ARTICLE IN PRESS

NeuroImage xxx (xxxx) xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

A joint network optimization framework to predict clinical severity from resting state functional MRI data

N.S. D'Souza [a,*], M.B. Nebel [b], N. Wymbs [b], S.H. Mostofsky [b,c,d], A. Venkataraman [a]

[a] Department of Electrical and Computer Engineering, Johns Hopkins University, USA
[b] Center for Neurodevelopmental & Imaging Research, Kennedy Krieger Institute, USA
[c] Department of Neurology, Johns Hopkins School of Medicine, USA
[d] Department of Psychiatry and Behavioral Science, Johns Hopkins School of Medicine, USA

ARTICLE INFO

Keywords:
Matrix factorization
Dictionary learning
Functional magnetic resonance imaging
Clinical severity

ABSTRACT

We propose a novel optimization framework to predict clinical severity from resting state fMRI (rs-fMRI) data. Our model consists of two coupled terms. The first term decomposes the correlation matrices into a sparse set of representative subnetworks that define a network manifold. These subnetworks are modeled as rank-one outer-products which correspond to the elemental patterns of co-activation across the brain; the subnetworks are combined via patient-specific non-negative coefficients. The second term is a linear regression model that uses the patient-specific coefficients to predict a measure of clinical severity. We validate our framework on two separate datasets in a ten fold cross validation setting. The first is a cohort of fifty-eight patients diagnosed with Autism Spectrum Disorder (ASD). The second dataset consists of sixty three patients from a publicly available ASD database. Our method outperforms standard semi-supervised frameworks, which employ conventional graph theoretic and statistical representation learning techniques to relate the rs-fMRI correlations to behavior. In contrast, our joint network optimization framework exploits the structure of the rs-fMRI correlation matrices to simultaneously capture group level effects and patient heterogeneity. Finally, we demonstrate that our proposed framework robustly identifies clinically relevant networks characteristic of ASD.

1. Introduction

Resting State fMRI (rs-fMRI) is a non-invasive neuroimaging modality that captures steady-state patterns of co-activation in the brain in the absence of a task paradigm. It is believed that these correlation patterns reflect the intrinsic communication between brain regions [FR07]. Consequently, rs-fMRI has become ubiquitous in the characterization of neuropsychiatric disorders such as Autism Spectrum Disorder (ASD) [MK10], Attention Deficit Hyperactivity Disorder (ADHD) [BVS05], and schizophrenia [NKS03]. Traditional rs-fMRI analysis has concentrated on comparing the statistics of the rs-fMRI data, or variations in these statistics, across individuals or between different cohorts. For example, statistical differences in rs-fMRI features between a patient cohort and neurotypical controls have been considered as biomarkers of a particular disorder. However, the high dimensionality of rs-fMRI data, along with the considerable inter-patient variability, make it extremely difficult to reliably predict clinical manifestations on a patient-specific basis.

There has been considerable work in developing statistical methods to analyze rs-fMRI data. A large number of these studies build on standard multivariate [Woo+17] or random effects models [HF98] to capture population level differences in functional connectivity. Although these studies identify functional connections affected by the disease, they often fail to generalize on a patient-specific level. Additionally, these techniques do not adequately characterize distributed impairments across multiple brain systems, which is crucial for understanding the complex pathologies associated with neuropsychiatric disorders [Kai+10; Kos+05; Rip+07]. This limitation has warranted the development of network-based models to study the inter and intra-subject variation across populations.

Network-based rs-fMRI studies typically group voxels in the brain into regions of interest (ROIs) using a standard anatomical or functional atlas. Further, the synchrony between the average regional time courses is summarized using a similarity matrix, which is the input for further analyses. This extraction procedure is demonstrated in Fig. 1. The works of [Spo+04; RS10; BS09] use graph theoretic notions of connectivity based on aggregate network measures, such as node degree, betweenness

* Corresponding author.
E-mail address: Shimona.Niharika.Dsouza@jhu.edu (N.S. D'Souza).

https://doi.org/10.1016/j.neuroimage.2019.116314
Received 16 July 2019; Received in revised form 23 October 2019; Accepted 24 October 2019
Available online xxxx
1053-8119/© 2019 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article as: D'Souza, N.S. et al., A joint network optimization framework to predict clinical severity from resting state functional MRI data, NeuroImage, https://doi.org/10.1016/j.neuroimage.2019.116314

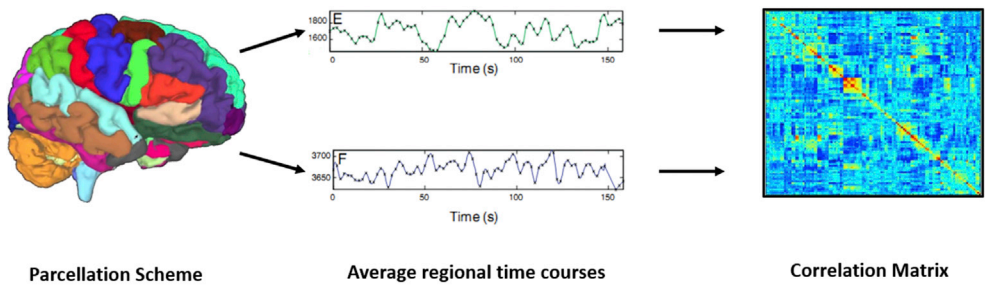

Fig. 1. We group voxels in the brain into ROIs defined by a standard atlas and compute the average time courses for each ROI. The correlation matrix captures the synchrony in the average time courses.

centrality, and eigenvector centrality to study the functional organization of the brain. These measures are extremely useful to compactly summarize the connectivity information onto a restricted set of nodes which map to brain regions. A more global network property is small-worldedness [BB06], which describes an architecture of sparsely connected clusters of nodes. Changes in small-worldedness have been implicated in many neurological disorders [Liu+08] [San+10]. These characterizations are quite successful at capturing global connectivity information, but often fail to illuminate the underlying etiological mechanisms.

To address the limitations of aggregate graph theoretic notions, recent focus has shifted towards mechanistic network models, which incorporate hierarchy onto existing graph connectivity notions. Community detection techniques are a class of population-level models which are used to identify highly interconnected subgraphs within a larger network. These techniques have become popular for understanding the organization of complex systems like the brain network architecture [Bar+16]. An application of this approach to identify regions having abnormal connectivity in schizophrenia patients can be found in [VKG13]. Similarly, Bayesian community detection algorithms developed in [Ven+16] have provided valuable insights in characterizing the social and communicative deficits associated with autism. An alternative network topology is the hub-spoke model, which targets regions associated with a large number of altered rs-fMRI connections [VKG13], [VKG12], [Ven+15]. However, the above methods focus on group characterizations, and even studies that consider patient variability [Ven+17] have little generalization power on new subjects.

Machine learning techniques cast the neuroimaging prediction problem as a two stage procedure. Essentially, the first step is a feature selection or a representation learning stage, while the second stage uses the output of the first to predict the subject characteristics. A simple representation learning framework entails a careful sub-selection of specialized biomarkers [Rav+16; Hon+17]. On a whole brain level, data-driven approaches treat the patient connectivity information as a *feature map* and estimate lower dimensional projections, typically through PCA, kernel-PCA [Sid+12] or ICA [Udd+13]. From here, the most popular classifier (i.e. a stage two algorithm) is a Support Vector Machine (SVM) [Eck+10], which optimizes the decision boundary between patients and neurotypical controls [Udd+13]. SVMs have also been shown to identify disease sub-types [Hon+17] from the lower dimensional features with high accuracy. Along similar lines, the work of [Hoyos-Idrobo, 2018] proposes an ensemble learning based encoder-decoder model, that is able to provide competitive performance on large fMRI datasets at discriminative tasks.

While this two stage pipeline has been successful in the classification realm, characterizing finer-grained measures of clinical severity in the fMRI literature has been restricted to associative analysis, as opposed to an actual prediction on unseen data. For example, the work of [Neb+16] identifies key visual and motor ICA components, which are then used to compute a visuo-motor measure that is significantly correlated with social-communicative and motor deficit measures in ASD. In the context of a continuous value prediction, [Rah+17] develops a modified random forest regression algorithm for stacked multi-output score estimation from multiple ROI-voxel correlation maps. They demonstrate that it outperforms single score prediction. This strategy, however, does not permit a straightforward interpretation of the co-activation patterns explaining an individual severity score. Rather, it identifies regions that explain the complete set of scores jointly. Finally, deep learning methods have become popular for several neuroimaging data analysis. These models have the ability to efficiently learn complex abstractions of the input data without requiring careful feature engineering. As a result, they have been quite successful in a number of case/control classification tasks [Pli+14]. However, a downside to these models is the requirement of large amounts of training data for adequate generalization, which is rarely the case with clinical neuroimaging. Consequently, there has been limited success in predicting behavior from rs-fMRI data using neural networks. In summary, the unification of rs-fMRI and behavioral severity prediction, remains an open challenge.

Dictionary learning [BTD12; Eav+13] methods move away from the pipelined representations, and have recently gained traction due to their ability to simultaneously model both group level and patient specific information. The work of [Eav+15] proposed a correlation matrix decomposition strategy, in which, multiple rank one outer products capture an underlying *generative* basis. The sparse basis representation identifies meaningful co-activation patterns common to all the patients, while patient-specific coefficients combine the subnetworks and model the individual variability in the dataset. An extension of their work [Eav+14] looks at classification of young adults versus children, again, by the addition of an SVM like hinge loss. Our work builds on this representation by using the *discriminative* nature of these coefficients to predict their clinical severity via a linear regression penalty. This Joint Network Optimization (JNO) framework combines both a generative and discriminative term, as opposed to a pipelined hyperparameter search. The generalizability of the model is indicated by the regression performance on unseen data, instead of the correlation fit as used in [Eav+15]. This refinement demonstrates the potential of our JNO framework in identifying patient-predictive biomarkers of a given disorder.

We validate our framework on an rs-fMRI study of Autism Spectrum Disorder (ASD). Patients with ASD are known to manifest a wide spectrum of impairments in terms of social reciprocity, communicative functioning, and repetitive/restrictive behaviours [SW80]. This variation is typically quantified by a clinical severity measure obtained from an expert assessment. We find that our method outperforms several graph theoretic and machine learning feature representation techniques in predicting these severity scores on unseen data. Additionally, our model automatically extracts key networks commonly associated with altered functioning in the ASD literature. Finally, we quantify the merit of our joint objective function by comparing the combined predictive performance with that of a similarly defined two-stage decomposition and regression. We demonstrate that the joint objective bridges the gap in the two representative views of the data, thus aiding the resting state ASD characterization.

A preliminary version of our work appeared in [DSo+18]. Here, we provide a detailed analysis of our model. We demonstrate the predictive

performance of our algorithm on three different clinical severity measures, which capture varied social, behavioral and cognitive deficits associated with ASD. We evaluate our model on two clinical datasets to demonstrate the reproducibility of the method. We identify resting state networks explaining the different behavioral manifestations, and accordingly discuss the robustness of our brain basis characterization across behavioral measures. Lastly, we study hyperparameter sensitivity, generalizability in a test-retest setting and include mitigation strategies to improve the robustness of the framework.

2. Materials and methods

2.1. A joint model for connectomics and clinical severity

Fig. 2 presents a graphical overview of our model. The two inputs to our model are the rs-fMRI similarity matrices (upper left) and the scalar clinical severity scores for each patient (lower right). As mentioned earlier, Fig. 1 illustrates the construction of the similarity matrix from the data. These matrices quantify the Pearson's Correlation Coefficient between the average time courses for each region of interest (ROI). The clinical scores are obtained from an expert evaluation and quantify the severity of the symptoms for the individual.

Notice that the correlation matrices in Fig. 2 have a dual representation. The generative part of the model is indicated in the purple box. Here, we decompose the correlation matrix into a basis term and a patient coefficient term. The columns of the basis capture ROI co-activation patterns common to the entire cohort, while the coefficients differ across patients and quantify the strength of each basis column in the matrix representation. The green box indicates the discriminative part of the model. Here, we leverage the information from the patient-specific coefficients to estimate a given measure of clinical severity via a linear regression model for each individual.

Rs-fMRI Data Representation. We define $\mathbf{\Gamma}_n \in \mathscr{R}^{P\times P}$ as the correlation matrix for patient n, where P is the number of regions given by the parcellation. As seen in Fig. 2, we model $\mathbf{\Gamma}_n$ using a group average basis representation and a patient-specific network strength term. The matrix $\mathbf{B} \in \mathscr{R}^{P\times K}$ is a concatenation of K elemental bases vectors $\mathbf{b}_k \in \mathscr{R}^{P\times 1}$, i.e. $\mathbf{B} := [\mathbf{b}_1 \quad \mathbf{b}_2 \quad \dots \quad \mathbf{b}_K]$, where $K \ll P$. These bases capture steady state patterns of co-activation across regions in the brain. While the bases are common to all patients in the cohort, the combination of these subnetworks is unique to each patient and is captured by the non-negative coefficients $\mathbf{c}_{nk}$. We include a non-negativity constraint $\mathbf{c}_{nk} \geq 0$ on the coefficients to preserve the positive semi-definite structure of the correlation matrices $\{\mathbf{\Gamma}_n\}$. Our complete rs-fMRI data representation is:

$$\mathbf{\Gamma}_n \approx \sum_k \mathbf{c}_{nk}\mathbf{b}_k\mathbf{b}_k^T \quad s.t. \quad \mathbf{c}_{nk} \geq 0 \tag{1}$$

As seen in Eq. (1), we model the heterogeneity in the cohort using a patient specific term in the form of $\mathbf{c}_n := [\mathbf{c}_{n1} \quad \dots \quad \mathbf{c}_{nK}]^T \in \mathscr{R}^{K\times 1}$. Taking $\mathbf{diag}(\mathbf{c}_n)$ to be a diagonal matrix with the K patient coefficients on the diagonal and off-diagonal terms set to zero, Eq. (1) can be re-written in matrix form as follows:

$$\mathbf{\Gamma}_n \approx \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T \quad s.t. \quad \mathbf{c}_{nk} \geq 0 \tag{2}$$

Overall, this formulation strategically reduces the high dimensionality of the data, while providing a patient level description of the correlation matrices.

Modeling Behavioral Scores. As shown in the green box of Fig. 2, the patient coefficients $\{\mathbf{c}_{nk}\}$ from the representation term, are used to model the clinical severity score $\mathbf{y}_n$ using a linear regression vector $\mathbf{w} \in \mathscr{R}^{K\times 1}$

$$\mathbf{y}_n \approx \mathbf{c}_n^T\mathbf{w} \tag{3}$$

Concatenating the vectors $\mathbf{c}_n$ into a matrix $\mathbf{C} := [\mathbf{c}_1 \quad \dots \quad \mathbf{c}_N] \in \mathscr{R}^{K\times N}$, and the severity scores into a vector $\mathbf{y} := [\mathbf{y}_1 \quad \dots \quad \mathbf{y}_N]^T \in \mathscr{R}^{N\times 1}$, Eq. (3) can be equivalently represented in matrix form:

$$\mathbf{y} \approx \mathbf{C}^T\mathbf{w} \tag{4}$$

Joint Objective for Representation and Prediction. We combine the two contrasting viewpoints described above into a joint optimization function by summing the contributions of Eq. (2) and Eq. (4) below:

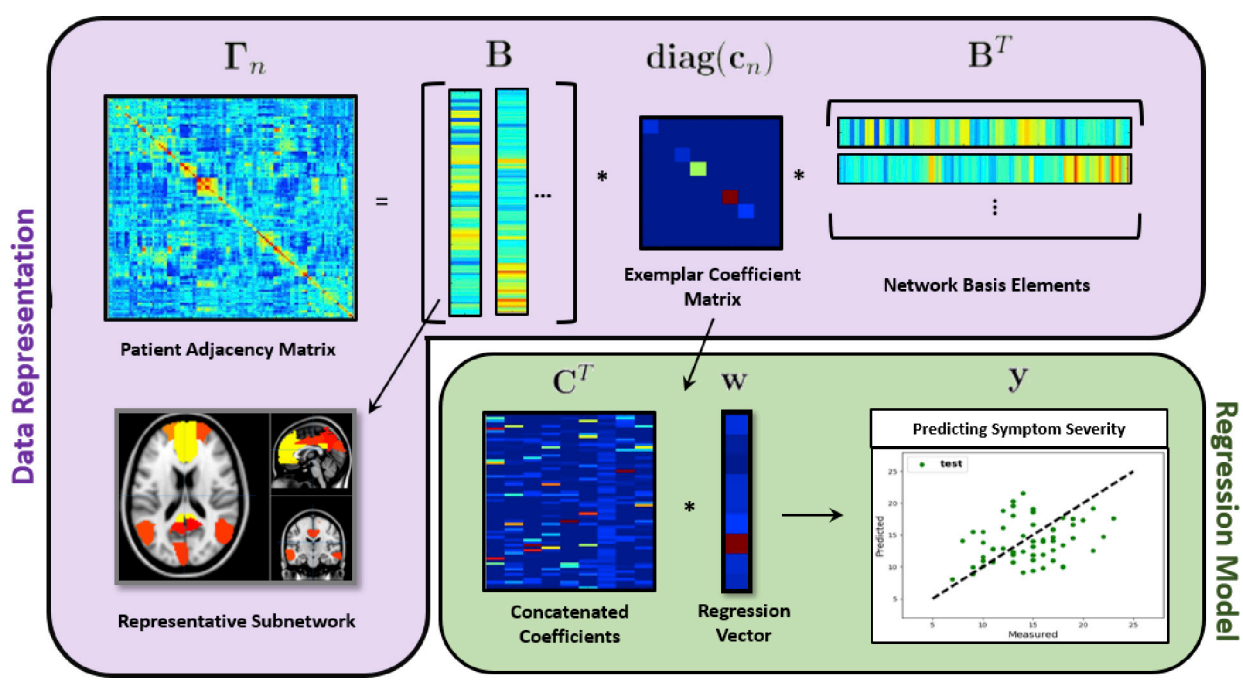

Fig. 2. A two level joint model for connectivity and prediction. **Purple Box:** Depicts the functional data representation or 'generative' term. The correlation matrix is decomposed into a group basis term and a patient specific coefficient term. The columns of the basis matrix correspond to individual subnetworks when projected onto the brain. We stack these coefficients into a matrix. **Green Box:** Prediction of symptom severity via linear regression.

$$\mathscr{J}(\mathbf{B},\mathbf{C},\mathbf{w}) = \sum_n \left|\left|\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T\right|\right|_F^2 + \gamma\left|\left|\mathbf{y} - \mathbf{C}^T\mathbf{w}\right|\right|_2^2 \quad s.t. \quad \mathbf{c}_{nk} \geq 0, \tag{5}$$

Here, $\sum_n \left|\left|\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T\right|\right|_F^2$ is the total error in the representation of the N patient correlation matrices, and $\left|\left|\mathbf{y} - \mathbf{C}^T\mathbf{w}\right|\right|_2^2$ is the prediction error for the behavioral data. Finally, γ is the trade-off between the rs-fMRI data-representation and score prediction terms.

Regularization Penalties. Since we wish to capture a compact, yet clinically informative subnetwork representations, we add an ℓ_1 penalty to encourage sparsity in $\mathbf{B}$. Intuitively, this regularizer will sub-select a small number of nonzero entries in $\mathbf{B}$ that explain the data. From an optimization perspective, notice that scaled solution pairs $\{\mathbf{B},\mathbf{C}\}$ and $\left\{\alpha\mathbf{B}, \frac{1}{\alpha^2}\mathbf{C}\right\}$, as well as $\{\mathbf{C},\mathbf{w}\}$ and $\left\{\beta\mathbf{C}, \frac{1}{\beta}\mathbf{w}\right\}$ give rise to equivalent data representations. As a result, we introduce a quadratic penalty on $\mathbf{C}$ to act as a bound constraint. Similarly, we add an ℓ_2 regularization term to the regression vector $\mathbf{w}$ analogous to ridge regression. Mathematically, the three regularizers can be written as:

$$\lambda_1\left|\left|\mathbf{B}\right|\right|_1 + \lambda_2\left|\left|\mathbf{C}\right|\right|_F^2 + \lambda_3\left|\left|\mathbf{w}\right|\right|_2^2 \tag{6}$$

The penalty terms in Eq. (6) are added to the main objective in Eq. (5). The final joint objective is as follows:

$$\mathscr{J}(\mathbf{B},\mathbf{C},\mathbf{w}) = \sum_n \left|\left|\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T\right|\right|_F^2 + \gamma\left|\left|\mathbf{y} - \mathbf{C}^T\mathbf{w}\right|\right|_2^2 + \lambda_1\left|\left|\mathbf{B}\right|\right|_1 + \lambda_2\left|\left|\mathbf{C}\right|\right|_F^2 + \lambda_3\left|\left|\mathbf{w}\right|\right|_2^2 \quad s.t. \quad \mathbf{c}_{nk} \geq 0, \tag{7}$$

The parameter λ_1 controls the number of nonzero elements in $\mathbf{B}$ by scaling the contribution of the ℓ_1 penalty. Similarly, λ_2 and λ_3 relate to element wise bounds on the entries in $\mathbf{C}$ and $\mathbf{w}$ since they scale the contribution of their respective ℓ_2 norms.

2.2. Optimization algorithm

We employ an alternating minimization technique in order to infer the set of latent variables $\{\mathbf{B},\mathbf{C},\mathbf{w}\}$. Here, we optimize the JNO objective function from Eq. (7) for each output variable, while holding the estimates of the other unknowns constant.

Proximal gradient descent [P+14] is an attractive algorithm to handle the non-differentiable sparsity penalty on $\mathbf{B}$ in Eq. (7), when the supporting terms in the variable of interest are convex. However, from Eq. (7), we see that the Frobenius norm terms expand to a biquadratic representation in $\mathbf{B}$, which is non-convex. We circumvent this problem by introducing N constraints of the form $\mathbf{D}_n = \mathbf{B}\mathbf{diag}(\mathbf{c}_n)$. We enforce these constraints using the Augmented Lagrangian [BS15], denoting the set of Lagrangian matrices by $\{\mathbf{\Lambda}_n\}$. The modified objective function in Eq. (7) takes the form:

$$\begin{aligned}\mathscr{J}(\mathbf{B},\mathbf{C},\mathbf{w},\mathbf{D}_n,\mathbf{\Lambda}_n) &= \sum_n \left|\left|\mathbf{\Gamma}_n - \mathbf{D}_n\mathbf{B}^T\right|\right|_F^2 \\ &+\gamma\left|\left|\mathbf{y} - \mathbf{C}^T\mathbf{w}\right|\right|_2^2 + \sum_n \mathrm{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n))\right] \\ &+\sum_n \frac{1}{2}\left|\left|\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\right|\right|_F^2 + \lambda_1\left|\left|\mathbf{B}\right|\right|_1 \\ &+\lambda_2\left|\left|\mathbf{C}\right|\right|_F^2 + \lambda_3\left|\left|\mathbf{w}\right|\right|_2^2 \qquad s.t. \quad \mathbf{c}_{nk} \geq 0\end{aligned} \tag{8}$$

Such that $\mathrm{Tr}[\mathbf{M}]$ is the trace operator, which sums the diagonal elements of the argument matrix $\mathbf{M}$. The additional Frobenius norm terms $\left|\left|\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\right|\right|_F^2$ act as regularizers for the trace constraints. Observe that Eq. (8) is now convex in both $\mathbf{B}$ and the set $\{\mathbf{D}_n\}$, which allows us to optimize them via standard procedures.

Fig. 3 provides an overview of the alternating minimization strategy employed. Each individual block in our optimization is described below. We refer the interested reader to Appendix A, which systematically delineates the supporting calculations from this section.

Proximal Gradient Descent on $\mathbf{B}$. Given the fixed learning rate parameter t, the proximal update for $\mathbf{B}$ is:

$$\mathbf{B}^{k+1} = \mathbf{sgn}(\mathbf{L}).^*(\mathbf{max}(|\mathbf{L}| - t, 0)) \quad s.t. \quad \mathbf{L} = \mathbf{B}^k - (t/\lambda_1)\frac{\partial\mathscr{J}}{\partial\mathbf{B}} \tag{9}$$

Here, $-\frac{\partial\mathscr{J}}{\partial\mathbf{B}}$ is a descent direction for the $\mathbf{B}$ update, and t controls the magnitude of the step we take in this direction. In practice, we fix t at 10^{-4} for stable convergence. The derivative of $\mathscr{J}$ with respect to $\mathbf{B}$, is computed as :

$$\frac{\partial\mathscr{J}}{\partial\mathbf{B}} = \sum_n \left[2\left[\mathbf{B}\mathbf{D}_n^T\mathbf{D}_n - \mathbf{\Gamma}_n\mathbf{D}_n\right] - \mathbf{D}_n\mathbf{diag}(\mathbf{c}_n)\right] + \sum_n \left[\mathbf{B}\mathbf{diag}(\mathbf{c}_n)^2 - \mathbf{\Lambda}_n\mathbf{diag}(\mathbf{c}_n)\right]$$

At a high level, Eq. (9) performs an iterative shrinkage thresholding operation to handle the non-smoothness of the $\left|\left|\mathbf{B}\right|\right|_1$ using a locally smooth quadratic model.

Optimizing C using Quadratic Programming. The objective is quadratic in $\mathbf{C}$ when $\mathbf{B}$ and $\mathbf{w}$ are held constant. Furthermore, the $\mathbf{diag}(\mathbf{c}_n)$ term decouples the updates for $\mathbf{c}_n$ across patients. We use N quadratic solvers of the form given below to estimate the vectors $\{\mathbf{c}_n\}$:

$$\frac{1}{2}\mathbf{c}_n^T\mathbf{H}_n\mathbf{c}_n + \mathbf{f}_n^T\mathbf{c}_n \quad s.t. \quad \mathbf{A}_n\mathbf{c}_n \leq \mathbf{b}_n \tag{10}$$

The objective and constraint matrices for our quadratic programming solvers are given by:

$$\begin{aligned}\mathbf{H}_n &= \mathscr{I}_K \circ (\mathbf{B}^T\mathbf{B}) + 2\gamma\mathbf{w}\mathbf{w}^T + 2\lambda_2\mathscr{I}_K \\ \mathbf{f}_n &= -2\left[\mathscr{I}_K \circ \left(\mathbf{D}_n^T + \mathbf{\Lambda}_n^T\right)\mathbf{B}\right]1 - 2\gamma y_n\mathbf{w}; \\ \mathbf{A}_n &= -\mathscr{I}_K \quad \mathbf{b}_n = 0\end{aligned}$$

Here, we use $\circ$ to denote the Hadamard product between two matrices and $\mathbf{1}$ to denote a vector of all ones. This strategy helps us find the globally optimal solutions for $\mathbf{c}_n$. The constraint matrices $\mathbf{A}_n$ and $\mathbf{b}_n$ project the solutions onto the K dimensional space of positive reals.

Closed Form Update for $\mathbf{w}$. The global minimizer of $\mathbf{w}$ can be computed by setting the gradient of Eq. (8) equal to zero. Thus, the closed form solution for $\mathbf{w}$ is given by:

$$\mathbf{w} = \left(\mathbf{C}\mathbf{C}^T + \frac{\lambda_3}{\gamma}\mathscr{I}_K\right)^{-1}(\mathbf{C}\mathbf{y}) \tag{11}$$

This is analogous to a **regularized linear regression, i.e. ridge regression** update for $\mathbf{w}$.

Optimizing the Constraint Variables $\mathbf{D}_n$ ***and*** $\mathbf{\Lambda}_n$. Similar to the case of $\mathbf{w}$, each of the primal variables $\{\mathbf{D}_n\}$ has a closed form solution given by:

$$\mathbf{D}_n = (\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T + 2\mathbf{\Gamma}_n\mathbf{B} - \mathbf{\Lambda}_n)(\mathscr{I}_K + 2\mathbf{B}^T\mathbf{B})^{-1} \tag{12}$$

We update the dual variables $\{\mathbf{\Lambda}_n\}$ via gradient ascent:

$$\mathbf{\Lambda}_n^{k+1} = \mathbf{\Lambda}_n^k + \eta_k(\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)) \tag{13}$$

The updates for $\mathbf{D}_n$ and $\mathbf{\Lambda}_n$ ensure that the proximal constraints are satisfied with increasing certainty at each iteration. The learning rate parameter η_k for the gradient ascent step of the augmented Lagrangian is chosen to guarantee sufficient decrease for every iteration of alternating

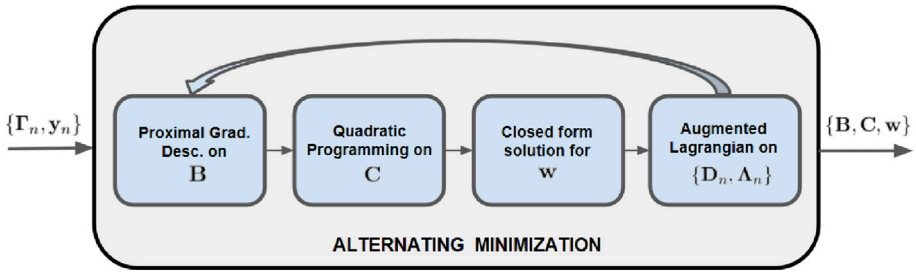

Fig. 3. Our Optimization Strategy, we iterate through four main steps until global convergence.

minimization. In practice, we initialize this value to 10^{-3}, and scale it by 0.5 at each iteration.

2.3. Prediction on an unseen patient

In order to estimate the coefficients $\widehat{\mathbf{c}}$ for a new patient, we re-solve the quadratic program in Eq. (10) using the $\{\mathbf{B}^*, \mathbf{w}^*\}$ computed from the training data via the procedure outlined in Section 2.2. We explicitly set the contribution from the data term in Eq. (5) to 0, since the corresponding value of $\widehat{\mathbf{y}}$ is unknown for the new patient. We also implicitly assume that the conditions for the proximal operator hold, i.e. the constraint $\widehat{\mathbf{D}} = \mathbf{B}^*\mathbf{diag}(\widehat{\mathbf{c}})$ is exactly satisfied. The estimation of the unseen patient's coefficients are thus mathematically formulated as follows:

$$\begin{aligned} \widehat{\mathbf{c}} = \arg\min_{\mathbf{c}} \; & \left|\left|\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c})\mathbf{B}^T\right|\right|_F^2 + \lambda_2 \left|\left|\mathbf{c}\right|\right|_2^2 \\ s.t. \quad & \mathbf{c}_k \geq 0 \end{aligned} \tag{14}$$

Once again, Eq. (14) can be formulated as a quadratic program. The parameters from Eq. (10) correspond to:

$$\begin{aligned} \mathbf{H}_n &= 2(\mathbf{B}^T\mathbf{B}) \circ (\mathbf{B}^T\mathbf{B}) + 2\lambda_2 \mathscr{I}_K \\ \mathbf{f}_n &= -2[\mathscr{I}_K \circ (\mathbf{B}^T\mathbf{\Gamma}_n\mathbf{B})]1; \\ \mathbf{A}_n &= -\mathscr{I}_K \qquad \mathbf{b}_n = 0 \end{aligned}$$

The estimate for the behavioral score for the test patient is given by the vector product $\widehat{\mathbf{y}} = \widehat{\mathbf{c}}^T\mathbf{w}^*$.

2.4. Baseline comparison techniques

We evaluate the performance of our method against a set of well established statistical, graph theoretic, and data-driven frameworks that have been used to provide rich feature representations. Fig. 4 describes a general two stage pipeline for our task. The first stage is a representation learning step used for feature extraction. Stage 2 is a regression model to map the learned features to behavioral data. We evaluate our method against several choices of linear and non-linear algorithms for Stage 1. These are combined with a regularized linear regression in Stage 2, similar to our method. Additionally, we evaluate the performance obtained by omitting a Stage 1 and training a deep neural network end-to-end on the input correlation features. Lastly, we demonstrate the advantage provided by combining the neuroimaging and behavioral representations in the JNO framework. For this, we present a comparison where the feature learning and prediction stages are decoupled, similar to the baselines.

2.4.1. Machine learning approach (PCA)

We start with the $P \times P$ correlation matrix $\mathbf{\Gamma}_n$ for each patient. Since this matrix is symmetric, we have $M = \frac{P\times(P-1)}{2}$ distinct rs-fMRI correlation pairs between various communicating sub-regions. Accordingly, the features from every individual are composed into a descriptor matrix $\mathbf{X} \in \mathscr{R}^{M\times N}$. We further concentrate these feature into a small number of representative bases. The basis extraction procedure in Stage 1 corresponds to a linear mapping in the original correlation space via a **Principal Component Analysis (PCA)**. In Stage 2, we construct a **regularized linear regression (ridge regression)** on the projected features to predict the clinical severity. PCA projects the observations onto a set of uncorrelated *principal component basis* by means of an orthogonal linear decomposition. Mathematically, PCA poses the following dimensionality reduction problem:

$$\begin{aligned} \mathscr{F}(\mathbf{U}, \mathbf{Z}, \boldsymbol{\mu}) = \arg\min_{\boldsymbol{\mu}, \mathbf{U}, \mathbf{Y}} \; & \left|\left|\mathbf{X} - \boldsymbol{\mu}1^T - \mathbf{U}\mathbf{Z}\right|\right|_F^2 \\ s.t. \quad & \mathbf{U}^T\mathbf{U} = \mathscr{I}_d, \quad \mathbf{Z}1 = 0 \end{aligned} \tag{15}$$

Here, $\mathbf{U} \in \mathscr{R}^{N\times d}$ is the d dimensional subspace basis which best approximates the information from $\mathbf{X}$ in the Frobenius norm sense, computed by calculating the eigenvectors of the sample covariance matrix $\mathbf{X}\mathbf{X}^T$. Consequently, $\mathbf{Z} \in \mathscr{R}^{d\times N}$ is a compact d dimensional representation of $\mathbf{X}$, where $d \ll M$. $\mathbf{1}$ is a d dimensional vector of ones. The constraint $\mathbf{Z}1 = 0$ centers the representation $\mathbf{Z}$.

2.4.2. Statistical approach (ICA)

Here, we use **Independent Component Analysis (ICA)** as the Stage 1 algorithm combined with **ridge regression**. ICA operates on the raw time series data to extract representative spatial patterns that explain rs-fMRI connectivity. ICA has become ubiquitous for identifying group level as well as individual-specific connectivity signatures. It decomposes a multivariate signal into 'independent' non-Gaussian components based on the statistics of the data. Mathematically, ICA models the components $\{\mathbf{y}_k\}$ of the observed signal $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]$ as a sum of n independent components $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_n]$ combined via the mixing matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$

$$\mathbf{y} = \sum_{i=1}^{n} \mathbf{s}_i\mathbf{a}_i \qquad i.e. \quad \mathbf{Y} = \mathbf{A}\mathbf{S} \tag{16}$$

$\mathbf{s}$ can be recovered by multiplying the observed signals $\mathbf{Y}$ with the inverse of the mixing matrix $\mathbf{W} = \mathbf{A}^{-1}$. We adaptively estimate both the mixing matrix $\mathbf{A}$ and the components $\mathbf{s}$ by setting up a cost function that maximizes the non-gaussianity of $\mathbf{s}_i = \mathbf{w}_i^T\mathbf{y}$ or minimizes the mutual information.

Group ICA extends this algorithm to a multi-subject analysis for extracting independent spatial patterns common across patients, but combined via individual time courses. We use the GIFT [CLA09] software in order to perform Group-ICA to derive independent spatial maps for each patient. The correlation values between the identified components are fed to the regression model.

2.4.3. Graph theoretic approach (node degree)

Each correlation matrix $\mathbf{\Gamma}_n$ can be thresholded and considered a graph adjacency matrix, which we denote by $\mathbf{\Psi} \in \mathscr{R}^{P\times P}$. The element $\mathbf{\Psi}_{ij}$ gives the strength of association between two communicating sub-regions i and j. The underlying graph topology can be summarized using node/edge based importance measures [Spo+04] [BB06]. Again, we use a regularized linear regression technique to estimate the severity score from the reduced representation. This treatment closely parallels the machine

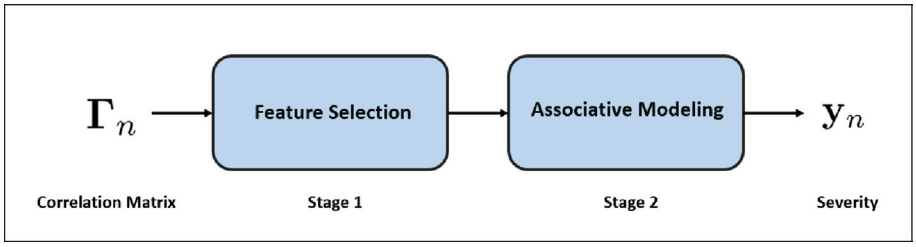

Fig. 4. A typical two stage baseline. We input the correlation matrices to Stage 1, which performs Feature Extraction on the raw correlations. This step could be a technique from machine learning, graph theory or a statistical measure. Stage 2 fits an associative regression model to the output representation of Stage 1.

learning approach, as we can view the graph measures as a dimensionality reduction. We compute **Node Degree (D_N)** from the adjacency graph followed by a **ridge regression** on the features.

Given the adjacency matrix $\mathbf{\Psi}$, the degree of region v is equal to the number of edges incident on v, with loops counted twice. Mathematically, the degree $\mathbf{D}_N(v)$ is computed as follows:

$$\mathbf{D}_N(v) = \sum_{j \neq v} 1\left(\mathbf{\Psi}_{jv} > 0\right) \tag{17}$$

where, $1(.)$ is the indicator function, which takes the value 1 if the condition is satified, and 0 otherwise. This metric captures the importance of each node in explaining the graph, which in our case, corresponds to the average connectivity strength of each region in the brain.

2.4.4. A neural network approach

Recently, there has been an upsurge in using neural networks to investigate neuroimaging correlates of developmental disorders [Kai+10]. Here, we test the efficacy of a simple **Artificial Neural Network (ANN)** for predicting the severity score from the correlation feature matrix $\mathbf{X}$ defined above. The network architecture encodes a series of non-linear transformations of the input correlations to approximate the severity score. Recall that the size of the input is dependent on our choice of parcellation, which could be of considerable width (of the order of ≈ 5000 connections for $P = 100$). After evaluating several architectures, we employ a two hidden layer network with widths 8000 and 10 respectively. We use a Rectified Linear Unit (ReLU) non-linearity after the first hidden layer and a Tanh non-linearity after the second hidden layer. We used the ADAM optimizer with an initial learning rate of 10^{-4}, scaled by 0.9 per 10 epochs, and a momentum of 0.9 to train the network.

3. Experiments

3.1. Validation on synthetic data

As a sanity check, we first sample data from the generative model in Eq. (7) and use the optimization outlined in Section 2.2 to estimate the unknowns $\{\mathbf{B}, \mathbf{C}, \mathbf{w}\}$. This procedure helps us analyze the performance of the algorithm under different noise scenarios. The inputs to our model are the correlation matrices $\{\mathbf{\Gamma}_n\}$ and the clinical scores $\{\mathbf{y}_n\}$. We note that the model gives a complete description of each $\mathbf{\Gamma}_n$ in terms of the basis vectors $\{\mathbf{b}_k\}$ and the patient coefficients $\{\mathbf{c}_n\}$. Since the data representation terms for each patient are coupled solely through the basis representation, the coefficient descriptors are independent of each other. In a similar observation, each score $\mathbf{y}_n$ is explained by the corresponding $\mathbf{c}_n$, independent of the remaining subjects, when we fix the regression vector $\mathbf{w}$. We use this information to describe the *observed data* $\{\mathbf{\Gamma}_n, \mathbf{y}_n\}$ using a generative model with the likelihood model based on the *hidden variables* $\{\mathbf{B}, \mathbf{C}, \mathbf{w}\}$.

Notice that, when treated as a Bayesian log-likelihood (i.e. taking a negative exponent of the objective), the ℓ_2 norms in Eq. (7) translate into Gaussian distributions, and the ℓ_1 norm is equivalent to a Laplacian prior. The corresponding graphical model is shown in Fig. 5. The observed variables are indicated by the shaded circles. The white circles contain the hidden variables. The distribution parameters for the hidden variables are indicated in the corresponding rectangle pointing to the variable. The Laplacian parameter σ_B controls the overlap in the patterns of sparsity in $\mathbf{B}$, which relates to λ_1. $\mathbf{C}$ and $\mathbf{w}$ are described by Gaussians with means zero (i.e. ℓ_2 norm offset). The variances $\sigma^2_{\mathbf{C}}$ and $\sigma^2_{\mathbf{w}}$ are related to the penalty parameters λ_2 and λ_3 respectively. The non-negativity constraint on $\mathbf{c}_n$ is handled by folding (i.e. taking the absolute value of) the normal distribution to restrict the $\mathbf{c}_n$ values to be positive reals. The observed variable $\{\mathbf{y}_n\}$, translates to a Gaussian with mean $\boldsymbol{\mu}_{\mathbf{y}_n} = \mathbf{c}_n^T\mathbf{w}$, and variance parameters $\sigma_{\mathbf{y}_n}$. This is again folded to reflect positive values of $\mathbf{y}_n$. The correlation matrices $\{\mathbf{\Gamma}_n\}$ are drawn from a Gaussian distribution with mean $\boldsymbol{\mu}_{\mathbf{\Gamma}_n} = \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T$ (which is positive semi-definite by construction) and variance $\boldsymbol{\sigma}_{\mathbf{\Gamma}_n}$.

There are two sources of noise for the observed variables, which include the error in the correlation matrices $\mathbf{\Gamma}_n$, and the error in the severity scores $\mathbf{y}_n$. These scenarios can be directly related to controlling the variance parameters $\sigma_{\mathbf{\Gamma}_n}$ and $\sigma_{\mathbf{y}_n}$ respectively. Additionally, we are interested in the performance of the algorithm under varying levels of overlap in the sparsity patterns in $\mathbf{B}$.

We evaluate the performance using an average inner-product measure of similarity S between each recovered network, $\widehat{\mathbf{b}}_k$, and its corresponding best matched generating network, $\mathbf{b}_k$, both normalized to unit norm, i.e.:

$$S = \frac{1}{K}\sum_k \frac{\left|\mathbf{b}_k^T \widehat{\mathbf{b}}_k\right|}{\left|\left|\mathbf{b}_k\right|\right|_2 \left|\left|\widehat{\mathbf{b}}_k\right|\right|_2}. \tag{18}$$

Fig. 6 depicts the performance of the algorithm in these three cases. The x-axis corresponds to increasing the levels of noise, while the y-axis indicates the similarity metric S computed for the particular setting. In the leftmost plot, an x-axis value close to 0 indicates high percentage of sparsity in $\mathbf{B}$, while increasing values correspond to denser basis matrices. Throughout this experiment, the values of the other free parameters in the generative model were held constant. The middle plot evaluates subnetwork recovery when the noise in the scores, i.e. $\boldsymbol{\sigma}_{\mathbf{y}_n}$ is increased. The x-axis reports normalized values of $\boldsymbol{\sigma}_{\mathbf{y}_n}$ while the remaining free parameters were held constant. Similarly, the rightmost plot in Fig. 9 indicates performance under varying noise in the correlation matrices $\mathbf{\Gamma}_n$. Again, normalized $\boldsymbol{\sigma}_{\mathbf{\Gamma}_n}$ values are reported on the x-axis. All numerical results have been aggregated over 100 independent trials.

As expected, increasing the noise in the correlation matrices and scores worsens the recovery performance of the algorithm. This is indicated by the decay in the similarity measure with increasing noise parameters as well as an increase in the corresponding variance. Additionally, the algorithm performs better when there is lesser overlap in the columns of $\mathbf{B}$, i.e. when the generating basis is sparse. However, we observe that our algorithm is robust in the noise regime estimated from the real-world rs-fMRI data $(0.01 - 0.2)$ and recovered sparsity levels $(0.1 - 0.4)$. In addition, we identify the stable parameter settings for the algorithm which guide our real world experiments.

3.2. A population study of autism

We evaluate the efficacy of our JNO framework on two separate cohorts. Our first dataset consists of 58 children with high functioning ASD acquired at the Kennedy Krieger Institute in Baltimore, USA. We refer to this as the KKI dataset. The age of the subjects ranged from 10.06 ± 1.26 and the IQ as 110 ± 14.03. The second cohort is a subset of the publicly available Autism Brain Imaging Data Exchange (ABIDE I) [+Di +14+] acquired at the New York University (NYU) site consisting of 63 patients. Social and communicative deficits in autism are believed to arise from abnormal interactions between functionally linked regions in the brain [PYM14]. Therefore, identifying long-range correlative patterns in the brain directly linked to clinical severity is an important stepping stone to understanding and quantifying the neural underpinnings of the disorder.

Neuroimaging Data. For the KKI dataset, rs-fMRI scans were acquired on a Phillips 3*T* Achieva scanner using a single shot, partially parallel gradient-recalled EPI sequence with TR/TE $=$ 2500/30 ms, flip angle 70, res $= 3.05 \times 3.15 \times 3$mm, having 128 or 156 time samples. The children were instructed to relax with eyes open and focus on a central cross-hair while remaining still for the duration of the scan.

Slice time correction, rigid body realignment, and normalization to the EPI version of the MNI template was performed as a part of preprocessing using SPM [Pen+11]. The time courses were temporally detrended in order to remove gradual trends in the data. From here, we used a CompCorr [Beh+07] strategy for the estimation and removal of spatially coherent noise from the white matter and ventricles, along with the linearly detrended versions of the six rigid body realignment parameters and their first derivatives. We performed a spatial smoothing with a 6 mm FWHM Gaussian kernel followed by a temporal filtering using a $0.01 - 0.1$ Hz pass band. Lastly, we removed spikes from the data via tools from the AFNI package [Cox96] as an alternative to motion scrubbing.

For the NYU cohort, the rs-fMRI data was pre-processed using the configurable pipeline for the analysis of connectomes, that has been integrated with ABIDE. The pre-processing steps involved are skull-stripping, global mean intensity normalization, spatial normalization to the MNI template, nuisance regression and CompCorr, followed by bandpass filtering, but without global signal regression.

This work relies on the Automatic Anatomical Labelling (AAL) atlas [Tzo+02], which defines 116 cortical, subcortical and cerebellar regions. We compute a 116×116 correlation matrix for each patient based on the Pearson's Correlation Coefficient between the average time series for these regions. Empirically, we observed a consistent noise component with nearly constant contribution from all brain regions and low predictive power for both datasets. Therefore, we subtracted out the first eigenvector contribution from each of the correlation matrices and used the residuals as the inputs $\{\mathbf{\Gamma}_n\}$ to the algorithm and the baselines.

Behavioral Data. We analyzed three independent measures of clinical severity. These include:

1 Autism Diagnostic Observation Schedule, Version 2 (ADOS-2) total raw score
2 Social Responsiveness Scale (SRS) total raw score
3 Praxis total percent correct score

ADOS and SRS are standard assessments and are available for both the KKI and NYU datasets, while the Praxis has been collected for the KKI dataset only.

The ADOS consists of different sub-scores which quantify the social and commumunicative deficits of the patient along with the restrictive/repetitive behaviors [Lor+00]. The test is administered by a trained clinician who evaluates the child against a set of guidelines. The total score is computed by adding the individual sub-scores. The dynamic range for ADOS is between 0 and 30, with higher score indicating greater impairment.

The SRS scale characterizes the social responsiveness of an individual

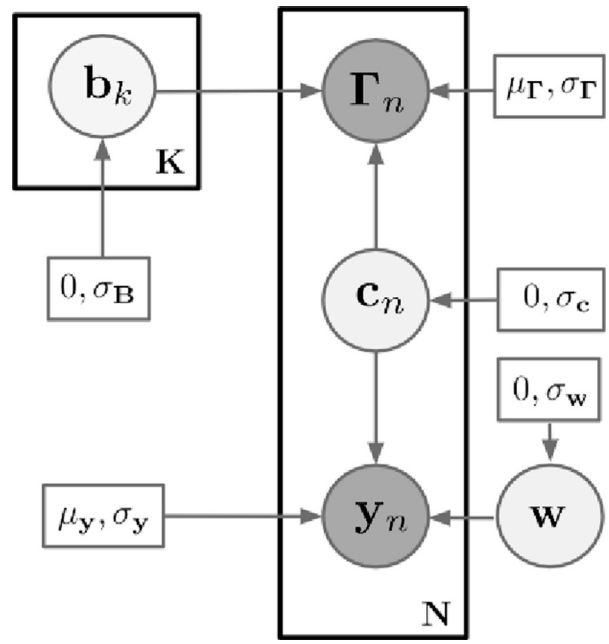

Fig. 5. The graphical model for the joint objective. For our synthetic experiments, we fix the model parameters. $\boldsymbol{\sigma}_\mathbf{C} = 2, \boldsymbol{\sigma}_\mathbf{w} = 0.2$

[BPC08]. Typically, a parent/care-giver or teacher completes out a standardized questionnaire assessing various aspects of the child's behavior. SRS reporting tends to be more variable across patients compared to ADOS, since the responses are heavily biased by the parent/teacher attitudes. The SRS dynamic range is between $70 - 200$ for ASD patients, with higher values corresponding to higher severity in terms of social responsiveness.

Praxis was assessed using the Florida Apraxia Battery (modified for children) [Mos+06], which assesses ability to perform skilled motor gestures to command, to imitation, and with actual tool use. Several studies [Mos+06], [Dzi+07], [DMM09], [Neb+16], etc) reveal that children with ASD show marked impairments in Praxis i.e., developmental dyspraxia, and that impaired Praxis correlates with impairments in core autism social-communicative and behavioral features. Performance is videotaped and later scored by two trained research-reliable raters, with total percent correctly performed gestures as the dependent variable of interest. Scores therefore range from $0 - 100$, with higher scores indicating better Praxis performance. This measure was available for 52 of the total patients in the KKI dataset.

3.3. Evaluating predictive performance

We wish to compare the performance of our JNO framework against a wide class of algorithms described in Section 2.4. In this paper, we use regularized linear regression (i.e. ridge regression) for all baselines, except the ANN. Our Supplementary Results document includes further comparison with a non-linear (Random Forest) regression model. We emphasize that the baseline performance is nearly identical for both the linear and the non-linear models.

We characterize the performance of each method using a 10 fold cross validation strategy as illustrated in Fig. 7. For a given parameter setting, we first split the data set into 10 training and test folds. For each of the folds, we train the models on a 90 percent training set split of the data. We report the score prediction on the held out 10 percent, which constitutes the testing set for that fold. Note that each datapoint is a part of the test set in exactly one of the 10 folds.

We report two quantitative measures of performance. Median Absolute Error (MAE) quantifies the absolute distance between the measured

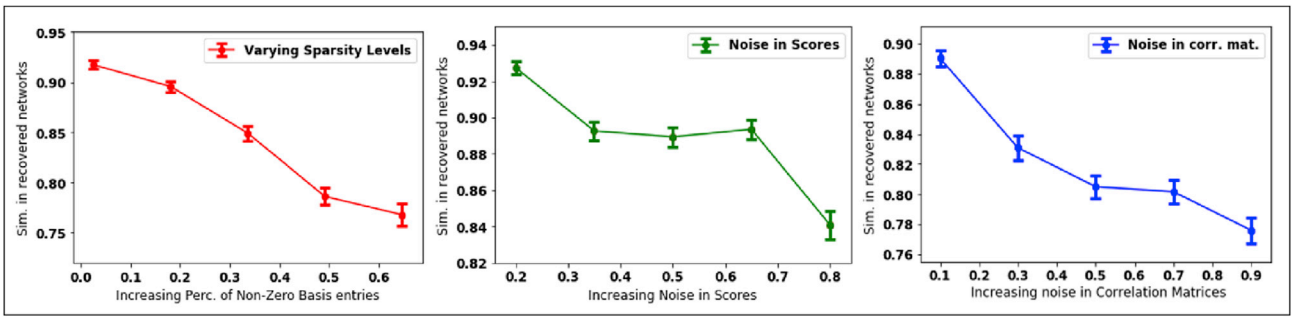

Fig. 6. Performance on synthetic experiments. (**L**): Varying the level of sparsity ($\sigma_{\Gamma_n} = 0.4$, $\sigma_{\mathbf{y}_n} = 0.2$), (**M**): Varying the level of noise in $\mathbf{y}_n$ ($\sigma_{\mathbf{B}} = 0.2$, $\sigma_{\Gamma_n} = 0.4$), (**R**): Varying the level of noise in $\mathbf{\Gamma}_n$ under ($\sigma_{\mathbf{B}} = 0.2$, $\sigma_{\mathbf{y}_n} = 0.2$) Values on the x-axis have been normalized to reflect a $[0-1]$ range by dividing by the maximum value of the variable. Deviations from the mean recovered similarity for each parameter setting is indicated in the figure and have been reported as a standard error value. The reported *x*-axis range reflects the regimes within which the algorithm converges to a local solution.

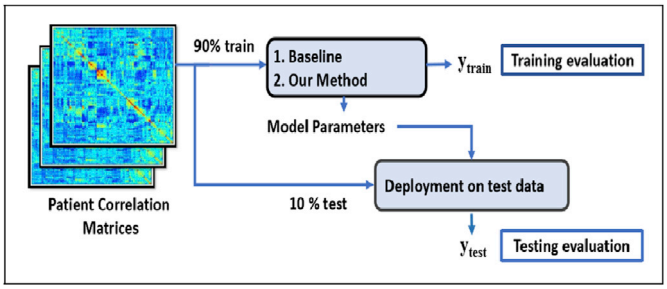

Fig. 7. A ten-fold cross validation for evaluating performance.

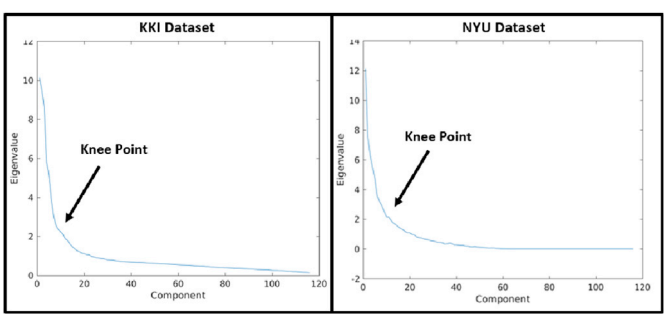

Fig. 8. Scree Plot of the correlation matrices to corrobrate the selected values for K. **(L)** KKI Dataset **(R)** NYU Dataset.

and predicted scores:

$$\text{MAE} = \text{median}(|\widehat{\mathbf{y}} - \mathbf{y}|),$$

where the median is computed across the set of patients. We report MAE values with the standard deviation of the error. Lower MAE indicates better testing performance.

Normalized Mutual Information (NMI) assesses the similarity in the distribution of the predicted and observed score distributions across test patients. NMI is computed as follows:

$$\text{NMI}(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{H(\mathbf{y}) + H(\widehat{\mathbf{y}}) - H(\mathbf{y}, \widehat{\mathbf{y}})}{\min\{H(\mathbf{y}), H(\widehat{\mathbf{y}})\}}$$

where $H(\mathbf{y})$ denotes the entropy of $\mathbf{y}$ and $H(\mathbf{y}, \widehat{\mathbf{y}})$ is the joint entropy between $\mathbf{y}$ and $\widehat{\mathbf{y}}$. NMI ranges from $0-1$ with higher values indicating a better agreement between predicted and measured score distributions, and thus characterizing improved performance.

Along with MAE and NMI, we perform a statistical test on the error distribution to evaluate the performance gain of our framework. We first calculate the Cumulative Distribution Function (CDF) of errors and then report the Kolmogorov-Smirnov statistic on the CDF. This statistic helps us quantify the differences in the error distributions of each baseline compared to our algorithm. A Kolmogorov-Smirnov statistic lower than $\alpha = 0.05$ is widely accepted in literature as statistically significant. We indicate comparisons which fall within this threshold in bold and near misses using an underline.

3.3.1. *Parameter settings*

Our method has five user-specified parameters $\{\gamma, \lambda_1, \lambda_2, \lambda_3, K\}$. Recall that K is the number of basis networks, γ is the penalty tradeoff between the representation and regression terms, λ_1 is the sparsity penalty, while λ_2 and λ_3 are the regularization penalties on the coefficients $\mathbf{C}$ and regression weights $\mathbf{w}$ respectively.

We use the knee point of the eigenspectrum of the correlation matrices $\mathbf{\Gamma}_n$ to select the number of bases ($K = 8$ for both datasets). For reference, we have included the scree-plots in Fig. 8. Empirically, the JNO model is insensitive to the choice of λ_3 and γ, so we fix both at one. Effectively, we are left with two free parameters, which we optimize by performing a bivariate grid search. We note that the generalization accuracy is dependent on the dynamic range of the scores and is sensitive to λ_1 and λ_2. As described in Section 3.6, we have identified a stable range of operation across a single order of magnitude for these parameters. Based on the cross validation results, we finally use the following settings in our experiments: For the KKI dataset, $\{\lambda_2 = 0.2, \lambda_1 = 30\}$ for ADOS, $\{\lambda_2 = 0.9, \lambda_1 = 50\}$ for SRS and $\{\lambda_2 = 0.6, \lambda_1 = 20\}$ for Praxis; for the NYU dataset, $\{\lambda_2 = 0.1, \lambda_1 = 20\}$ for ADOS, $\{\lambda_2 = 0.9, \lambda_1 = 40\}$ for SRS.

To provide a fair comparison with our JNO framework, we use a joint grid search on the Stage 1 hyperparameters and the Stage 2 ridge penalty to optimize these values for every baseline method. Again, we report the best performance in a ten fold cross validation setting.

We select 10 PCA components for the KKI dataset, and 15 for the NYU dataset. For ICA, we obtained good performance for 35 spatial maps obtained from GIFT [CLA09]. For the graph theoretic baseline, we threshold the correlation matrices $\{\mathbf{\Gamma}_n\}$ at 0.2 to obtain valid adjacency matrices $\{\mathbf{\Psi}_n\}$. In conjunction with these, the ridge penalty parameter was swept across four orders of magnitude.

As described in Section 2.4.4, we fixed the network architecture to be a two hidden layer network with widths 8000 and 10 respectively, having a Rectified Linear Unit (ReLU) non-linearity after the first hidden layer and a Tanh non-linearity after the second hidden layer. We use a standard weight decay regularizer, with the regularization parameter varied over three orders of magnitude for each baseline comparison. We trained the network using the ADAM optimizer with an initial learning rate of 10^{-4}, scaled by 0.9 per 10 epochs, and a momentum of 0.9.

Finally, we include the performance upon decoupling the ridge regression and the matrix decomposition in Eq. (5) as a sanity check. This is akin to the two stage treatment in the baselines where the two terms are not explicitly coupled as in the JNO objective.

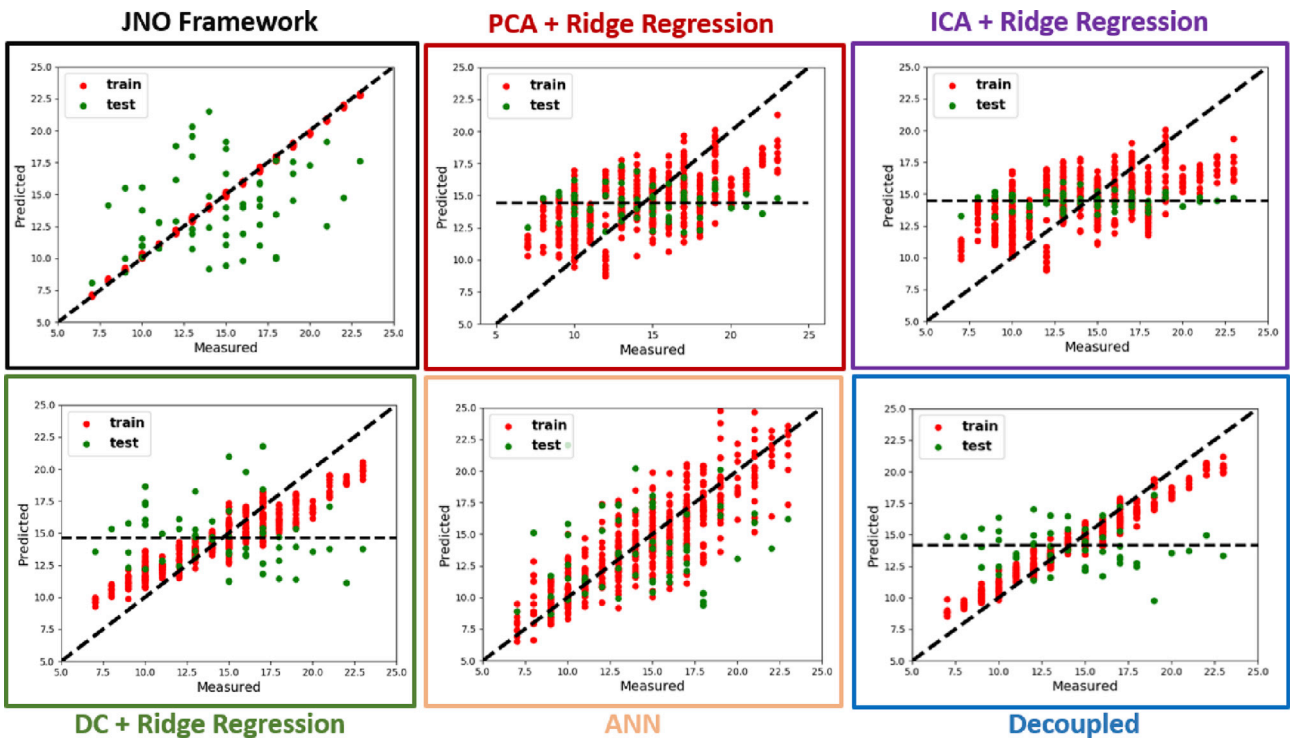

Fig. 9. KKI dataset: Prediction performance for the ADOS score for **Black Box:** JNO Framework. **Red Box:** PCA and ridge regression **Purple Box:** ICA and ridge regression **Green Box:** Node degree centrality and ridge regression **Orange Box**: ANN on correlation features **Blue Box:** Decoupled matrix cactorization and ridge Regression.

3.4. Performance on real world data

Fig. 9 – 11 compare the performance of our method against the baselines described in Section 2.4 for the prediction of ADOS, SRS and Praxis respectively for the KKI dataset. Similarly, Fig. 12 and Fig. 13 illustrate the performance comparison on the NYU dataset for ADOS and SRS respectively. We plot the score predicted by the algorithm on the y-axis against the measured ground truth score on the x-axis. The bold $x = y$ line indicates ideal performance. The red points correspond to training data, while the green points represent the held out testing data for all the folds in the cross validation. Our method is indicated at the top left corner in each plot. We observe that, although the training performance of the baselines is good (i.e. the red points follow the $x = y$ line), the JNO achieves the best training performance in all cases. Furthermore, we notice that all the two stage baseline testing performances track the mean value of the held out data (indicated by the black horizontal line). Our method clearly outperforms the baselines and is able to capture a trend in the data, beyond a mean value estimation in case of both datasets for all scores. This can be observed by the spread of the green points about the $x = y$ line in the case of the JNO method. Through our experiments, we noticed that the testing performance of the ANN is dependent on the choice of architecture. For example, the architecture chosen in Section 2.4 performs well on predicting ADOS for the KKI dataset, but performs poorly on all other comparisons. Our empirical evaluations could not identify a single architecture that performed well in all cases, like our JNO framework. The failure of the two stage decomposition in the bottom right comparison figures strengthens our hypothesis that a joint modeling of the neuroimaging and behavioral data is necessary in the context of generalization onto unseen data.

The lackluster generalization performance of the baselines is testament to the difficulty of the task at hand. The number of connections or features available to us are of the order of a 6670 dimensional vector representation for 58 or 63 patients. Both the machine learning and graph theoretic techniques we selected for a comparison are well known in literature for being able to robustly provide compact characterizations for high dimensional datasets. However, we see that PCA and ICA are unable to estimate a reliable projection of the data that is particularly indicative of clinical severity. Similarly, the node degree measure heavily rely on being able to accurately identify informative network topologies from the observed correlation matrices. However, its aggregate nature captures general trends and is not successful in characterizing subtle patient level differences. The failure of the decoupled matrix factorization and ridge regression makes a strong case for including the regression term as a part of our JNO objective. The basis directions obtained in this case are not indicative of clinical severity, due to which the regression performance suffers. Despite sweeping parameters across several orders of magnitude, we observe that the baselines are only good at capturing group level information, as is indicated by the training fit. However, they fail to characterize patient level differences for an unseen subject and simply predict the mean of the given cohort.

On the other hand, the generalization power of the ANN is contingent on the model order choice. This is demonstrated by its inability to perform well on comparisons outside of ADOS for the KKI dataset. Said another way, we have to change the network architecture for different severity measures across datasets. This is a major computational disadvantage when compared with our method.

A key difference between the JNO framework and the baselines is that we utilize the structure of the correlation matrices to guide the predictive model. In essence, we optimize for the tradeoff between the neuroimaging and behavioral data representations jointly, instead of posing it as a two stage problem. The matrix decomposition we employ explicitly models the group information through the basis, and the patient differences through the coefficients. The limited number of basis elements we employ to decompose the data provides us with compact representations which explain the connectivity information well. The regularization terms and constraints ensure that the problem is well posed, while providing clinically meaningful and informative representations about the data. We also quantify the performance indicated in these figures in Table 1 (KKI dataset) and Table 2 (NYU dataset) based on the validation metrics mentioned earlier.

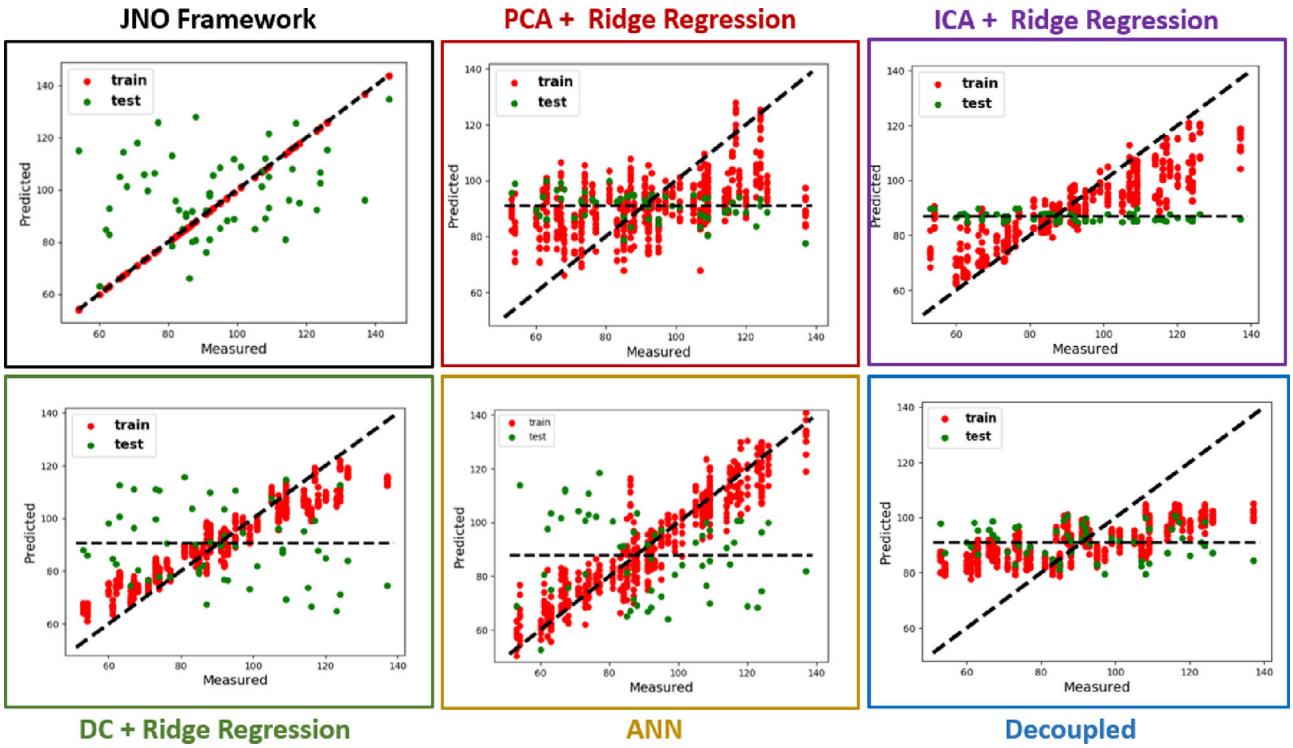

Fig. 10. KKI dataset: Prediction performance for the SRS score for **Black Box:** JNO Framework. **Red Box:** PCA and ridge regression **Purple Box:** ICA and ridge regression **Green Box:** Node degree centrality and ridge regression **Orange Box**: ANN on correlation features **Blue Box:** Decoupled matrix factorization and ridge regression.

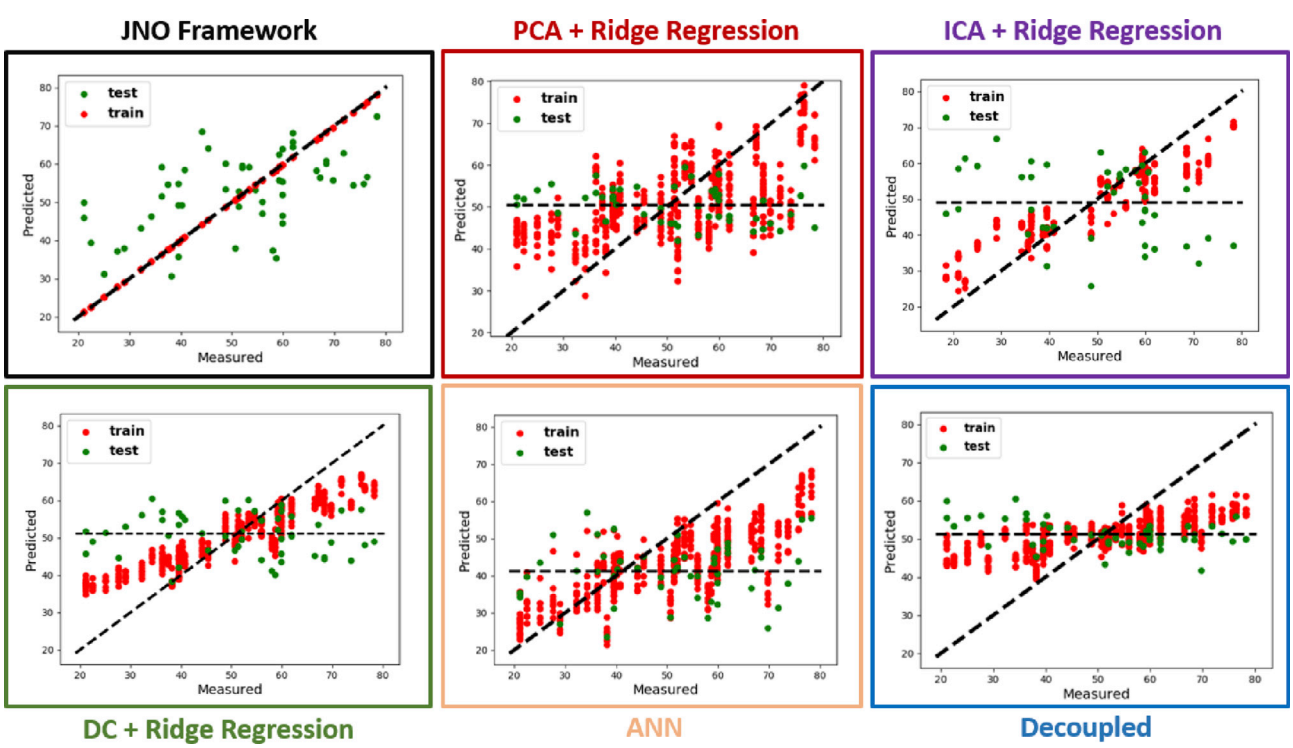

Fig. 11. KKI dataset: Prediction performance for the Praxis score for **Black Box:** JNO Framework. **Red Box:** PCA and ridge regression **Purple Box:** ICA and ridge regression **Green Box:** Node degree centrality and ridge regression **Orange Box**: ANN on correlation features **Blue Box:** Decoupled matrix factorization and ridge regression.

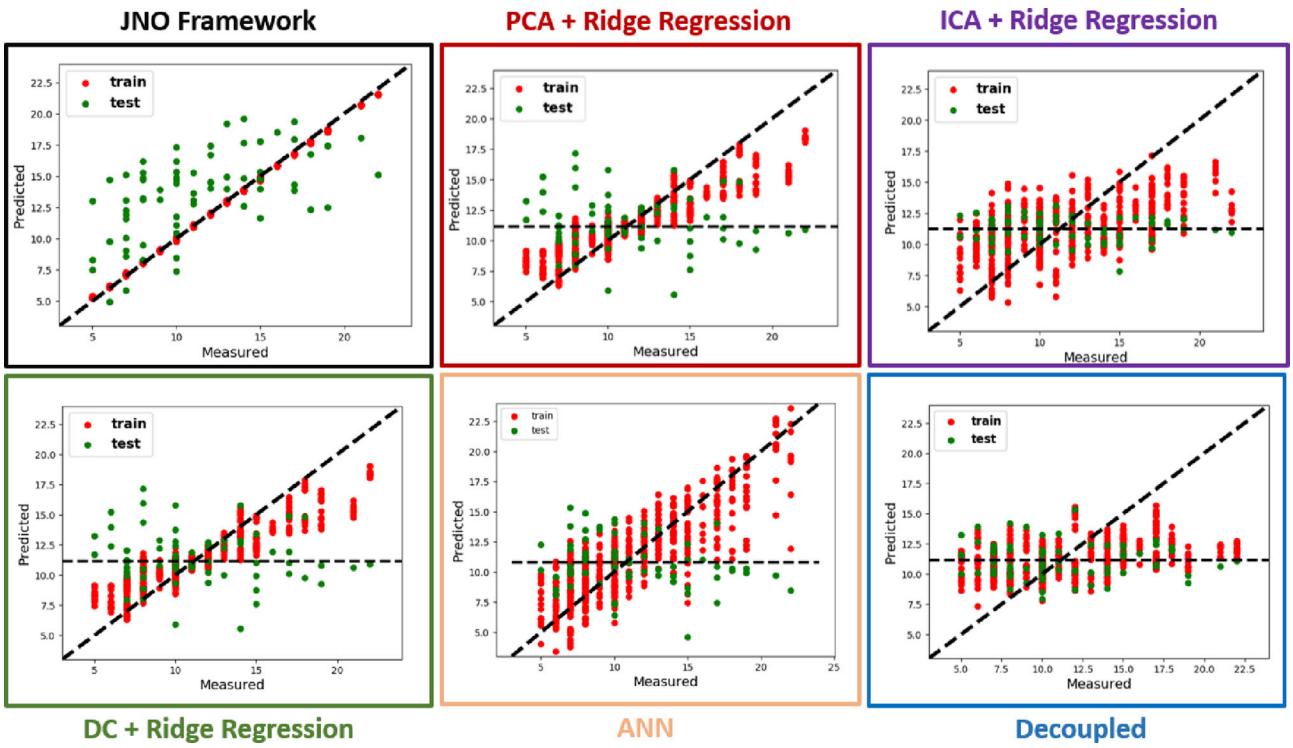

Fig. 12. NYU dataset: Prediction performance for the ADOS score for **Black Box:** JNO Framework. **Red Box:** PCA and ridge regression **Purple Box:** ICA and ridge regression **Green Box:** Node degree centrality and ridge regression **Orange**: ANN on correlation features **Blue Box:** Decoupled matrix factorization and ridge regression.

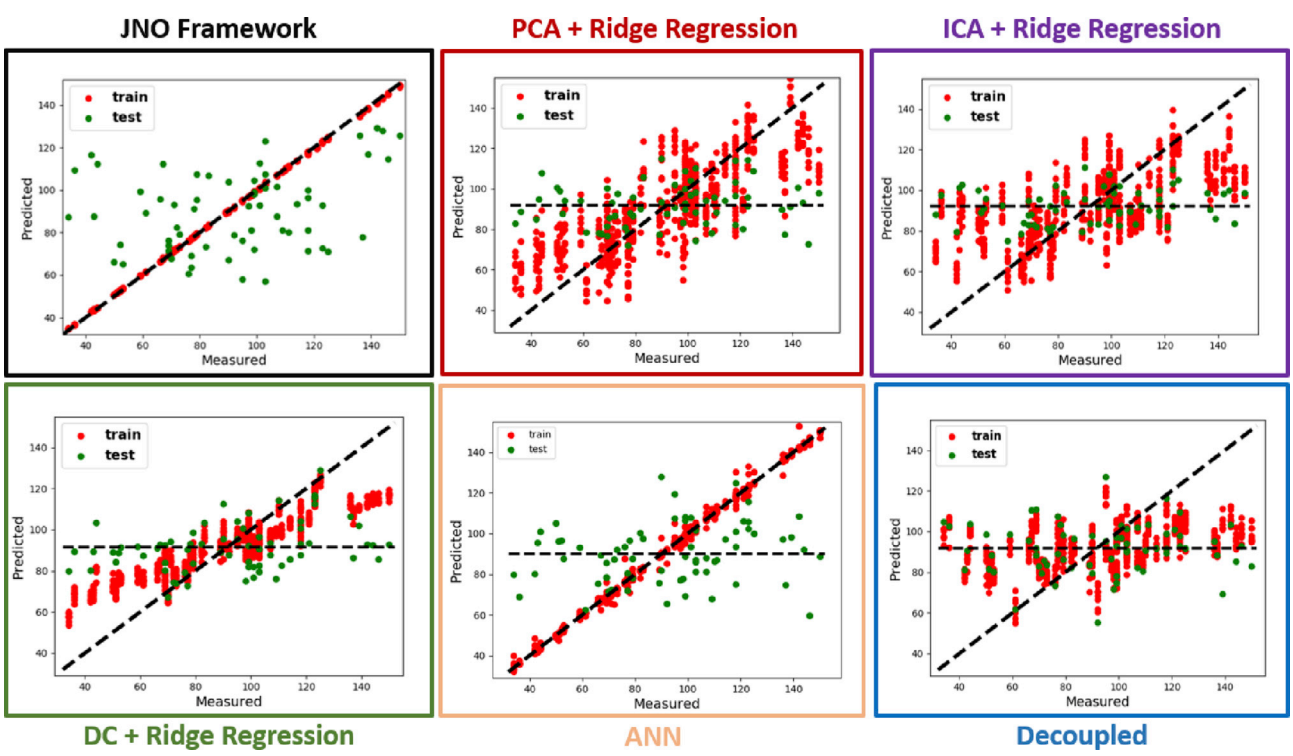

Fig. 13. NYU dataset: Prediction performance for the SRS score for **Black Box:** JNO Framework. **Red Box:** PCA and ridge regression **Purple Box:** ICA and ridge regression **Green Box:** Node degree centrality and ridge regression **Orange**: ANN on correlation features **Blue Box:** Decoupled matrix factorization and ridge regression.

Table 1
KKI Dataset: Performance evaluation using **Median Absolute Error (MAE)** and **Normalized Mutual Information (NMI)** fit, both for testing & training. Lower MAE & higher NMI score indicate better performance. We have highlighted the best performance in bold. The **CDF-KS** column indicates the Kolmogorov-Smirnoff statistic on the CDF comparison against our method. The instances highlighted in bold indicate that the performance is within the accepted 0.05 threshold. Near misses have been underlined.

Score	Method	MAE Train	MAE Test	NMI Train	NMI Test	CDF-KS
ADOS	PCA & ridge	2.18 ± 2.2	2.99 ± 1.71	0.22	0.18	0.06
	ICA & ridge	2.13 ± 1.1	3.01 ± 1.90	0.31	0.23	**0.027**
	D_N & ridge	1.22 ± 0.91	3.68 ± 2.53	0.45	0.39	**0.015**
	ANN	2.68 ± 2.21	**2.28±1.30**	0.91	**0.58**	0.088
	Decoupled	2.36 ± 2.33	2.63 ± 1.90	0.15	0.30	0.083
	JNO Framework	**0.088±0.13**	2.53 ± 1.86	**0.99**	0.52	–
SRS	PCA & ridge	12.92 ± 10.48	19.09 ± 12.48	0.64	0.39	**0.032**
	ICA & ridge	7.96 ± 6.35	20.8 ± 17.3	0.83	0.63	**0.041**
	D_N & ridge	5.77 ± 4.88	19.63 ± 17.23	0.85	0.59	0.089
	ANN	4.77 ± 4.09	21.25 ± 14.63	0.81	0.56	0.093
	Decoupled	12.06 ± 10.04	18.5 ± 16.4	0.74	0.37	**0.014**
	JNO Framework	**0.13±0.07**	**13.27±10.85**	**0.99**	**0.78**	–
Praxis	PCA & ridge	9.44 ± 6.83	12.83 ± 8.84	0.64	0.37	0.17
	ICA & ridge	4.79 ± 4.17	13.08 ± 13.07	0.73	0.63	**0.035**
	D_N & ridge	4.78 ± 3.24	13.93 ± 8.14	0.68	0.56	**0.017**
	ANN	9.34 ± 7.21	14.90 ± 10.06	0.69	0.39	**0.01**
	Decoupled	10.17 ± 7.96	13.24 ± 10.38	0.68	0.29	0.10
	JNO Framework	**0.11±0.065**	**10.18±6.58**	**0.99**	**0.79**	–

Table 2
NYU Dataset: Performance evaluation using **Median Absolute Error (MAE)** and **Normalized Mutual Information (NMI)** fit, both for testing & training. Lower MAE & higher NMI score indicate better performance. We have highlighted the best performance in bold. The **CDF-KS** column indicates the Kolmogorov-Smirnoff statistic on the CDF comparison against our method. The instances highlighted in bold indicate that the performance is within the accepted 0.05 threshold. Near misses have been underlined.

Score	Method	MAE Train	MAE Test	NMI Train	NMI Test	CDF-KS
ADOS	PCA & ridge	1.68 ± 1.53	3.46 ± 2.21	0.30	0.28	**0.0019**
	ICA & ridge	2.75 ± 1.88	3.41 ± 2.34	0.15	0.17	0.068
	D_N & ridge	1.18 ± 1.19	3.17 ± 3.05	0.50	0.39	**0.0025**
	ANN	1.40 ± 1.39	3.36 ± 2.89	0.31	0.28	0.081
	Decoupled	2.62 ± 2.54	3.32 ± 2.27	0.21	0.12	0.07
	JNO Framework	**0.10±0.088**	**2.63±2.51**	**0.99**	**0.54**	–
SRS	PCA & ridge	10.64 ± 12.60	18.22 ± 12.43	0.87	0.54	0.24
	ICA & ridge	16.88 ± 15.71	18.11 ± 13.9	0.71	0.49	0.19
	D_N & ridge	7.32 ± 5.76	23.18 ± 18.44	0.87	0.66	0.081
	ANN	1.50 ± 1.39	19.04 ± 17.69	0.85	0.08	**0.009**
	Decoupled	16.42 ± 15.66	22.43 ± 18.79	0.80	0.42	0.06
	JNO Framework	**0.46±0.36**	**16.61±12.43**	**0.96**	**0.72**	–

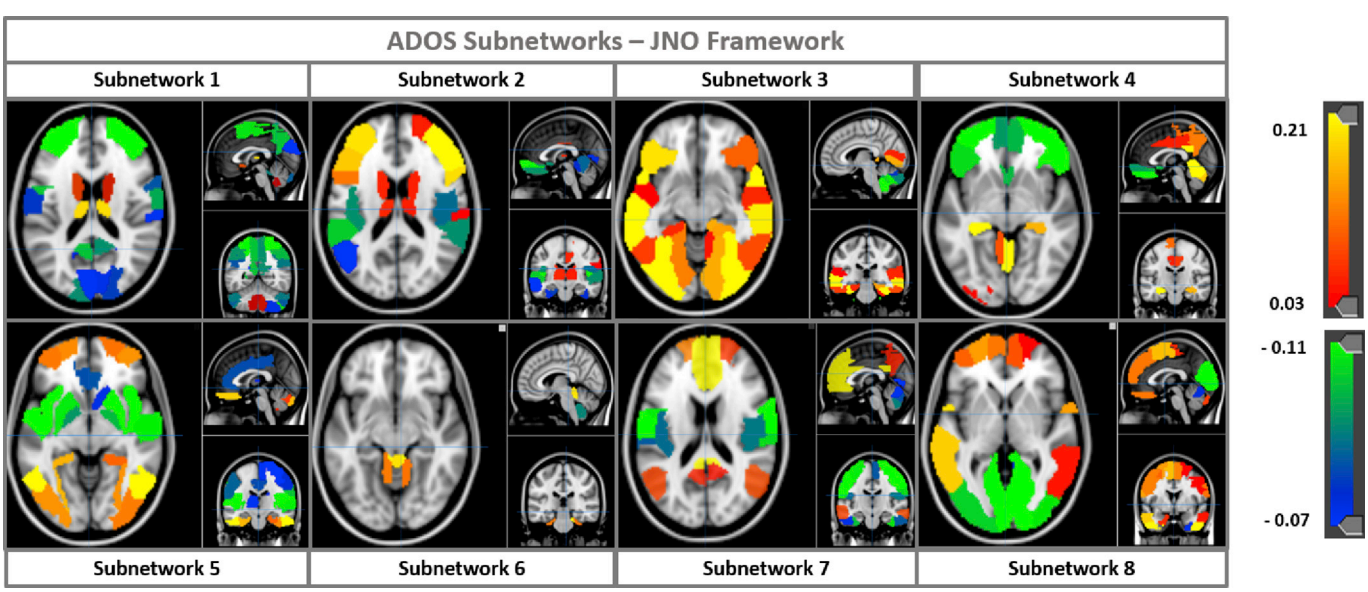

Fig. 14. Subnetworks estimated to predict the ADOS score by the JNO. Regions having negative contributions are anti-correlated with areas having positive values.

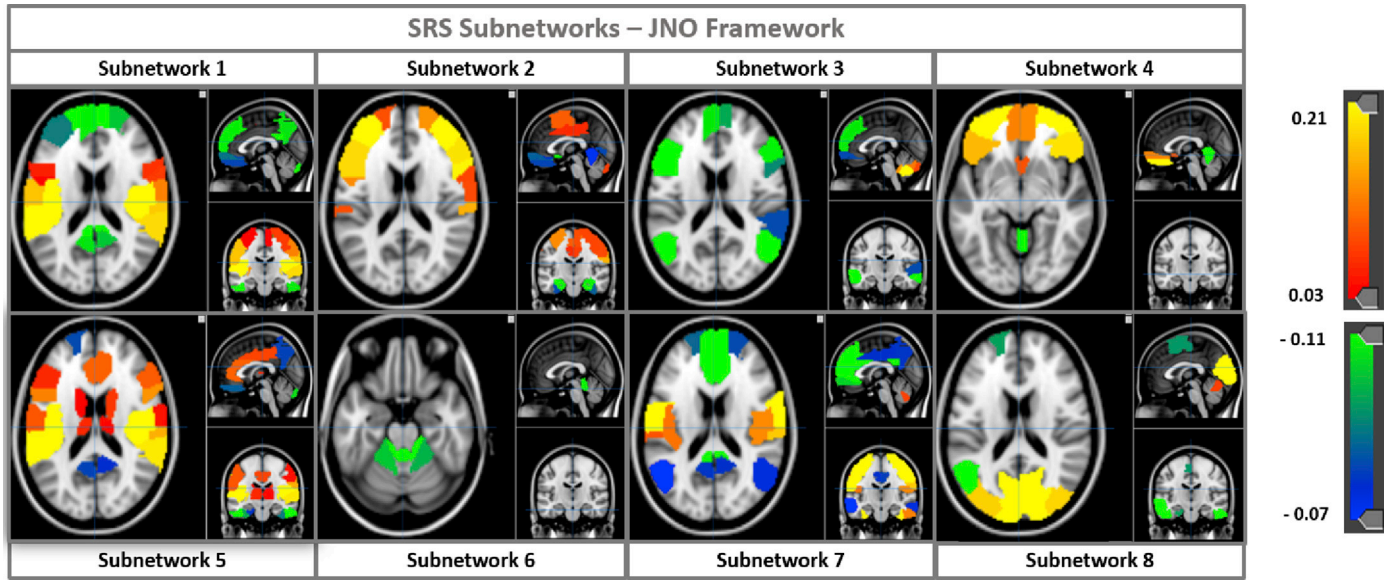

Fig. 15. Subnetworks estimated to predict the SRS score. Regions having negative contributions are anti-correlated with areas having positive values.

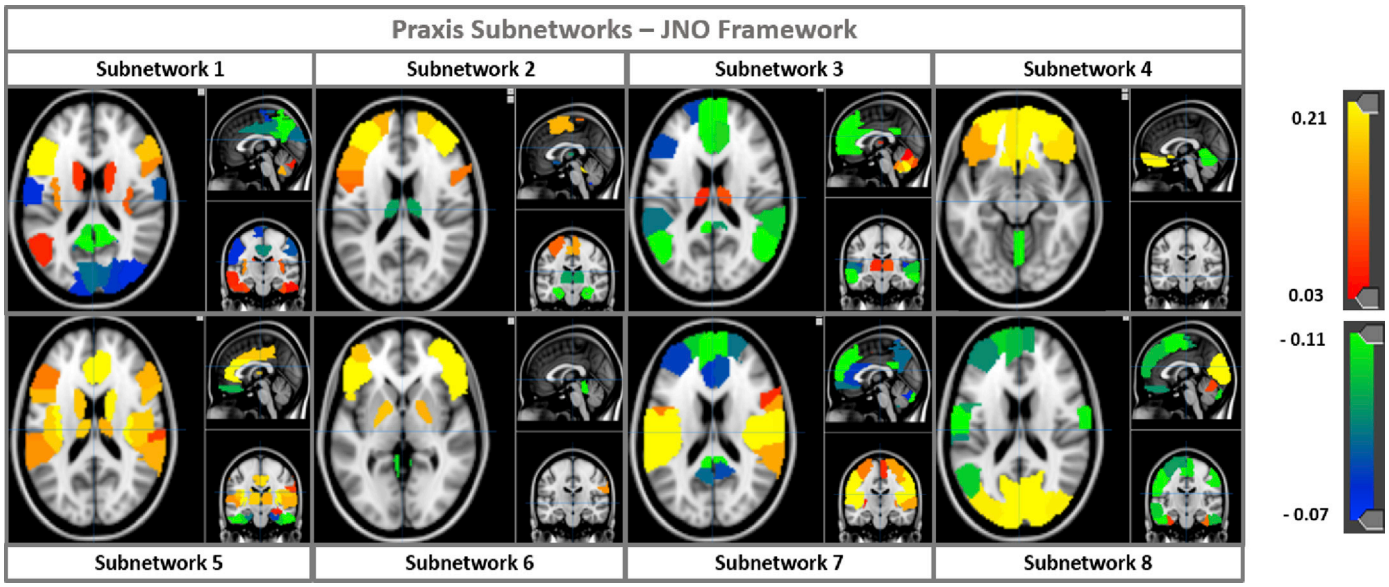

Fig. 16. Subnetworks estimated to predict the Praxis score. Regions having negative contributions are anti-correlated with areas having positive values.

3.5. Subnetwork identification

Figs. 14–16 illustrate the subnetworks in **B**, as trained on the ADOS, SRS and Praxis in the KKI dataset, respectively. Since each column of the basis corresponds to a set of co-activated subregions, we plot the normalized values stored in these columns onto the corresponding AAL ROIs. The colorbar indicates subnetwork contribution to the AAL regions. Regions colored as negative values are anticorrelated with regions storing positive ones. We rank the 8 subnetworks obtained from SRS and Praxis according to their overlap with the subnetworks from ADOS. As seen from these figures, corresponding subnetworks show considerable overlap in regional co-activation patterns. The individual variations can arise from the fundamental differences in the behavioral traits that each score is trying to capture.

From a clinical standpoint, Subnetwork 7 includes competing i.e. anticorrelated contributions from regions of the default mode network (DMN) and somatomotor network (SMN). Abnormal connectivity within the DMN and SMN has been previously reported in ASD [Lyn+13; Neb+16]. Subnetwork 5 comprises of competing contributions from SMN regions. Additionally, it includes higher order visual processing areas in the occipital and temporal lobes, which is consistent with behavioral reports of reduced visual-motor integration in the ASD literature [Neb+16]. Subnetwork 1 has competing from prefrontal and subcortical contributions, mainly the thalamus, amygdala and hippocampus. The thalamus is responsible for relaying sensory and motor signals to the cerebral cortex in the brain. The hippocampus is known to play a crucial role in the consolidation of long and short term memory, along with spatial memory to aid navigation. Altered memory functioning has been shown to manifest in children diagnosed with ASD [WGM06]. Along with the amygdala, which is known to be associated with emotional responses, these areas may be crucial for social-emotional regulation in ASD. Finally, Subnetwork 2 is comprised of competing contributions from the central executive control network and the insula, which is thought to be critical for switching between self-referential and

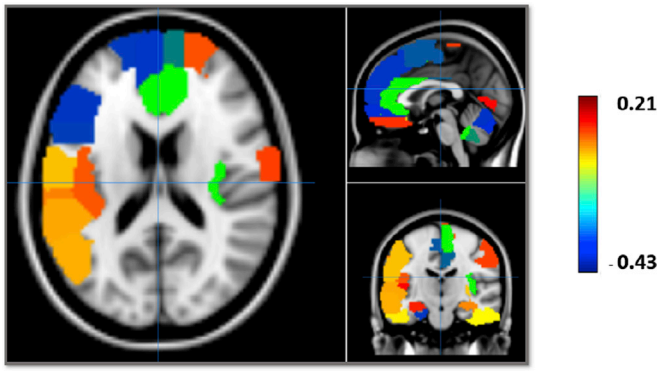

Fig. 17. Representation learned from the prediction of ADOS by Node degree centrality + ridge regression. The colorbar indicates the weight of the ROI assigned by the ridge regression.

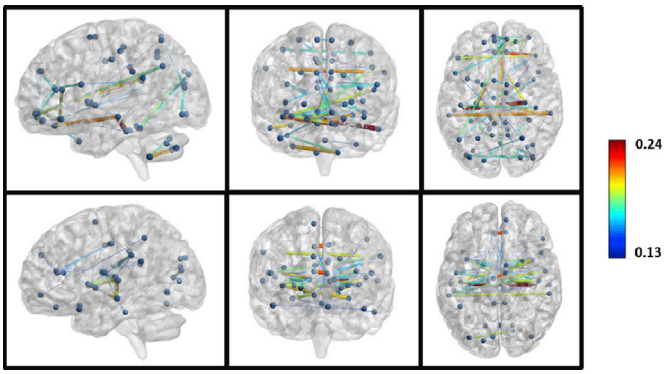

Fig. 18. Top two subnetworks identified by the prediction of the ADOS score by PCA + ridge regression. The colorbar indicates the weight of the connection.

goal-directed behavior [SLM08].

3.5.1. Robustness in subnetwork recovery

Notice that we estimate a different basis matrix **B** for each cross validation fold. Therefore, one important property to verify is that these subnetworks are similar across different cohorts of the data.

We observed an average similarity of 0.79 ± 0.06 for the ADOS networks, 0.86 ± 0.04 for the SRS networks, and 0.76 ± 0.06 for the Praxis networks across their cross validation runs. Additionally, upon a cross comparison between the ADOS and SRS networks, we obtained an average similarity of 0.82 ± 0.07. Similarly, the overlap between ADOS and Praxis is 0.79 ± 0.04, and between SRS and Praxis is 0.77 ± 0.06. For a convenient visual inspection, we have arranged the networks in Fig. 15 (SRS) and Fig. 16 (Praxis) in the order of their inner product similarity with the ADOS networks in Fig. 14. This finding strengthens the hypothesis that our model is successful at capturing the stable underlying mechanisms which explain the different sets of deficits of the disorder.

3.5.2. Comparing subnetwork representations

In this section, we compare the subnetworks identified by the JNO to the representations learned by the baseline methods. Recall that we have used a regularized linear regression as the Stage 2 predictor for the baselines. Therefore, we can probe the learned regression weights to characterize the baseline network representations.

Degree centrality looks at the relative importance of each brain region or 'node' to the overall representation. To visualize the pattern identified by the degree centrality + ridge regression baseline, we display the regression weights on the brain surface plots in Fig. 17, normalized to unit norm. The colorbar indicates the strength of co-activation. Regions storing negative values are anticorrelated with regions storing positive

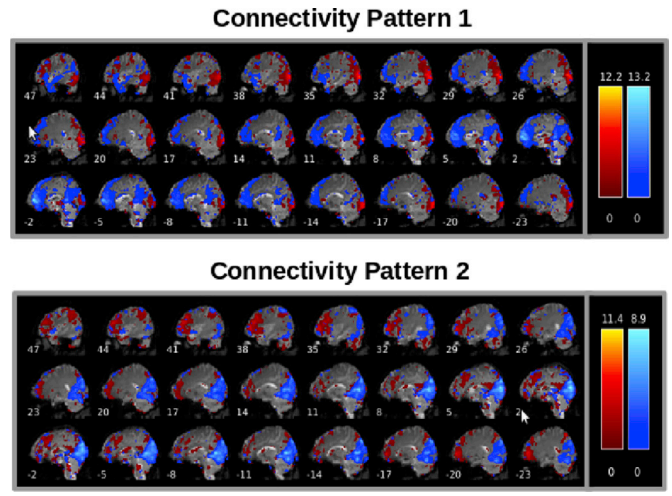

Fig. 19. Connectivity patterns identified as important in the prediction of the ADOS score by ICA + ridge regression. Each plot displays 2 spatial components contributing to the correlation feature. The colorbar indicates the weight of the connection.

weights. We again observe patterns from the DMN in the subnetwork plot. Note that the DMN was also a key connectivity pattern identified by the JNO. However, several other subnetworks identified by the JNO do not figure in this representation.

On the other hand, for the PCA + ridge regression baseline, the regression weights inform us of the relative importance of the principal components in prediction. Since the features fed into PCA are the $M = (P \times (P-1))/2$ correlation values, we are left with a 6670 dimensional edge connectivity representation for the AAL per component. We first examine the absolute value of the regression weights learned, and then display the connectivity in the top 2 basis components in Fig. 18. We render this connectivity measure using the BrainNet Viewer [XWH13] software. For clarity, we have chosen to display the top 5 percent of the connections obtained. The solid edges signify retained connections, while the blue spheres correspond to nodes of the AAL regions. The colorbar to the right indicates the strength of the connections. We notice that the components consist of several crossing connections spread across different regions of the brain. As compared to our model, which pinpoints key subnetworks already known to be associated with ASD, the representation obtained is not immediately interpretable.

In the ICA + ridge regression baseline, the input to the regression model are the correlation values between the components identified by ICA. After the model is fit, we sort the input correlations based on the learned regression weights. This helps us identify the features important for prediction. In Fig. 19, we display the spatial maps of the top 2 connections identified by the algorithm. We again, observe patterns from the DMN and visual areas. However, it fails to capture several other subnetwork patterns that the JNO identifies as important for ASD.

For the ANN, we use the weight matrix learned at the input layer to

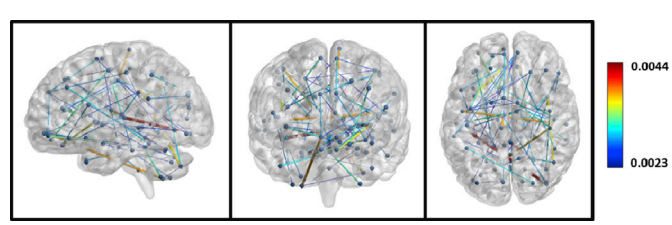

Fig. 20. Connectivity patterns identified in the prediction of the ADOS score by the ANN. The colorbar indicates the weight of the connections. The narrow range of values are indicative that the ANN assigns equal weighting to most connections on an average.

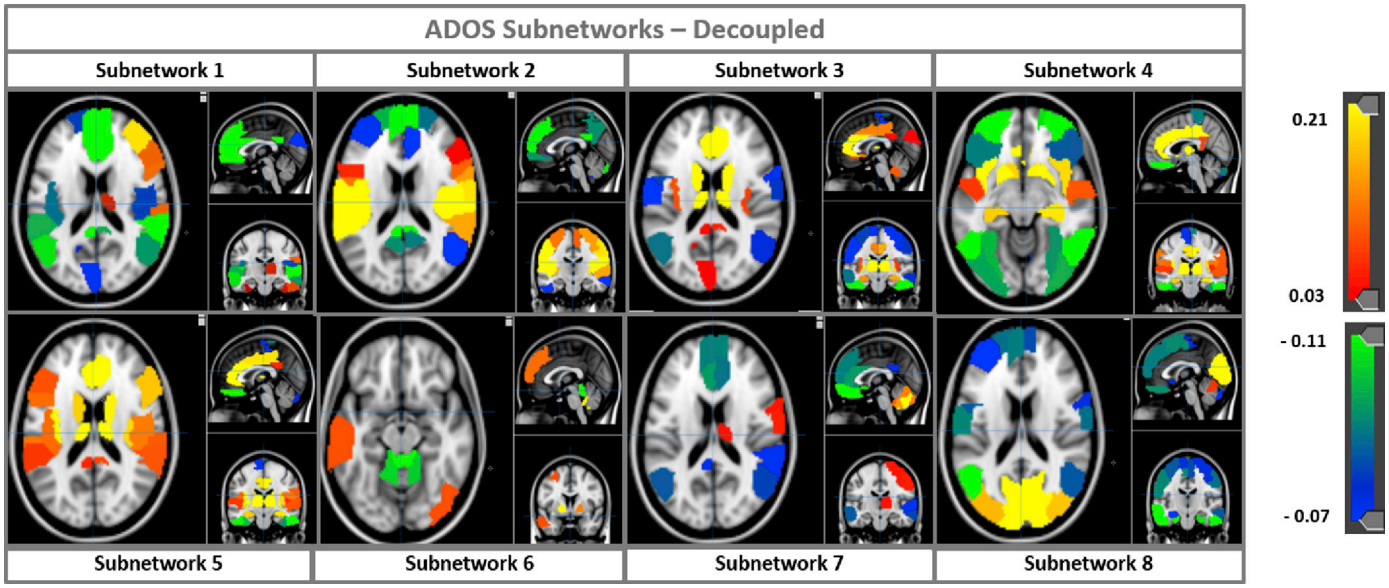

Fig. 21. Subnetworks estimated to predict ADOS score by decoupling the matrix decomposition and ridge regression. Regions having negative contributions are anti-correlated with areas having positive values.

inform us of the subnetwork connectivity. Recall that this matrix is of dimension $M \times D$, where $M = (P \times (P-1))/2 = 6670$ for the AAL atlas. For our application, D is of width 8000. We first take the absolute values of these weights, and then normalize the columns of this matrix to unit norm. We then average across the rows to obtain a single 6670 dimensional edge-edge connectivity vector. Again, we use the BrainNet connectivity plots to display this information in Fig. 20. We have chosen to display the top 1 percent of the connections obtained. The solid edges signify retained connections, while the blue spheres correspond to nodes of the AAL regions. The colorbar to the right indicates the strength of the connections. We observe several overlapping connectivity patterns spread across the entire brain despite applying a stringent threshold. Additionally, the narrow range of values indicates that the ANN assigns nearly equal weight to all connections on an average. Similar to the PCA baseline, this representation is unable to capture interpretable connectivity patterns which explain behavior.

Finally, we examine the representation learned by performing the matrix decomposition and prediction separately, i.e. the decoupled case. Note that the learned basis matrix **B** follows the same interpretation as that of the JNO. We display the corresponding co-activation patterns in Fig. 21. Again, the colorbar indicates the strength of activation of the AAL ROIs. Negative regions are anticorrelated with the positive regions. For convenience, we have ordered the 8 subnetworks according to their similarity with the ADOS subnetworks identified in Fig. 14. Since we use the same matrix decompotion as the JNO, we observe several similarities in the learned representations. We also notice subtle differences in the

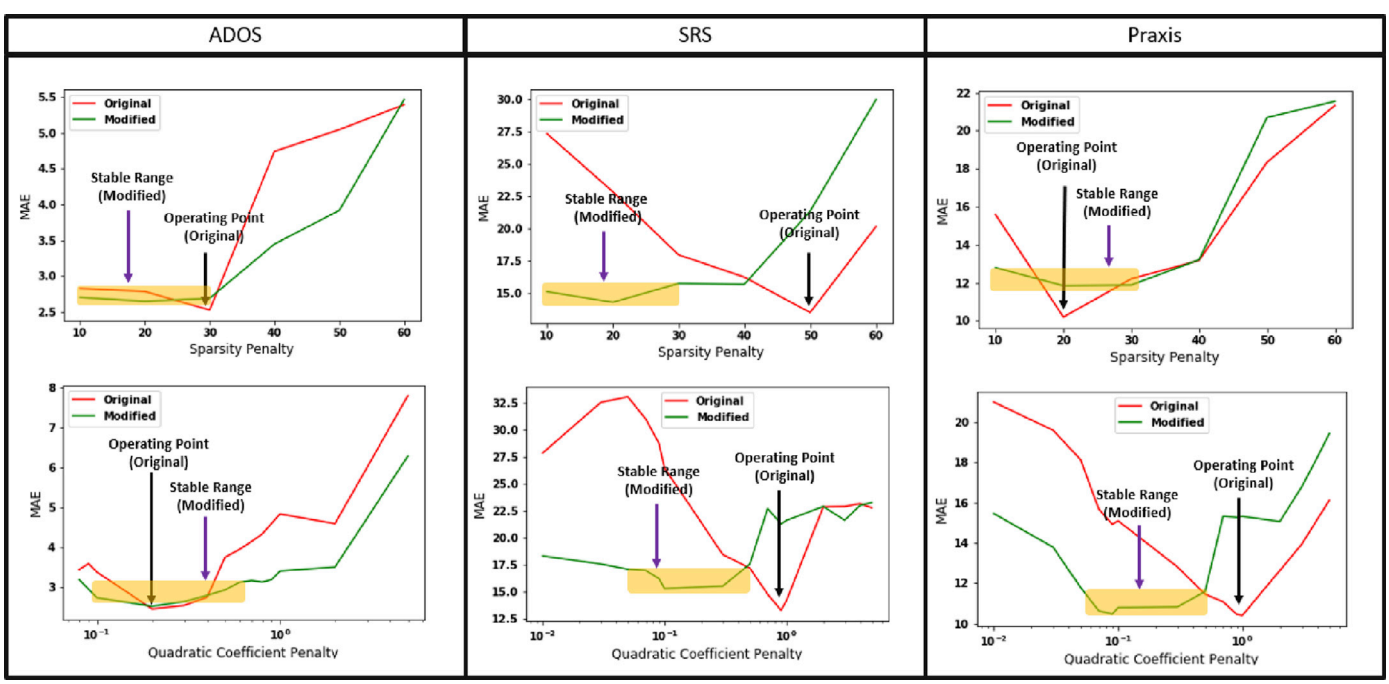

Fig. 22. Comparing the sensitivity of the JNO framework with the modified objective in Eq. (19). Prediction performance with varying **Top**λ_1 for **(L–R):** ADOS, SRS and Praxis **Bottom**λ_2 for **(L–R):** ADOS, SRS and Praxis.

patterns on account of the coupling with the predictive term in the JNO. We conjecture that these learned differences are what gives the JNO the leverage to generalize to unseen data.

3.6. Evaluating model generalizability

We have shown both predictive power and interpretabilitiy of our model thus far. Furthermore, characterizing model generalizability is important for future application of our framework. Here, we first examine the sensitivity of our prediction results with respect to the model hyperparameters. Then, we show robustness of our formulation to two common obstacles in generalizabilitiy for rs-fMRI analysis, namely choice of parcellation scheme and test-retest reliability. Lastly, we show mitigation strategies to handle hyperparameter sensitivity that make our framework more robust.

3.6.1. Hyperparameter sensitivity

As initially described in Section 3.4, our JNO framework is insensitive to the regression tradeoff γ and ridge penalty λ_3. We also have a natural way to set the number of subnetworks K. However, we observe that our JNO framework is fairly sensitive to the sparsity on $\mathbf{B}$ and the ridge penalty on the coefficients $\mathbf{c}_n$, i.e. λ_1 and λ_2 respectively. Fig. 22 represents the MAE recovery performance of the algorithm for varying settings of λ_1 and λ_2, holding the remaining parameter settings constant when evaluated on the KKI dataset. The red plots in each case indicate the performance of the JNO framework. The green plots allude to our mitigation strategy which we will present later in Subsection 3.6.3. The **x**-axis denotes the parameter value, while the **y**-axis quantifies the MAE from cross validation. Observe that the best λ_1 and λ_2 settings for the individual scores are different, i.e ADOS-$\{\lambda_1 = 30, \lambda_2 = 0.2\}$, SRS-$\{\lambda_1 = 50, \lambda_2 = 0.9\}$, and Praxis-$\{\lambda_1 = 20, \lambda_2 = 0.6\}$. Additionally, the kinks in the plots (shown by the black arrow) also indicate that small changes in the sparsity and coefficient regularization lead to a dramatic change in performance, i.e. the operating points for these two parameters are narrow. We suspect that the hyperparameter differences can be partially attributed to the different dynamic range of each clinical score. Specifically, these differences impact the tradeoff between the representation learning and prediction terms in the JNO optimization. This in turn affects the generalization performance at a particular hyperparameter setting. These observations further illustrate the difficulty of the problem we are trying to address. Keeping these subtleties in mind, the next subsection focuses exclusively on studying the generalization performance of the JNO and the baselines at the optimal hyperparameter settings we obtained in Section 3.4.

3.6.2. Examining test-retest performance

Based on the observations above, we design two experiments to characterize the generalizability of the JNO with regards to its free parameters, i.e. λ_1 and λ_2, in a test-retest setting. The first of these experiments investigates the impact of varying the rs-fMRI parcellation scheme, which not only changes the dimensionality of the input correlation matrices, but also changes the distribution of correlation measures. The second experiment is a cross site comparison between KKI and NYU where the same parcellation (i.e. AAL) is maintained. However, site differences like acquisition protocol and pre-processing, impact the distribution of the input correlations. Overall, these two experiments validate generalizabilitiy and robustness of our model, as we show good performance under sub-optimal parameter settings.

For the first experiment, we compute correlation matrices for the KKI rs-fMRI dataset using the Brainetome-246 atlas. Note that the Brainetome [Fan, 2016] provides a much finer ROI resolution as compared to the AAL-116 atlas used previously. We then predict the clinical scores from the new correlation matrices using the AAL hyperparameter settings. This is repeated for all the baselines along with the JNO. The plots in Figs. 23–25 illustrate the prediction of ADOS, SRS and Praxis respectively by the JNO. In each figure, the left plots denote the Brainetome

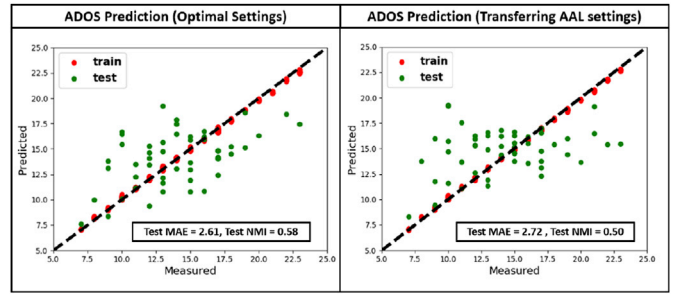

Fig. 23. Brainetome Parcellation: A performance comparison for ADOS prediction by the JNO using the **(L)** Optimal Settings for Brainetome **(R)** Transferring AAL settings.

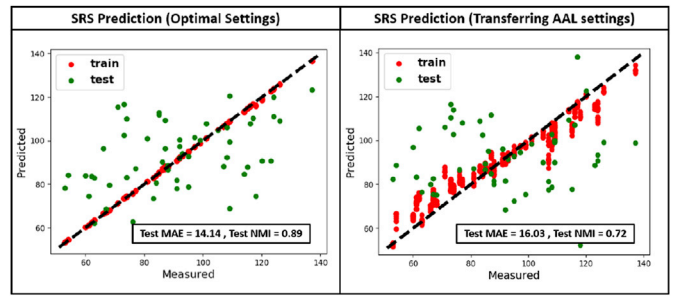

Fig. 24. Brainetome Parcellation: A performance comparison for SRS prediction by the JNO using the **(L)** Optimal Settings for Brainetome **(R)** Transferring AAL settings.

performance using the hyperparameter selection outlined in Section 3.4, while the right plots indicate the performance after transferring the hyperparameter settings from the AAL experiments. Notice that the performance is slightly worse than the case where they were explicitly set by cross validation (left plots). This is expected since we have not explicitly optimized for the new data characteristics. Table 3 provides the corresponding quantitative metrics. We emphasize that the performance trends are very similar to those in Section 3.4, namely, we outperform nearly all the baselines, while retaining the interpretability of the subnetwork decomposition. Since the ROI resolution differs here, we can treat this comparison as a surrogate for out of sample generalization. Additionally, this experiment suggests that the JNO is agnostic to the parcellation scheme used. We have included additional baseline comparisons in the supplementary results document. We believe that the joint optimization helps us balance the tradeoff between the representation learning and prediction terms. Said another way, the matrix factorization term regularizes the problem and helps provide stability while handling the changes in the input data distribution. As a result, we are able to

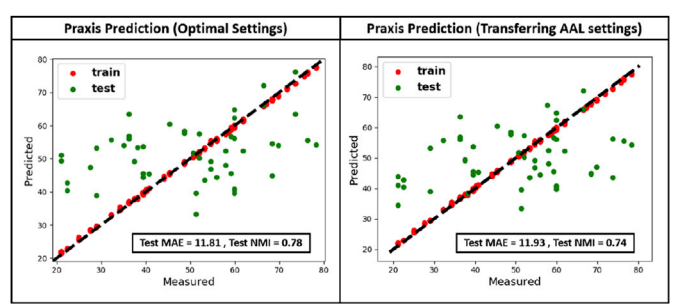

Fig. 25. Brainetome Parcellation: A performance comparison for Praxis prediction by the JNO using the **(L)** Optimal Settings for Brainetome **(R)** Transferring AAL settings.

Table 3
KKI Dataset on Brainetome parcellation: Performance evaluation using **Median Absolute Error (MAE)** and **Normalized Mutual Information (NMI)** fit, both for testing & training. Lower MAE & higher NMI score indicate better performance. We have highlighted the best performance in bold. The **CDF-KS** column indicates the Kolmogorov-Smirnoff statistic on the CDF comparison against our method. The instances highlighted in bold indicate that the performance is within the accepted 0.05 threshold. Near misses have been underlined. The parameter settings used were the same as those for the KKI dataset from the main manuscript.

Score	Method	MAE Train	MAE Test	NMI Train	NMI Test	CDF-KS
ADOS	PCA & ridge	1.19 ± 1.54	3.11 ± 2.99	0.30	0.22	0.061
	ICA & ridge	1.48 ± 1.11	3.13 ± 3.01	0.58	0.43	0.08
	D_N & ridge	2.26 ± 1.88	**2.41±2.69**	0.53	0.33	0.059
	Decoupled	1.44 ± 1.33	3.15 ± 2.96	0.41	0.32	0.23
	JNO Framework	**0.10±0.096**	2.72 ± 2.34	**0.99**	**0.50**	–
SRS	PCA & ridge	6.73 ± 5.94	19.62 ± 14.30	0.84	0.46	0.19
	ICA & ridge	8.44 ± 7.19	17.43 ± 11.17	0.80	0.55	**0.02**
	D_N & ridge	7.11 ± 6.24	24.43 ± 19.18	0.85	0.62	0.083
	Decoupled	8.19 ± 7.95	19.16 ± 14.13	0.84	0.46	0.08
	JNO Framework	**2.11±1.19**	**16.03±14.58**	**0.91**	**0.72**	–
Praxis	PCA &ridge	6.91 ± 6.04	13.67 ± 8.36	0.79	0.51	0.063
	ICA & ridge	7.96 ± 5.99	13.12 ± 9.36	0.80	0.51	**0.012**
	D_N & ridge	6.41 ± 5.99	13.21 ± 7.96	0.81	0.49	0.09
	Decoupled	11.56 ± 11.78	17.96 ± 16.11	0.69	0.48	0.068
	JNO Framework	**0.36±0.27**	**11.93±9.05**	**0.95**	**0.74**	–

transfer the learned hyperparameter settings, yet outperform the baselines at severity prediction.

In our second experiment, we again use the hyperparameter settings learned from the KKI dataset, but this time, predict clinical scores from the NYU rs-fMRI correlation matrices. The parcellation scheme i.e. AAL, is maintained for both datasets. Figs. 26 and 27 compare these predictions (right sub-plots) with those from Section 3.4 (left sub-plots) for ADOS and SRS respectively. Table 4 delineates the quantitative comparisons against the baselines. Again, we observe that the JNO still outperforms the baselines, though there is a slight reduction in quantitative performance. This reduction is expected, especially given the differences in scanning protocols across sites. As in the previous experiment, we are able to use the transferred parameter settings, and still optimize for a representation which is predictive of clinical severity. The cross dataset comparison is a good indicator of the robustness of the JNO to its hyperparameters. The observations from both experiments suggest that the JNO has the potential to capture robust and interpretable phenomenon related to the neurological disorder of interest.

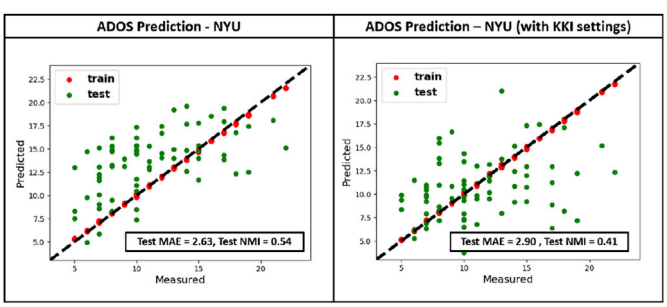

Fig. 26. A performance comparison for ADOS prediction on the NYU Dataset by the JNO using the settings learned from cross validation on the **(L)** NYU **(R)** KKI dataset.

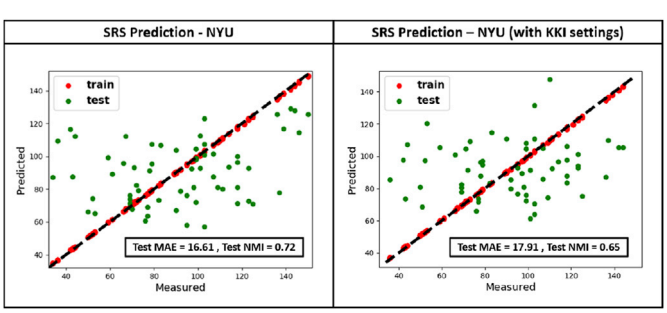

Fig. 27. A performance comparison for SRS prediction on the NYU Dataset by the JNO using the settings learned from cross validation on the **(L)** NYU **(R)** KKI dataset.

Lastly, both the test-retest experiments study the overall generalizability of the method by examining a form of out of sample prediction, either by varying the parcellation scheme, or via a cross dataset comparison. This is very similar to the principles of nested cross validation, which aims to reduce bias in parameter selection.

3.6.3. *Mitigating parameter sensitivity*

Finally, we propose two main modifications to tackle the observed hyperparameter sensitivity in λ_1 and λ_2. Given that the dynamic ranges of the scores are quite different and potentially impact generalization, our first mitigation strategy is to rescale the measures to a fixed interval. Since ADOS is the most widely accepted observational measure of clinical autism severity, we have scaled and offset the remaining scores to have a range of 0–30 (similar to ADOS). To mitigate the narrow 'operating point', we include an extra template average correlation term in Eq. (5). We now model the residual outer-product terms as deviations around a mean template correlation matrix $\mathbf{B}_{avg}$. The rationale behind this additional term is that it encourage sparsity in the basis matrix along with the explicit ℓ_1 penalty. The modified objective is as follows:

$$\begin{aligned}\mathscr{J}\left(\mathbf{B}, \mathbf{B}_{avg}, \mathbf{C}, \mathbf{w}\right) = \sum_n \left|\left|\mathbf{\Gamma}_n - \mathbf{B}_{avg} - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T\right|\right|_F^2 \\ +\gamma\left|\left|\mathbf{y} - \mathbf{C}^T\mathbf{w}\right|\right|_2^2 + \lambda_1\left|\left|\mathbf{B}\right|\right|_1 \\ +\lambda_2\left|\left|\mathbf{C}\right|\right|_F^2 + \lambda_3\left|\left|\mathbf{w}\right|\right|_2^2 \quad s.t. \quad \mathbf{c}_{nk} \geq 0,\end{aligned} \tag{19}$$

Notice that $\mathbf{B}_{avg}$ has a closed form update, which does not add much computational overhead. The updates for the remaining variables follow the same procedure as described in Section. 2.2, except that the term, $\{\mathbf{\Gamma}_n\}$ is replaced with $\{\mathbf{\Gamma}_n - \mathbf{B}_{avg}\}$ in every update.

The green plots in Fig. 22 illustrate the cross validated performance of the modified JNO framework from Eq. (19). The operating point $\{\lambda_1, \lambda_2\}$ for the modified framework is fairly consistent across the scores. Moreover, the green plots exhibit a larger stable range (highlighted in yellow) compared to the red plots. Accordingly, we identify the settings $\{\lambda_1 = 10 - 30, \lambda_2 = 0.08\text{–}0.6\}$ as the operating range for the modified JNO objective, which is roughly an order of magnitude larger than the original formulation and does not exhibit any kinks. Fig. 28 and Fig. 29 illustrate the best generalization performance for SRS and Praxis using the two algorithms.

Table 4
NYU Dataset: Performance evaluation using **Median Absolute Error (MAE)** and **Normalized Mutual Information (NMI)** fit, both for testing & training. Lower MAE & higher NMI score indicate better performance. We have highlighted the best performance in bold. The **CDF-KS** column indicates the Kolmogorov-Smirnoff statistic on the CDF comparison against our method. The instances highlighted in bold indicate that the performance is within the accepted 0.05 threshold. Near misses have been underlined. The parameter settings used were the same as those from the KKI dataset in the main manuscript.

Score	Method	MAE Train	MAE Test	NMI Train	NMI Test	CDF-KS
ADOS	PCA & ridge	1.51 ± 1.49	3.52 ± 3.21	0.41	0.23	**0.039**
	ICA & ridge	2.15 ± 2.16	3.52 ± 2.91	0.17	0.11	0.07
	D_N & ridge	2.16 ± 2.39	3.91 ± 3.05	0.39	0.31	0.052
	ANN	1.40 ± 1.39	3.36 ± 2.89	0.31	0.28	**0.001**
	Decoupled	2.61 ± 2.13	3.51 ± 3.17	0.29	0.09	0.21
	JNO Framework	**0.08±0.06**	**2.90±2.18**	**0.99**	**0.41**	–
SRS	PCA & ridge	11.19 ± 14.16	**16.25±14.11**	0.89	0.52	0.16
	ICA & ridge	16.98 ± 16.62	18.90 ± 15.14	0.73	0.43	**0.041**
	D_N & ridge	8.91 ± 6.51	23.52 ± 16.10	0.88	0.51	0.054
	ANN	1.50 ± 1.39	19.04 ± 17.69	0.85	0.08	0.072
	Decoupled	15.11 ± 14.36	24.19 ± 19.17	0.75	0.39	0.059
	JNO Framework	**0.59±0.47**	17.91 ± 14.15	**0.95**	**0.65**	–

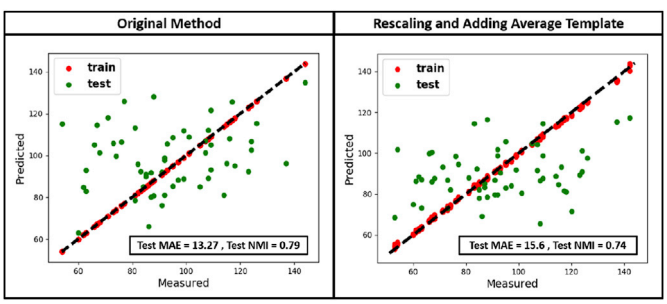

Fig. 28. A performance comparison for SRS prediction after modifying the objective according to Eq. (18). **(L)** Original Method **(R)** After re-scaling and average template addition.

Notice that the modified JNO has a slight tradeoff in regression performance at the expense of the gain in parameter stability. We highlight the importance of this exploration, as future applications of our work include applying our method to rs-fMRI and severity scores from a variety of neurological disorders. Our modified formulation provides additional flexibility in this sense, and extends the overall generalizability of our model.

4. Discussion

Our JNO model cleverly exploits the structure intrinsic to rs-fMRI correlation matrices through an outer product representation. The regression term further guides the basis decomposition to explain the group level and patient specific information. The compactness of our representation serves as a dimensionality reduction step that is related to the clinical score of interest, unlike the pipelined treatment commonly found in the literature. As seen from the results, our JNO framework outperforms a wide range of well established baselines from the machine learning and graph theoretic methods ubiquitous in fMRI analysis on two separate real world datasets.

We conjecture that the baseline techniques fail to extract representative patterns from the correlation data, and learn only the group level representation for the cohort. Consequently, they overfit the training set, despite sweeping the parameters across several orders of magnitude. Any patient level symptomatic and connectivity level differences are lost due to the restrictive pipelined procedure and the group level confounds.

Our Joint Network Optimization Framework is agnostic to the choice of parcellation scheme. This was demonstrated by our additional experiments on the KKI dataset, where we chose the 246 region Brainnetome parcellation to extract correlation matrices (Section 3.6). We emphasize that our framework makes minimal assumptions on the data. Provided we have access to a valid behavioral and network similarity measure, this analysis can be easily adapted to other neurological disorders and even predictive network models outside the medical realm. This greatly broadens the scope of the method to numerous potential applications.

Finally, notice that the training examples (red points) in Fig. 9-13 follow the $\mathbf{x} = \mathbf{y}$ line nearly perfectly. Here, we explain this (potentially misleading) phenomenon in terms of the parametrizatization of our joint objective in Eq. (7).

Recall that Section 2.3 describes the procedure for calculating the coefficients for an unseen patient $\bar{\mathbf{c}}_n$ from the training solution set $\{\mathbf{B}^*, \mathbf{w}^*\}$. Recall that we explicitly set the contribution from the data term in Eq. (5) to 0. Since the patient is not a part of the training set, the corresponding value of $\hat{\mathbf{y}}$ is unknown. In contrast, the training performance is computed based on the estimated coefficients $\mathbf{c}_n$, which have access to the severity scores. Here, we examine the effect of removing the severity information when calculating the coefficients for the training patients. In other words, we estimate the corresponding severity $\mathbf{y}$ excluding the ridge regression term. Accordingly, Fig. 30 highlights the differences in training fit with and without this term is not included in estimating $\mathbf{c}_n$. Notice that in the latter, the training accuracy has the same distribution as the testing points in Fig. 9-13. Taken together, we conclude that, the linear predictive term overparamterizes the search space of solutions for $\mathbf{c}_n$ to yield a near perfect fit. We use this observation to emphasize that the subnetworks and regression model learned by our JNO framework are capturing the underlying data distribution and not simply 'overfitting' the training data.

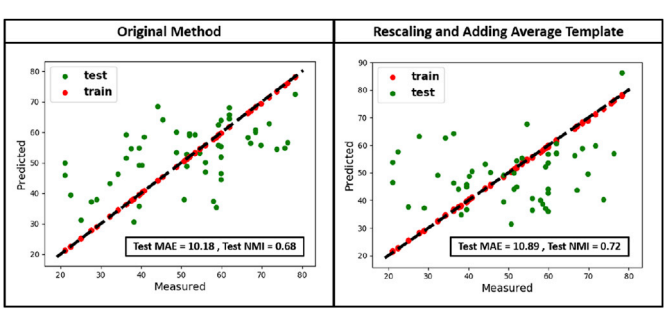

Fig. 29. A performance comparison for Praxis prediction after modifying the objective according to Eq. (18). **(L)** Original Method **(R)** After re-scaling and average template addition.

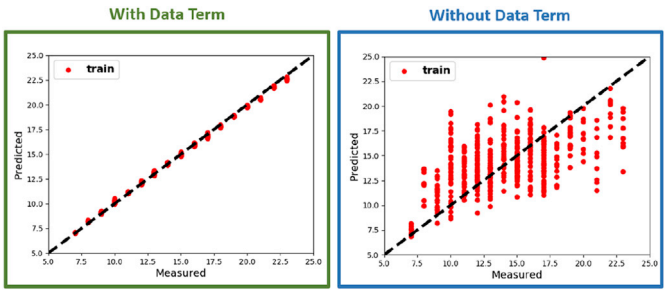

Fig. 30. Prediction Performance of the JNO for ADOS on training data when **(L)** The data term is included in computing $\mathbf{c}_n$ **(R)** The data term is excluded from the computation of. $\mathbf{c}_n$

5. Conclusion

We present an elegant matrix decomposition strategy to combine neuroimaging and behavioral data information. As opposed to generic prediction frameworks, our model directly captures key representative information from the correlation matrices. In contrast, conventional analysis methods dramatically fall short of unifying the two data viewpoints reliably enough to implicate predictive functional patterns in the brain. Our joint optimization framework robustly identifies brain networks characterizing ASD and provides a key link to quantifying and interpreting the clinical spectrum of manifestation of the disorder across patients. Moreover, our evaluation on two separate real world dataset supports the reproducibility of the framework.

We are working on a multi-score extension which can incorporate data from different behavioral domians. In the future, we will explore extensions of this model that learn a patient versus controls distinction in addition to predicting symptom severity. A potential extension of this model includes replacing the linear regression term in Eq. (5) with its non-linear counterpart, thus providing us with the flexibility to model more complex decision functions which can better map the behavioral space.

Another avenue for refinement is to incorporate structural connectivity information in the form of anatomical priors. Typically, structural modalities such as Diffusion Tensor Imaging (DTI) are used to define and track existing anatomical pathways in the brain. Incorporating this information into the network optimization model could be an important step towards unifying anatomical, functional and behavioural domains to better understand altered brain functioning in the context of neurological disorders such as Autism, ADHD, and Schizophrenia.

Our experiment on the KKI dataset using the Brainetome parcellation supports the observation that the method can be tuned with different connectivity and behavioral information, which is an added flexibility. Thus, it could be used to characterize the efficacy of behavioral therapies for neuropsychiatric disorders, as well as, for the development of patient-specific disease biomarkers. We believe that the all of benefits offered by our JNO framework could make it an important diagnostic tool for personalized medicine in the future.

Acknowledgements

This work was supported by the National Science Foundation CRCNS award 1822575 and CAREER award 1845430, the National Institute of Mental Health (R01 MH085328-09, R01 MH078160-07, K01 MH109766 and R01 MH106564), the National Institute of Neurological Disorders and Stroke (R01NS048527-08), and the Autism Speaks foundation.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.116314.

Appendix A

In this section, we derive the alternating minimization updates from Section 2.2:

Optimizing* B *via Proximal Gradient Descent. We first write out the optimization problem with respect to **B** when the estimates of $\{\mathbf{C}, \mathbf{w}\}$ are held constant:

$$\mathbf{B}^{k+1} = \arg\min_{\mathbf{B}} \lambda_1 \left|\left|\mathbf{B}\right|\right|_1 + \sum_n \left|\left|\mathbf{\Gamma}_n - \mathbf{D}_n\mathbf{B}^T\right|\right|_F^2$$

$$+\sum_n \mathrm{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n))\right]$$

$$+\sum_n \frac{1}{2}\left|\left|\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\right|\right|_F^2$$

$$\mathbf{B}^{k+1} = \arg\min_{\mathbf{B}} \left|\left|\mathbf{B}^k\right|\right|_1 + \frac{1}{\lambda_1}\mathscr{G}(\mathbf{B}^k)$$

We see that the proximal gradient iteration is the solution to the following fixed point problem:

$$0 \in \frac{1}{\lambda_1}\frac{\partial \mathscr{G}}{\partial \mathbf{B}} + \partial(\left|\left|\mathbf{B}\right|\right|_1)$$

The derivative of $\mathscr{G}$ with respect to **B**, is computed as:

$$\frac{\partial \mathscr{G}}{\partial \mathbf{B}} = \sum_n \left[2\left[\left[\mathbf{B}\mathbf{D}_n^T - \mathbf{\Gamma}_n\right]\mathbf{D}_n\right] - \mathbf{D}_n\mathbf{diag}(\mathbf{c}_n)\right]$$

$$+\sum_n \left[\mathbf{B}\mathbf{diag}(\mathbf{c}_n)^2 - \mathbf{\Lambda}_n\mathbf{diag}(\mathbf{c}_n)\right]$$

Given the fixed learning rate parameter t, the proximal update for **B** is easily computed as:

$$\mathbf{B}^{k+1} = \mathbf{sgn}(\mathbf{X}) \circ (\mathbf{max}(|\mathbf{X}| - t, 0))$$

$$\mathbf{X} = \mathbf{B}^k - (t / \lambda_1) \frac{\partial \mathscr{G}}{\partial \mathbf{B}}$$

This step first estimates a locally smooth quadratic model at each iterate $\mathbf{B}^k$ and applies a step of iterative shrinkage thresholding to the compute the local solution of $\mathbf{B}$. The resulting iterative algorithm is computationally efficient compared to the counterpart sub-gradient based descent methods and arrives at a good local solution for an appropriate choice of the learning rate.

Optimizing **C** ***using Quadratic Programming.*** The objective is quadratic in $\mathbf{C}$ when $\mathbf{B}$, and $\mathbf{w}$ are held constant. Furthermore, the $\mathbf{diag}(\mathbf{c}_n)$ term decouples the updates for $\mathbf{c}_n$ across patients. Each $\mathbf{c}_n$ is the solution to the a separate optimization problem of the following form:

$$\mathbf{c}_n^{k+1} = \underset{\mathbf{c}_n \in \mathscr{R}^{K+}}{\arg\min} \mathrm{Tr}\left[\mathbf{\Lambda}_n^T \left(\mathbf{D}_n - \mathbf{Bdiag}\left(\mathbf{c}_n^k\right)\right)\right] + \lambda_2 \left|\left|\mathbf{c}_n^k\right|\right|_2^2$$

$$+\frac{1}{2}\left|\left|\mathbf{D}_n - \mathbf{Bdiag}\left(\mathbf{c}_n^k\right)\right|\right|_F^2 + \gamma\left(\left(\mathbf{c}_n^k\right)^T \mathbf{w} - \mathbf{y}_n\right)^2$$

Hence, we use N quadratic programs (QP) of the form below to solve for the vectors $\{\mathbf{c}_n\}$:

$$\frac{1}{2}\mathbf{c}_n^T \mathbf{H}_n \mathbf{c}_n + \mathbf{f}_n^T \mathbf{c}_n \quad s.t. \quad \mathbf{A}_n \mathbf{c}_n \leq \mathbf{b}_n$$

The QP parameters for our problem are given by:

$$\mathbf{H}_n = \mathscr{I}_K \circ (\mathbf{B}^T \mathbf{B}) + 2\gamma \mathbf{w}\mathbf{w}^T + 2\lambda_2 \mathscr{I}_K$$
$$\mathbf{f}_n = -2\left[\mathscr{I}_K \circ \left(\mathbf{D}_n^T + \mathbf{\Lambda}_n^T\right)\mathbf{B}\right] 1 - 2\gamma y_n \mathbf{w};$$
$$\mathbf{A}_n = -\mathscr{I}_K \quad \mathbf{b}_n = 0$$

The non-negativity constraint requires us to project the quadratic programming solution to the space of positive reals in K dimensions for each $\mathbf{c}_n$ through $\mathbf{A}_n$ and $\mathbf{b}_n$. Since the Hessians $\{\mathbf{H}_n\}$ for our problem are positive definite, there exist polynomial time algorithms for solving the bound constrained QPs to the global optimum value. The decoupling of the $\{\mathbf{c}_n\}$ allows us to solve for each coefficient vector in parallel.

Closed Form Update for **w**. The global minimizer of $\mathbf{w}$ is computed at the first order stationary point of the convex objective, which is:

$$\mathscr{J}(\mathbf{w}) = \lambda_3 \left|\left|\mathbf{w}\right|\right|_2^2 + \gamma \left|\left|\mathbf{C}^T \mathbf{w} - \mathbf{y}\right|\right|_2^2$$

$$\frac{\partial \mathscr{J}}{\partial \mathbf{w}} = 0 = 2\lambda_3 \mathbf{w} + 2\gamma(\mathbf{C}\mathbf{C}^T \mathbf{w} - \mathbf{C}\mathbf{y})$$

The closed form update can be expressed as:

$$\mathbf{w} = \left(\mathbf{C}\mathbf{C}^T + \frac{\lambda_3}{\gamma} \mathscr{I}_K\right)^{-1} (\mathbf{C}\mathbf{y})$$

Thus, the ratio $\frac{\lambda_3}{\gamma}$ acts as a regularizer for the matrix inversion in our estimate, ensuring that the update for $\mathbf{w}$ is well defined at each iterate.

Optimizing the Constraint Variables $\mathbf{D}_n$ ***and*** $\mathbf{\Lambda}_n$. A closed form solution for the primal variables $\{\mathbf{D}_n\}$ can be obtained by setting their first derivatives to zero:

$$\frac{\partial \mathscr{J}}{\partial \mathbf{D}_n} = 0 = \mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T + 2\mathbf{\Gamma}_n \mathbf{B}$$
$$-\mathbf{\Lambda}_n - \mathbf{D}_n - 2\mathbf{D}_n \mathbf{B}^T \mathbf{B}$$
$$\mathbf{D}_n = (\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T + 2\mathbf{\Gamma}_n \mathbf{B} - \mathbf{\Lambda}_n)(\mathscr{I}_K + 2\mathbf{B}^T \mathbf{B})^{-1}$$

The gradient ascent update on $\{\mathbf{\Lambda}_n\}$ is as follows:

$$\frac{\partial \mathscr{J}}{\partial \mathbf{\Lambda}_n} = \mathbf{D}_n - \mathbf{Bdiag}(\mathbf{c}_n)$$

$$\mathbf{\Lambda}_n^{k+1} = \mathbf{\Lambda}_n^k + \eta_k \frac{\partial \mathscr{J}}{\partial \mathbf{\Lambda}_n}$$

Similar to the case of the coefficients $\mathbf{c}_n$, each of the N pairs of updates $\{\mathbf{D}_n, \mathbf{\Lambda}_n\}$ are decoupled from each other, and can be solved in parallel. Overall, the sets of $\mathbf{\Lambda}_n$ gradient ascent updates ensure that the respective set of constraints $\mathbf{D}_n = \mathbf{Bdiag}(\mathbf{c}_n)$ is satisfied with increasing certainty at each iteration. The Augmented Lagrangian construct $\left|\left|\mathbf{D}_n - \mathbf{Bdiag}(\mathbf{c}_n)\right|\right|_F^2$ prevents trivial Lagrangian $\mathbf{\Lambda}_n$ solutions.

References

Bardella, Giampiero, et al., 2016. Hierarchical organization of functional connectivity in the mouse brain: a complex network approach. In: Scientific Reports, vol. 6, p. 32060.
Bassett, Danielle Smith, Bullmore, E.D., 2006. Small-world brain networks. The Neuroscientist 12 (6), 512–523.
Behzadi, Yashar, et al., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage 37 (1), 90–101.
Bolte, Sven, Fritz, Poustka, Constantino, John N., 2008. Assessing autistic traits: cross-cultural validation of the social responsiveness scale (SRS). Autism Res. 1 (6), 354–363.
Bullmore, Ed, Sporns, Olaf, 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10 (3), 186.

Bertsekas, Dimitri P., Athena Scientific, 2015. Convex Optimization Algorithms. Athena Scientific Belmont.
Batmanghelich, Nematollah K., Taskar, Ben, Davatzikos, Christos, 2012. Generative-discriminative basis learning for medical imaging. IEEE Trans. Med. Imaging 31 (1), 51–69.
Bush, George, Valera, Eve M., Sei- dman, Larry J., 2005. Functional neuroimaging of attention- deficit/hyperactivity disorder: a review and suggested future directions. Biol. Psychiatry 57 (11), 1273–1284.
D Calhoun, Vince, Liu, Jingyu, Adali, Tiilay, 2009. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. Neuroimage 45 (1), S163–S172.
Cox, Robert W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29 (3), 162–173.
Di Martino, Adriana, et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19 (6), 659.
Dowell, Lauren R., Mahone, E Mark, Mostofsky, Stewart H., 2009. Associations of postural knowledge and basic motor skill with dyspraxia in autism: implication for abnormalities in distributed connectivity and motor learning. Neuropsychology 23 (5), 563.
D'Souza, Niharika Shimona, et al., 2018. A generative- discriminative basis learning framework to predict clinical severity from resting state functional MRI data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 163–171.
Dziuk, M.A., et al., 2007. Dyspraxia in autism: association with motor, social, and communicative deficits. Dev. Med. Child Neurol. 49 (10), 734–739.
Eavani, Harini, et al., 2013. Unsupervised learning of functional network dynamics in resting state fMRI. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 426–437.
Eavani, Harini, et al., 2014. Discriminative sparse connectivity patterns for classification of fMRI data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 193–200.
Eavani, Harini, et al., 2015. Identifying sparse connectivity patterns in the brain using resting-state fMRI. Neuroimage 105, 286–299.
Ecker, Christine, et al., 2010. Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. Neuroimage 49 (1), 44–56.
Fan, L., et al., 2016. The human **brainnetome** atlas: a new brain atlas based on connectional architecture. Cerebral cortex 8, 3508–3526. In press.
Fox, Michael D., Raichle, Marcus E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat. Rev. Neurosci. 8 (9), 700.
Holmes, A.P., Friston, K.J., 1998. Generalisability, random E ects & population inference. Neuroimage 7, S754.
Hong, Seok-Jun, et al., 2017. Multidimensional neuroanatomical subtyping of autism spectrum disorder. Cerebr. Cortex 1–11.
Hoyos-Idrobo, A., et al., 2018. FReM-scalable and stable decoding with fast regularized ensemble of models. Neurolmage 180, 160–172.
Kaiser, Martha D., et al., 2010. Neural signatures of autism. In: Proceedings of the National Academy of Sciences, p. 201010412.
Koshino, Hideya, et al., 2005. Functional connectivity in an fMRI working memory task in high-functioning autism. Neuroimage 24 (3), 810–821.
Liu, Yong, et al., 2008. Disrupted small-world networks in schizophrenia. Brain 131 (4), 945–961.
Lord, Catherine, et al., 2000. The Autism Diagnostic Observation Schedule-Generic: a standard measure of social and communication deficits associated with the spectrum of autism. J. Autism Dev. Disord. 30 (3), 205–223.
Lynch, Charles J., et al., 2013. Default mode network in childhood autism: posteromedial cortex heterogeneity and relationship with social deficits. Biol. Psychiatry 74 (3), 212–219.
Minshew, Nancy J., Keller, Timothy A., 2010. The nature of brain dysfunction in autism: functional brain imaging studies. Curr. Opin. Neurol. 23 (2), 124.
Mostofsky, Stewart H., et al., 2006. Developmental dyspraxia is not limited to imitation in children with autism spectrum disorders. J. Int. Neuropsychol. Soc. 12 (3), 314–326.
Nebel, Mary Beth, et al., 2016. Intrinsic visual-motor synchrony correlates with social deficits in autism. Biol. Psychiatry 79 (8), 633–641.
Niznikiewicz, Margaret A., Kubicki, Marek, Shenton, Martha E., 2003. Recent structural and functional imaging findings in schizophrenia. Curr. Opin. Psychiatr. 16 (2), 123–147.
Parikh, Neal, Boyd, Stephen, et al., 2014. Proximal algorithms. In: Foundations and Trends® in Optimization, 1.3, pp. 127–239.
Penny, William D., et al., 2011. Statistical Parametric Mapping: the Analysis of Functional Brain Images. Elsevier.
Plis, Sergey M., et al., 2014. Deep learning for neuroimaging: a validation study. In: Frontiers in Neuroscience, vol. 8, p. 229.
Pelphrey, Kevin A., Yang, Daniel Y-J., McPartland, James C., 2014. Building a social neuroscience of autism spectrum disorder. In: The Neurobiology of Childhood. Springer, pp. 215–233.
Rahim, Mehdi, et al., 2017. Joint prediction of multiple scores captures better individual traits from brain images. Neurolmage 158, 145–154.
Ravishankar, Hariharan, et al., 2016. Recursive feature elimination for biomarker discovery in resting-state functional connectivity. In: Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the. IEEE., pp. 4071–4074
Rippon, Gina, et al., 2007. Disordered connectivity in the autistic brain: challenges for the 'new psychophysi- ology. Int. J. Psychophysiol. 63 (2), 164–172.
Rubinov, Mikail, Sporns, Olaf, 2010. Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52 (3), 1059–1069.
Sanz-Arigita, Ernesto J., et al., 2010. Loss of 'small-world' networks in Alzheimer's disease: graph analysis of FMRI resting-state functional connectivity. PLoS One 5 (11), e13788.
Sidhu, Gagan S., et al., 2012. Kernel Principal Component Analysis for dimensionality reduction in fMRI-based diagnosis of ADHD. Front. Syst. Neurosci. 6, 74.
Sridharan, Devarajan, Levitin, Daniel J., Menon, Vinod, 2008. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. Proc. Natl. Acad. Sci. 105 (34), 12569–12574.
Sporns, Olaf, et al., 2004. Organization, development and function of complex brain networks. Trends Cogn. Sci. 8 (9), 418–425.
Spitzer, Robert L., Williams, Janet BW., 1980. Diagnostic and statistical manual of mental disorders. Am. Psychiatr. Assoc. Citeseer.
Tzourio-Mazoyer, Nathalie, et al., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15 (1), 273–289.
Uddin, Lucina Q., et al., 2013. Salience network-based classification and prediction of symptom severity in children with autism. JAMA psychiatry 70 (8), 869–879.
Venkataraman, Archana, et al., 2015. An unbiased Bayesian approach to functional connectomics implicates social- communication networks in autism. Neurol. Clin. 8, 356–366.
Venkataraman, Archana, et al., 2016. Bayesian community detection in the space of group-level functional differences. IEEE Trans. Med. Imaging 35 (8), 1866–1882.
Venkataraman, Archana, et al., 2017. A unified bayesian approach to extract network-based functional differences from a heterogeneous patient cohort. In: International Workshop on Connectomics in Neuroimaging. Springer, pp. 60–69.
Venkataraman, Archana, Kubicki, Marek, Golland, Polina, 2012. From brain connectivity models to identifying foci of a neurological disorder. In: International Conference on Medical Image Computing and ComputerAssisted Intervention. Springer, pp. 715–722.
Venkataraman, Archana, Kubicki, Marek, Golland, Polina, 2013. From connectivity models to region labels: identifying foci of a neurological disorder. IEEE Trans. Med. Imaging 32 (11), 20782098.
Williams, Diane L., Goldstein, Gerald, Min- shew, Nancy J., 2006. The profile of memory function in children with autism. Neuropsychology 20 (1), 21.
Woo, Choong-Wan, et al., 2017. Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci. 20 (3), 365.
Xia, Mingrui, Wang, Jinhui, He, Yong, 2013. BrainNet Viewer: a network visualization tool for human brain connectomics. PLoS One 8 (7), e68910.