an empirical study focusing on predictive modeling and biomarker discovery; we apply established algorithms (TN-PCA, VAE) rather than deriving new theoretical theorems.

   Guidelines:

   - The answer NA 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: All of the details needed to reproduce the results are available on GitHub where there are multiple options to do so.

Guidelines:

- The answer NA 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: We utilize the publicly available Human Connectome Project (HCP) release and provide a GitHub repository containing our preprocessing scripts, VAE architecture, and analysis notebooks.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail the hardware (standard CPU/GPU environment via Docker), the software stack (Python 3.10, PyTorch, Scikit-learn), and the specific preprocessing pipeline (log-transform, MinMax scaling) used for all models.

   Guidelines:

   - The answer NA 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: We report AUC scores with cross-validation results to assess stability.

   Guidelines:

   - The answer NA 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: The experiment can be reproduced on Google Colab with one button and without the use of GPU.

Guidelines:

- The answer NA 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: Our research adheres to the NeurIPS Code of Ethics. We analyze anonymous data to extract insights in the brain structure and alcohol abuse.

   Guidelines:

   - The answer NA 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: This research could positively improve early intervention for alcohol abuse by using brain connectivity to identify biomarkers.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our models (SVM, VAE) are standard predictive tools on de-identified medical data and do not pose high risks for dual-use or misuse like generative AI or surveillance systems.

    Guidelines:

    - The answer NA 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: We cite the Human Connectome Project (HCP) data use terms and acknowledge the open-source libraries (Scikit-learn, PyTorch) used in our analysis.

    Guidelines:

    - The answer NA 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: [Yes]

    Justification: We release our processed feature matrices (TN-PCA, VAE Latent Dimensions), the Docker configuration file, and the project package to facilitate reproduction and extension of our work.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: We did not recruit or interact with human subjects directly.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: All authors completed the required HCP agreement and used data collected by HCP.

    Guidelines:

    - The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We used LLMs (like Gemini) for code debugging and LaTeX formatting assistance, but the core scientific hypotheses, methodology, and data interpretation were generated by the authors.

    Guidelines:

- The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.