PAPER • OPEN ACCESS

Sparse coupled logistic regression to estimate co-activation and modulatory influences of brain regions

To cite this article: Thomas A W Bolton *et al* 2020 *J. Neural Eng.* **17** 065003

View the article online for updates and enhancements.

Journal of Neural Engineering

Sparse coupled logistic regression to estimate co-activation and modulatory influences of brain regions

Thomas A W Bolton[1,2,3,7] , Eneko Uruñuela[4] , Ye Tian[5] , Andrew Zalesky[5,6] ,
César Caballero-Gaudes[4] and Dimitri Van De Ville[1,2]

1 Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
2 Department of Radiology and Medical Informatics, University of Geneva (UNIGE), Geneva, Switzerland
3 Department of Decoded Neurofeedback, ATR Computational Neuroscience Laboratories, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
4 Basque Center on Cognition, Brain and Language, San Sebastian, Spain
5 Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne and Melbourne Health, Carlton South, Melbourne, Victoria, Australia
6 Department of Biomedical Engineering, The University of Melbourne, Parkville, Melbourne, Victoria, Australia

E-mail: thomas.bolton@epfl.ch

Abstract

Accurate mapping of the functional interactions between remote brain areas with resting-state functional magnetic resonance imaging requires the quantification of their underlying dynamics. In conventional methodological pipelines, a spatial scale of interest is first selected and dynamic analysis then proceeds at this hypothesised level of complexity. If large-scale functional networks or states are studied, more local regional rearrangements are then not described, potentially missing important neurobiological information. Here, we propose a novel mathematical framework that jointly estimates resting-state functional networks and spatially more localised cross-regional modulations. To do so, the changes in activity of each brain region are modelled by a logistic regression including co-activation coefficients (reflective of network assignment, as they highlight simultaneous activations across areas) and causal interplays (denoting finer regional cross-talks, when one region active at time t modulates the t to $t + 1$ transition likelihood of another area). A two-parameter ℓ_1 regularisation scheme is used to make these two sets of coefficients sparse: one controls overall sparsity, while the other governs the trade-off between co-activations and causal interplays, enabling to properly fit the data despite the yet unknown balance between both types of couplings. Across a range of simulation settings, we show that the framework successfully retrieves the two types of cross-regional interactions at once. Performance across noise and sample size settings was globally on par with that of other existing methods, with the potential to reveal more precise information missed by alternative approaches. Preliminary application to experimental data revealed that in the resting brain, co-activations and causal modulations co-exist with a varying balance across regions. Our methodological pipeline offers a conceptually elegant alternative for the assessment of functional brain dynamics and can be downloaded at https://c4science.ch/source/Sparse_logistic_regression.git.

1. Introduction

How the brain is structurally wired at its most global spatial scale and how information flows between remote processing centres, are essential questions to improve our mechanistic understanding of high-level behaviours [1]. When it comes to functional magnetic resonance imaging (fMRI), the mapping of brain function is commonly performed from resting-state (RS) recordings through the computation of *functional connectivity* (FC), that is, the statistical interdependence between different time courses reflective of regional activity [2], as can

---

7 Author to whom any correspondence should be addressed.

be assessed from an array of measures [3]. This approach has revealed the presence of a set of RS networks (RSNs) [4–6], whose properties are critical landmarks of healthy and perturbed cognition [7–9].

Over the past decade, it has become increasingly clear that quantifying FC between two brain regions throughout a full scanning session as only one scalar is an overly simplistic approach; indeed, it does not characterise the numerous reconfigurations that occur at the time scale of seconds [10]. Accordingly, many methodological pipelines have been developed to dig into time-resolved FC and map brain function dynamically (see [11–14] for contemporary reviews).

One of the most notorious family of dynamic approaches simplifies the originally voxel-wise fMRI data into a state-level representation: first, whole-brain FC is computed over successive temporal sub-windows and then, the concatenated data across the full subject population at hand is decomposed into a set of dynamic FC (dFC) states. Each of them is temporally recurring, short-lived and renders a distinct set of correlational relationships across individual brain parcels, or—if spatial independent component analysis (ICA) is performed prior to sliding window computations—across RSNs [15–17].

In other analytical schemes, whole-brain voxel-wise activity [18], or activity transients [19], undergo clustering instead of FC patterns; in this case, each of the retrieved centroids directly stands for an RSN. If temporal ICA is directly cascaded to spatial ICA, temporally mutually independent functional modes that highlight specific RSN combinations are retrieved [20]. Finally, the use of a hidden Markov model (HMM) also enables to derive hidden states reflective of RSNs, or of their interplays, which are parameterised as (sparse) FC patterns [21–23] or vectors of activation [24].

In all the above cases, one assumes that the fMRI data can be efficiently understood in terms of a restricted set of RSNs and that functional brain dynamics should be investigated from a fixed and restricted set of spatial patterns. Recent results, however, challenge the sufficiency of such postulates: first, some brain regions do not remain attached to the same network throughout a scanning session, but instead adjust their modular allegiance over time in a way that relates to cognitive performance [25, 26]. Second, brain regions and networks morph spatially over time, with this spatial dynamics bearing promising clinical relevance [27, 28]. Third, spatio-temporal patterns have been suggested as more telling features extracted from the fMRI signal [29–31].

In order to capture such subtle propagation of activity at the finer regional scale, *effective connectivity* (EC) approaches have also been developed. In opposition to the above correlational tools, they explore the *causal* relationships (i.e., from time t to $t+1$) that link distinct brain areas. Notoriously,

dynamic causal modelling has been leveraged to the RS setting: the cross-spectral content of the data is described probabilistically—including haemodynamic effects—and model inversion yields the posterior probability density for each EC coefficient (i.e., the probability that it takes a given value knowing the cross-spectra). Recent technical improvements have pushed towards making such computationally greedy approaches applicable at the whole-brain scale [32, 33]. In other related work that did not employ a Bayesian framework, EC coefficients were derived from the empirical cross-spectral density of the data with an added ℓ_1 regularisation constraint, forcing the set of cross-regional causal relationships to be sparse [34].

An alternative to a spectral characterisation of the data is to remain in the temporal domain and explicitly enforce the causality of the system. If working in the continuous domain, with a multivariate Ornstein-Uhlenbeck model, regional activity can be described by a system of coupled ordinary differential stochastic equations reaching a steady-state and the EC coefficients that yield the best set of lagged covariance matrices (in terms of fitting empirical ones) are obtained by iterative updates [35, 36]. In the discrete domain, first-order multivariate autoregressive models have also been applied: with such causal tools, sliding window-based fluctuations in FC—a correlational measure, as highlighted above—could be well replicated [37]. In addition, autoregressive approaches have shown relevance in the characterisation of several facets of human behaviour [38].

It transpires from the above that at present, there are two conceptually distinct ways to view RS dFC: on the one hand, as sets of simultaneously activating regions that make networks and on the other hand, as effective connectivity between individual areas. Which of these two alternatives offers the best representation of RS dynamics and whether they describe overlapping or distinct facets of the data, remain important questions to explore. In the present work, we have attempted to progress in answering them by developing a novel methodological framework that characterises whole-brain activity through coupled logistic regression equations. Co-activations and causal couplings are jointly derived for each pair of brain regions and the inclusion of sparsity constraints in our model allows us to only extract a parsimonious array of coefficients, while enabling, at the same time, to modulate the trade-off in data fitting between both viewpoints.

2. Materials and Methods

2.1. Mathematical framework

Let us denote the activity of a region r (out of R in total) at time t as $h_t^{(r)}$. We hypothesise two possible states of activity: *baseline* ($h_t^{(r)} = 0$) or *active*

($h_t^{(r)} = +1$). We further posit that each region may interact with all the other areas $s \neq r$ in two ways: (1) showing simultaneous activity (that is, episodes of co-activation), or (2) being causally modulated. To jointly describe these two phenomena, we characterise the probability of a region r to switch between activity states via logistic regression [39]:

$$\begin{cases} \mathcal{P}(h_{t+1}^{(r)} = +1 | h_t^{(r)} = 0, \mathbf{h}_t^{(-\mathbf{r})}, \mathbf{h}_{t+1}^{(-\mathbf{r})}) \\ \quad = \dfrac{1}{1 + e^{-(\alpha_B^{(r)} + \boldsymbol{\gamma}_B^{(r)\top} \mathbf{h}_{t+1}^{(-\mathbf{r})} + \boldsymbol{\beta}_B^{(r)\top} \mathbf{h}_t^{(-\mathbf{r})})}} \\ \mathcal{P}(h_{t+1}^{(r)} = 0 | h_t^{(r)} = +1, \mathbf{h}_t^{(-\mathbf{r})}, \mathbf{h}_{t+1}^{(-\mathbf{r})}) \\ \quad = \dfrac{1}{1 + e^{-(\alpha_A^{(r)} + \boldsymbol{\gamma}_A^{(r)\top} \mathbf{h}_{t+1}^{(-\mathbf{r})} + \boldsymbol{\beta}_A^{(r)\top} \mathbf{h}_t^{(-\mathbf{r})})}} \end{cases} \quad (1)$$

The baseline-to-active transition is modelled by the first equation, while the return to baseline from an active state is governed by the second. Associated coefficients are respectively written with the $\cdot_B$ and $\cdot_A$ subscripts. In what follows, for the sake of clarity, we will omit these subscripts and only consider one set of equations, as the formulations are strictly equivalent for both types of transitions.

If all other regions are at a baseline level of activity at the start ($\mathbf{h}_t^{(-\mathbf{r})} = 0$) and end ($\mathbf{h}_{t+1}^{(-\mathbf{r})} = 0$) of the transition, only the scalar coefficient $\alpha^{(r)}$ plays a role in shaping the transition likelihood. The vector $\boldsymbol{\gamma}^{(r)} \in \mathbb{R}^{R-1}$ contains the co-activation coefficients for all regions $s \neq r$: positive-valued coefficients will enhance the likelihood of the transition of interest if $h_{t+1}^{(s)} = +1$ (that is, if regions r and s are co-active at time $t+1$). Negative-valued coefficients will, likewise, reduce the transition probability. The reasoning is similar for the vector $\boldsymbol{\beta}^{(r)} \in \mathbb{R}^{R-1}$, except that a modulatory effect is then exerted if $h_t^{(s)} = +1$ (i.e., region s is active before the transition in activity level of region r, resulting in a causal modulation instead of a co-activation).

The concomitant modelling of co-activations and causal modulations enables to jointly derive the two sets of coefficients. Given the fact that the resting brain is often characterised as a set of RSNs [4–6], we expect only a sparse subset of non-null co-activation coefficients. Similarly, only a restricted amount of areas or networks are believed to causally modulate each other [40, 41]. To fit these neurobiological priors, we can consider that the joint set of coefficients is sparse by imposing an ℓ_1 regularisation term on them:

$$(1 - \xi^{(r)})||\boldsymbol{\gamma}^{(r)}||_1 + \xi^{(r)}||\boldsymbol{\beta}^{(r)}||_1 < \rho^{(r)} \quad \forall \quad r = 1, \dots, R. \quad (2)$$

In the above, $\rho^{(r)}$ controls the extent of regularisation casted on all coefficients associated to region r (it relates to an inversely proportional parameter $\lambda^{(r)}$ in equation (3) below). The parameter $\xi^{(r)}$ enables to balance to what extent the co-activation and causal

sets are regularised for a given area: if $\xi^{(r)} = 0$, regularisation only operates on co-activation coefficients, while if $\xi^{(r)} = 1$, only causal coefficients are made sparse. This respectively amounts to a description of regional brain dynamics where causal influences, or co-activations, dominate. Note that, since each region is associated to dedicated regularisation parameters, it becomes possible to address nuanced differences in influence within the whole-brain circuitry and in causal/co-activation balance.

2.2. Implementation

Solving the above set of coupled logistic regression equations requires that the activity levels of all regions be known. To binarise the input time courses, we individually z-score each and set to $+1/0$ the time points with a value above/below 0. While binarisation may remove part of the insightful information from the original data, it has been used in recently developed methodological pipelines [42]. In the Discussion, we touch upon possibilities to make the framework amenable to a case with more than 2 states of activity.

After defining the activation states, initial parameter estimates can be computed. Co-activation and modulatory coefficients are all set to 0 and intrinsic transition probabilities are set to 0.5 (*i.e*, $\alpha^{(r)} = 0$).

Following [39], in a regularised logistic regression, one attempts to solve the following:

$$\min_{\alpha^{(r)}, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\beta}^{(r)}} - \mathcal{L}^{(r)}(\alpha^{(r)}, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\beta}^{(r)})$$
$$+ \lambda^{(r)} \left[(1 - \xi^{(r)})||\boldsymbol{\gamma}^{(r)}||_1 + \xi^{(r)}||\boldsymbol{\beta}^{(r)}||_1 \right], \quad (3)$$

where r is the assessed region and the log-likelihood is approximated as:

$$\mathcal{L}^{(r)}(\alpha^{(r)}, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\beta}^{(r)}) = -\frac{1}{2|\mathcal{T}|} \sum_{t \in \mathcal{T}} \omega_t^{(r)}(z_t^{(r)} - \alpha^{(r)}$$
$$- \boldsymbol{\gamma}^{(r)\top} \mathbf{h}_{t+1}^{(-\mathbf{r})} - \boldsymbol{\beta}^{(r)\top} \mathbf{h}_t^{(-\mathbf{r})}) + C. \quad (4)$$

The ensemble $\mathcal{T}$ contains all the data points for which the probed region is in the currently considered start state at time t (*e.g.*, baseline for the baseline-to-active transitions) and C is a constant. If we define the probability of the transition of interest as $p(\alpha^{(r)}, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\beta}^{(r)}, \mathbf{h}_t^{(-\mathbf{r})}, \mathbf{h}_{t+1}^{(-\mathbf{r})})$, the parameters $\omega_t^{(r)}$ and $z_t^{(r)}$ depend on the current estimates of the coefficients—which we denote with a tilda—as:

$$\begin{cases} \omega_t^{(r)} = p(\tilde{\alpha}^{(r)}, \tilde{\boldsymbol{\gamma}}^{(r)}, \tilde{\boldsymbol{\beta}}^{(r)}, \mathbf{h}_t^{(-\mathbf{r})}, \mathbf{h}_{t+1}^{(-\mathbf{r})}) \\ \quad - p(\tilde{\alpha}^{(r)}, \tilde{\boldsymbol{\gamma}}^{(r)}, \tilde{\boldsymbol{\beta}}^{(r)}, \mathbf{h}_t^{(-\mathbf{r})}, \mathbf{h}_{t+1}^{(-\mathbf{r})})^2 \\ z_t^{(r)} = \tilde{\alpha}^{(r)} + \tilde{\boldsymbol{\gamma}}^{(r)\top} \mathbf{h}_{t+1}^{(-\mathbf{r})} + \tilde{\boldsymbol{\beta}}^{(r)\top} \mathbf{h}_t^{(-\mathbf{r})} \\ \quad + \dfrac{1}{\omega_t^{(r)}} \left[y_t^{(r)} - p(\tilde{\alpha}^{(r)}, \tilde{\boldsymbol{\gamma}}^{(r)}, \tilde{\boldsymbol{\beta}}^{(r)}, \mathbf{h}_t^{(-\mathbf{r})}, \mathbf{h}_{t+1}^{(-\mathbf{r})}) \right] \end{cases} \quad (5)$$

where $y_t^{(r)}$ defines whether there was a change in activity level in region r from time t to $t+1$ or

not (respectively, $y_t^{(r)} = 1$ or $y_t^{(r)} = 0$). Coefficients are iteratively estimated by a coordinate-wise descent algorithm, following [43]: the initial estimates outlined above are used at the maximal regularisation level λ_{MAX} and individual coefficients are successively re-estimated in random order (note that for $\alpha^{(r)}$ coefficients, which do not enter the ℓ_1 regularisation term, soft shrinkage is not required). After cycling through all coefficients, the Euclidean distance of the whole coefficient vector with respect to the prior iteration is assessed and the algorithm stops either when the change across two iterations becomes lower than a defined tolerance threshold ε, or when n_{iter} iterations have been performed. The next regularisation level is then considered, using warm restarts to speed up computations (i.e., the estimates obtained at the end of a regularisation cycle are used as initial values for the following one).

In all the analyses performed therein, we compared five levels of trade-off between co-activation and causal coefficients ($\xi^{(r)} = \{0, 0.25, 0.5, 0.75, 1\}$ $\forall$ $r = 1, \ldots, R$) and used convergence parameters $\varepsilon = 10^{-2}$ and $n_{\mathrm{iter}} = 5$.

In our simulations, we considered a regularisation path with $\lambda^{(r)} \in [10000, 0]$ $\forall$ $r = 1, \ldots, R$ (80 logarithmically distributed values), while for our application to experimental data, we used $\lambda^{(r)} \in [50000, 0]$ $\forall$ $r = 1, \ldots, R$ (60 logarithmically distributed values). We always verified that at λ_{MAX}, all coefficients remained equal to 0.

2.3. Determination of final co-activation and causal modulation values

Upon solving, the framework yields an array of co-activation and causal coefficients across all examined regulariser values. In order to determine the optimal parameters for each region r, we resorted to cross-validation, using otherwise untouched data. In the cases considered in this work (both simulations and experimental data), the cross-validation dataset always had 60% of the training set size. Following z-scoring and binarisation of each regional time course in the same way as described in section 2.2, the exact log-likelihood was computed, for each candidate parameter set ($\xi^{(r)}, \lambda^{(r)}$), as:

$$\mathcal{L}^{(r)}(\xi^{(r)}, \lambda^{(r)}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} y_t^{(r)} (\alpha^{(r)} + \boldsymbol{\gamma}^{(r)\top} \mathbf{h}_{t+1}^{(-\mathbf{r})} + \boldsymbol{\beta}^{(r)\top} \mathbf{h}_t^{(-\mathbf{r})})$$
$$+ \log(1 + e^{\alpha^{(r)} + \boldsymbol{\gamma}^{(r)\top} \mathbf{h}_{t+1}^{(-\mathbf{r})} + \boldsymbol{\beta}^{(r)\top} \mathbf{h}_t^{(-\mathbf{r})}}),$$
(6)

where $\alpha^{(r)} = \alpha^{(r)}(\xi^{(r)}, \lambda^{(r)})$, $\boldsymbol{\gamma}^{(r)} = \boldsymbol{\gamma}^{(r)}(\xi^{(r)}, \lambda^{(r)})$ and $\boldsymbol{\beta}^{(r)} = \boldsymbol{\beta}^{(r)}(\xi^{(r)}, \lambda^{(r)})$ were used as short-hand notations for the sake of clarity and $y_t^{(r)}, \mathbf{h}_t^{(-\mathbf{r})}$ and $\mathbf{h}_{t+1}^{(-\mathbf{r})}$ are computed from the cross-validation set. For each region r, optimal coefficients were set as the ones maximising the above log-likelihood function.

Following this step, coefficients are converted into a probabilistic equivalent. Let two regions r and s; we can use equation (1) to define the probability for

region r to undergo a change in activity when region s is itself active and similarly, when it is not. The difference is then taken as the measure of interest. For co-activation, we contrast $h_{t+1}^{(s)} = +1$ and $h_{t+1}^{(s)} = 0$, while $h_t^{(s)} = 0$; by this mean, we selectively evaluate co-activation independently from causal regulation. This gives the following probability differential:

$$\Delta \mathcal{P}_{\Gamma, s \to r} = \mathcal{P}(h_{t+1}^{(r)} \neq h_t^{(r)} | h_{t+1}^{(s)}$$
$$= +1, \mathbf{h}_t^{(-\mathbf{r})} = 0, \mathbf{h}_{t+1}^{(-\mathbf{r}, -\mathbf{s})} = 0)$$
$$- \mathcal{P}(h_{t+1}^{(r)} \neq h_t^{(r)} | \mathbf{h}_t^{(-\mathbf{r})} = 0, \mathbf{h}_{t+1}^{(-\mathbf{r})} = 0).$$
(7)

In a similar vein, for causal modulations, we have:

$$\Delta \mathcal{P}_{B, s \to r} = \mathcal{P}(h_{t+1}^{(r)} \neq h_t^{(r)} | h_t^{(s)}$$
$$= +1, \mathbf{h}_t^{(-\mathbf{r}, -\mathbf{s})} = 0, \mathbf{h}_{t+1}^{(-\mathbf{r})} = 0)$$
$$- \mathcal{P}(h_{t+1}^{(r)} \neq h_t^{(r)} | \mathbf{h}_t^{(-\mathbf{r})} = 0, \mathbf{h}_{t+1}^{(-\mathbf{r})} = 0),$$
(8)

where this time we contrast the activity of region s at time t instead. The resulting values can be arranged in two matrices (one per type of coefficient), where the rth column contains the $R - 1$ influences onto region r (diagonal elements are left empty). Recall that this process is performed separately for two types of transitions: baseline to active and vice versa. Let us respectively denote the associated co-activation matrices by $\boldsymbol{\Gamma}_\mathbf{B}$ and $\boldsymbol{\Gamma}_\mathbf{A}$, while we term causal modulation matrices $\mathbf{B_B}$ and $\mathbf{B_A}$.

Considering an example coupling between regions s and r, a positive-valued $\boldsymbol{\Gamma}_\mathbf{B}(s, r)$ element means that when region s is active at time $t + 1$, region r will have a greater likelihood to transit to the active state from time t to $t + 1$. A positive-valued $\boldsymbol{\Gamma}_\mathbf{A}(s, r)$ value, however, means that upon activity of region s at time $t + 1$, region r is more likely to transit back from the active to the baseline state from time t to $t + 1$. Similar observations can be made for causal modulations.

Thus, a simple solution to aggregate both types of transition is to consider $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_\mathbf{B} - \boldsymbol{\Gamma}_\mathbf{A}$ and $\mathbf{B} = \mathbf{B_B} - \mathbf{B_A}$ as the final values of interest. Positive-valued entries then reflect up-regulatory influences, irrespective of the transition type. Figure 1 schematically recapitulates the undertaken steps to generate the examined features. Note that, while we stick to such a simplified representation throughout most of our work, in figure 5(C), we briefly touch upon the theoretical ability of our framework to reveal subtler types of dynamics that dissociate activity states.

2.4. Validation of the framework on simulated data

To verify the face validity of our framework and assess its flexibility under different settings, we first considered simulated data containing cross-regional causal modulations as well as co-activations. We simulated activity time courses for $R = 35$ regions (or $R = 40$ in a sub-case presented in figure 5(B)). To

Figure 1. Overview of the framework. (A) Example activity time courses for a set of 9 regions; each can transit between a baseline state of activity (symbolised by a grey circle) and an active state (red circle). The green, brown and blue underlays highlight the regions that belong to the same RSN and thus exhibit a similar transitory dynamics. (B) Coefficient matrices associated to the example presented in (A) for co-activations (top row) and causal modulations (bottom row). The left column pertains to the transition from the baseline to the active state: a positive-valued coefficient at element (s, r) means that when region s is active, it enhances the likelihood of a transition for region r at the same time point (for co-activations) or one time point later (for causal modulations). The middle column similarly characterises transitions from the active to the baseline state; thus, modulations that enhance the overall activity of an area are here reflected by negative-valued coefficients (i.e., the probability to go down in activity is lowered). The right column yields total influences summed across both transition types. (C) First, the sparse logistic regression framework is solved along a regularisation path (λ parameter), for different co-activation/causal coupling balances (ξ parameter), as described in section 2.2. Second, cross-validation is performed to extract, for each region r, the best set of coefficients that maximise the log-likelihood of the unseen data (see equation (6)). Third, coefficients are converted into cross-regional probabilistic modulations (see Equations (7) and 8). Fourth, the resulting values are arranged into matrices and information is aggregated across transition types by a subtraction. R: region. N: network.

match the experimental data case as much as possible, we considered $T = 1200$ time points per subject and we used a number of subjects S that would yield a similar amount of available data points for the estimation of each parameter of the model. In more details, we have:

$$S = \frac{n_{\mathrm{DP}}(2R + 4R(R-1))}{T}, \qquad (9)$$

where n_{DP} denotes the number of data points required for properly estimating one parameter and $2R + 4R(R-1)$ is the total number of parameters to estimate. The number of available subjects on experimental data ($S = 350$, for the estimation of $R = 94$ regions) is achieved with $n_{\mathrm{DP}} = 12$; applying the same equation to $R = 35$ (or $R = 40$) then yields $S = 50$ (or $S = 65$).

In our initial simulation (presented in figure 2) and the majority of ensuing ones, we considered $N = 7$ separate RSNs, a number that matches data from the RS literature [6]. In all simulations conducted with $N = 7$, each network contained between 4 and 7 areas (from network 1 to 7: 5, 4, 7, 6, 4, 5 and 4 regions) and time courses for all regions belonging to the same network were similar (prior to the addition

of noise). To examine the flexibility of our pipeline, we also explored some cases with $N = 3$ (presented in figures 3(A) and (B)), where networks 1, 2 and 3 respectively comprised 10, 14 and 11 regions. In the case examined in figure 5(B), a few additional regions were also set as *hubs* that jointly belong to two networks and activate as soon as one of the networks turns on.

Each simulated dynamics was associated to a probability to switch from the baseline to the active state and to a probability to transit from the active to the baseline state. In all examined simulation cases, both were set to 0.5. Causal modulations were introduced between a subset of networks: when a modulating network turned active, it could enhance the activity of the modulated network (both by enhancing the likelihood of a 0 to +1 transition and reducing that of a +1 to 0 one), as symbolised by a positive-valued causal coefficient, or decrease that activity, as reflected by a negative-valued element. We always considered a probability modulation equal to 0.4 and in one case examined in figure 5(C), also considered distinct pools of causal modulations between the baseline-to-active and active-to-baseline transition cases.

The number of simulated networks is directly related to the density of co-activation coefficients present in the problem, ρ_Γ (a larger N lowers ρ_Γ). The number of causal modulations across networks then defines the density of causal coefficients, ρ_B. In figure 3, we explored the robustness of our framework to the exact balance between co-activation and causal coefficients.

Eventually, all time courses were corrupted with Gaussian noise with $\sigma^2 = 2$, apart from the results shown in figure 4 where our framework is compared to alternative approaches (described in section 2.6) while exploring a range of possible noise variances. Indicative regional time courses for a simulated subject under such noise settings are presented in figure 2(A), where noise is sufficient not to be able to infer any cross-regional relationships by mere eyesight.

2.5. Quantification of the ability to retrieve the ground truth

In order to assess how accurately ground truth parameters could be retrieved, we considered an array of quality measures. Separately for co-activation and causal coefficients, we first computed Pearson's correlation coefficient between the ground truth coefficient matrix and the output from our pipeline (respectively, Γ and B). In what follows, we term this metric *similarity*.

For extracted co-activation coefficients, we also examined whether the contained information was sufficient to re-order the regions into their underlying networks, by computing Ward's linkage from the columns of Γ (having excluded diagonal elements). We separated all regions into N clusters using the constructed dendrogram and used the *purity* measure [44] to compare the obtained clusters to the ground truth. A purity of 1 denotes perfect agreement between both sets.

For causal modulations, we used the ground truth network structure to construct a directional graph: first, the elements of B that were associated to null values in either B_B or B_A were set to 0, so that the regularisation potential of our framework is fully exploited in yielding a sparse graph. Second, from this modified matrix $\hat{B}$, all probabilistic causal couplings associated to the same network-to-network modulation were joined together, resulting in an $N \times N$ graph. In doing so, we used the median operator instead of the mean to preserve sparsity. From the generated directional graph, we computed *sensitivity* and *specificity* in directional edge detection as two separate quality metrics.

2.6. Comparison of performance to other approaches

We compared our framework to four alternative approaches (two that derive co-activations and two that extract causal modulations), using the metrics introduced in section 2.5. For each of these methods,

all data points across subjects were jointly analysed in a population-level analysis, to match the application of our framework. In addition, the same number of samples was used for both the training and the cross-validation datasets.

For co-activation, we first selected the graphical lasso (GLasso) [45], which also leverages ℓ_1 regularisation and is widely applied for the estimation of static or dynamic FC in the literature. We performed cross-validation to extract the optimal regularisation parameter and were always able to locate a clear log-likelihood maximum within the interval of probed values.

As a second co-activation approach, we considered the point process analysis (PPA) put forward in [46], which derives a proxy for FC using only a subset of the available time samples per voxel or region. We found it interesting to compare our framework to another methodology that operates at the frame-wise level, without the reliance on second-order statistics. The approach relies on a thresholding parameter T_{PPA} to define the moments of interest in each activity time course (i.e., those that overcome this threshold): to set it, we performed cross-validation by computing Pearson's correlation coefficient between the estimated FC proxy from the training data and the FC matrix derived from the cross-validation dataset. In all examined cases, the probed range for T_{PPA} yielded a clear similarity maximum.

For the estimation of causal coefficients, we first selected the approach introduced by [34], which works at the level of the cross-spectral density (CSD) of the data with an added ℓ_1 regularisation constraint, making it somehow conceptually related to our framework. No parameter needed to be tuned, but to enhance sparsity of the output matrix (for which many non-null, but negligible values remained), coefficients lower than $\epsilon_{CSD} = 10^{-7}$ were set to zero. The rationale for this was to ease the generation of a network-to-network directional graph representation, which we use in the evaluated quality metrics.

As a second alternative, we considered the use of an order-1 multivariate autoregressive model (MAR) [37], for which we only considered the cross-regional coefficients. In order to enable sparsity of the outputs, we generated null realisations in which regional time courses were independently randomly shuffled across subjects and assessed significance of the coefficients at a Bonferroni-corrected p-value of $\frac{0.05}{N(N-1)}$ (that is, correcting for the known maximal number of possible cross-network couplings).

All five candidate approaches were examined across a series of noise values $\sigma^2 = 1, 2, 4, 9, 16, 25$, for a total number of subjects in the training dataset equal to $S = 1, 15, 30, 40, 50, 80$. We assessed how much performance would decrease with larger noise and/or less available subjects to derive coefficients. The results of these analyses are presented in figure 4.

2.7. Application of the framework to experimental fMRI data

We applied our framework to experimental RS fMRI data from the *Human Connectome Project* [47]. We considered one scanning session long of $T = 1200$ time points. The data from $S = 350$ subjects served to extract co-activation and causal coefficients and $S_{CV} = 207$ separate subjects were considered for cross-validation. Finally, a yet distinct pool of $S_{VAL} = 350$ subjects was leveraged to rerun the framework at optimal regularisation values, yielding new co-activation and causal coupling estimates that we compared to the original ones to gauge the generalisability of our findings.

The data was acquired at a fast TR of 720 ms, at a spatial resolution of $2 \times 2 \times 2$ mm^3; additional acquisition details can be found elsewhere [48]. We considered ICA-FIX denoised preprocessed voxel-wise time courses with extra Wishart rolloff filtering to improve signal to noise ratio, similarly to [49]. The data was originally parcellated into 376 separate areas (360 from the Glasser atlas [50] and 16 added subcortical regions), but as we did not have enough data to our disposal for properly estimating parameters from such a high-dimensional representation, we downscaled these 376 areas into $R = 94$ separate parcels. To do so, we computed a weighted average of the parcels from the Glasser atlas that overlapped with a given parcel from the AAL atlas [51], where the weights jointly reflected the size of the original parcels and their relative overlap with the output parcel.

As a final step, from the fully preprocessed data, we used a total variation-based denoising approach [52, 53] to derive cleaned *activity-inducing* signals freed from haemodynamic effects. We only included temporal regularisation in the process, without any spatial prior, to avoid the need to manually specify any parameter. By this deconvolution step, we hoped to minimise auto-correlation in the analysed time courses.

From columns of the co-activation matrix $\mathbf{\Gamma}$, Ward's linkage analysis was conducted to perform hierarchical clustering into distinct networks. To effectively look at network identities, we defined a distance cutoff by generating 10 000 null realisations in which each column was independently shuffled prior to linkage analysis. For each of these null cases, the maximal distance between columns was sampled and eventually, we used the 95th percentile of this null distribution as cutoff.

From the causal coupling matrices $\mathbf{B_B}$ and $\mathbf{B_A}$, a directional graph representation was generated as described in section 2.5, using the network assignments derived from the above hierarchical clustering process.

To compare our findings to those from the validation dataset, we considered (1) similarity between the co-activation matrices $\mathbf{\Gamma}$ and $\mathbf{\Gamma_{VAL}}$, (2) purity between the network assignments derived from the training data and the ones extracted by running hierarchical clustering on the validation data, using the training-inferred number of networks, (3) similarity between the matrices containing causal modulations ($\mathbf{B}$ and $\mathbf{B_{VAL}}$) and (4) comparison between the directional graphs obtained from the training data and from the validation data, using the training-inferred network assignments.
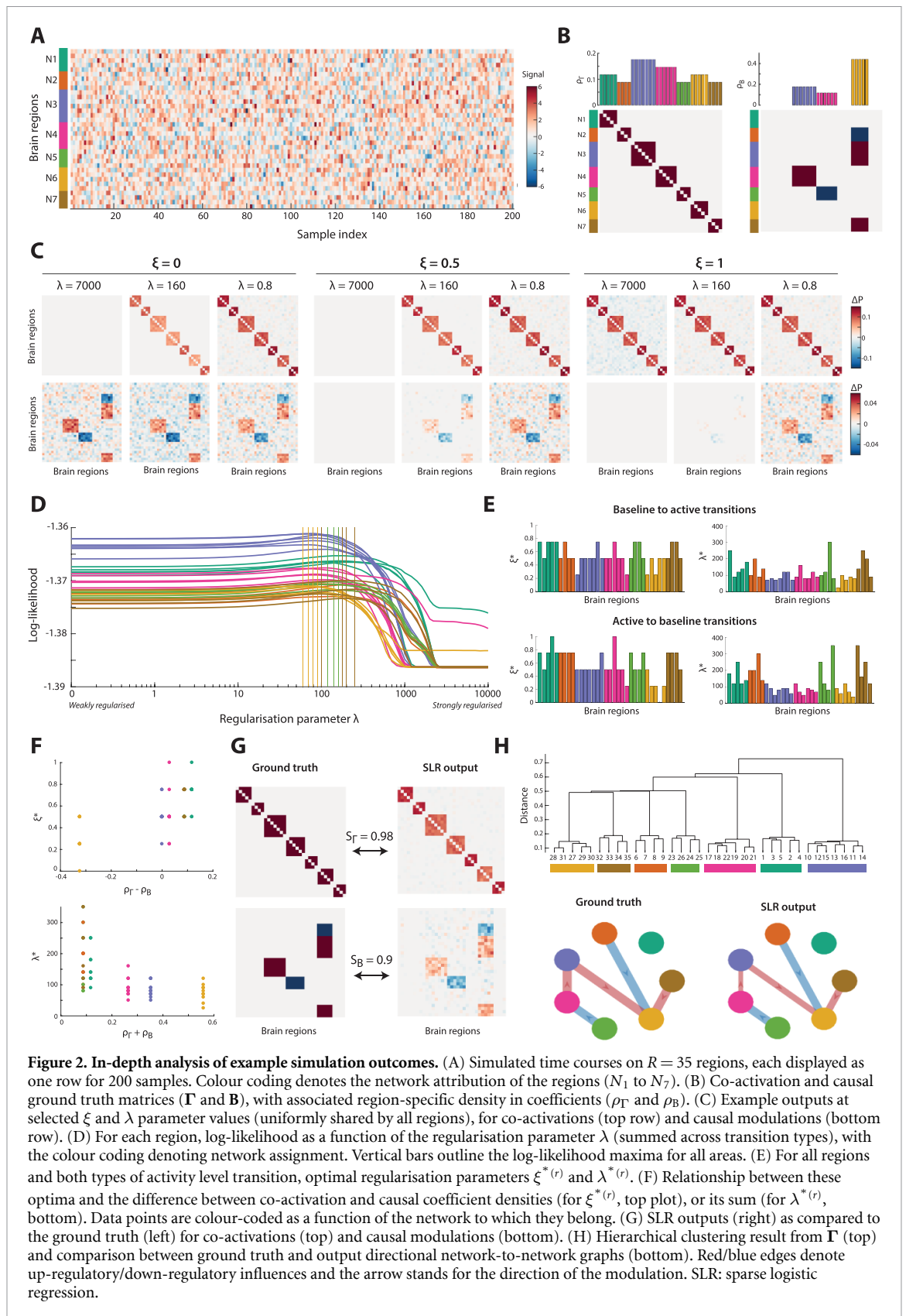
3. Results

3.1. Validation of the framework on simulated data

Figure 2 displays the results from an example simulation for which we go in depth into the information provided by our sparse logistic regression (SLR) framework. At the considered noise level $\sigma^2 = 2$, regional time courses cannot easily be assigned to their underlying networks (figure 2(A), left colour-coding), although in fact, there is an underlying organisation into $N = 7$ distinct systems (figure 2(B), bottom left $\mathbf{\Gamma_{GT}}$ matrix). In addition to the co-activation structure, there are also three positive-valued causal modulations and two negative-valued ones, across the networks (figure 2(B), bottom right $\mathbf{B_{GT}}$ matrix). As a function of their network assignment, different regions thus showcase distinct densities in co-activation coefficients ρ_Γ (depending on how many regions are part of the same network) and in causal ones (ρ_B; see figure 2(B), top half).

In figure 2(C), example outputs of the framework (i.e., $\mathbf{\Gamma}$ and $\mathbf{B}$ matrices) are provided when the same ξ and λ values are used across all regions. As anticipated, it can be seen that for $\xi = 0$, co-activations are more strongly attenuated, while for $\xi = 1$, causal modulations are more largely removed. Furthermore, as λ decreases, the overall extent of regularisation is lowered, yielding a less sparse set of coefficients. There are two further interesting points to note: first, co-activation probabilistic influences are generally larger than causal ones. Second, regardless of the exact ξ and λ values used, the SLR outputs strongly resemble the ground truth.

In figure 2(D), the log-likelihood (as computed on cross-validation data and summed across both transition types) is plotted for each region as a function of λ, at the region-specific optimum $\xi^{*(r)}$. Regardless of the region, the log-likelihood was lowest around the largest regularisation values (right of the graph), a scheme in which it can be seen from figure 2(C) that fittingly, the ground truth structure is then not captured. When λ became lower, the log-likelihood gradually increased, until it reached a clear peak at values around $\lambda = 100$, with the exact location differing from region to region (see the coloured vertical bars). Note that the regions belonging to network 6 (light brown colour) were those linked to the lowest optimal regularisation level, fitting the fact that they had the most elevated overall density in incoming

Figure 2. In-depth analysis of example simulation outcomes. (A) Simulated time courses on $R = 35$ regions, each displayed as one row for 200 samples. Colour coding denotes the network attribution of the regions (N_1 to N_7). (B) Co-activation and causal ground truth matrices (Γ and $\mathbf{B}$), with associated region-specific density in coefficients (ρ_Γ and ρ_B). (C) Example outputs at selected ξ and λ parameter values (uniformly shared by all regions), for co-activations (top row) and causal modulations (bottom row). (D) For each region, log-likelihood as a function of the regularisation parameter λ (summed across transition types), with the colour coding denoting network assignment. Vertical bars outline the log-likelihood maxima for all areas. (E) For all regions and both types of activity level transition, optimal regularisation parameters $\xi^{*(r)}$ and $\lambda^{*(r)}$. (F) Relationship between these optima and the difference between co-activation and causal coefficient densities (for $\xi^{*(r)}$, top plot), or its sum (for $\lambda^{*(r)}$, bottom). Data points are colour-coded as a function of the network to which they belong. (G) SLR outputs (right) as compared to the ground truth (left) for co-activations (top) and causal modulations (bottom). (H) Hierarchical clustering result from Γ (top) and comparison between ground truth and output directional network-to-network graphs (bottom). Red/blue edges denote up-regulatory/down-regulatory influences and the arrow stands for the direction of the modulation. SLR: sparse logistic regression.

influences. As regularisation became still weaker, the log-likelihood decreased back, because many noisy coefficients then pollute the estimates compared to the ground truth.

Figures 2(E) and (F) further disentangle the results from the log-likelihood computation by separating the optimal $\xi^{*(r)}$ and $\lambda^{*(r)}$ for both types of transitions (baseline to active, or active to baseline).

In figure 2(E), the regions that belong to the networks that do not receive any causal modulation (networks 1, 2, 5 and 7, respectively colour-coded in turquoise, orange, green and dark brown) are associated to higher $\xi^{*(r)}$ values (as co-activations then dominate in such settings) and to higher $\lambda^{*(r)}$ values as well (since they are associated to less coefficients overall). Conversely and fitting the above log-likelihood-based observations, regions from network 6 (light brown)—the most heavily causally modulated—show the lowest $\xi^{*(r)}$ and $\lambda^{*(r)}$ values. In figure 2(F), it can be seen that, expectedly given the mathematical underpinnings of the framework, regions with a higher overall density of coefficients ($\rho_\Gamma + \rho_B$) were linked to larger $\lambda^{*(r)}$. Meanwhile, $\xi^{*(r)}$ was lower/larger for regions with a balance in incoming modulations leaning towards the causal/co-activation case.

The final outputs of the SLR framework, when the probabilistic influences onto each region are sampled from its optimal $\xi^{*(r)}/\lambda^{*(r)}$ values, are depicted in figure 2(G) (right half) and compared to the ground truth (left half). Similarity was very elevated ($S_\Gamma = 0.98$ and $S_B = 0.9$). Accordingly, hierarchical clustering from $\mathbf{\Gamma}$ could separate all 7 networks with a perfect purity of 1 (figure 2(H), top half) and the directional graph representation generated from the SLR framework exactly matched the ground truth one (figure 2(H), bottom half).

Figure 3 considers the outputs from our framework upon different network structures and co-activation/causal balances. In figure 3(A), we considered $N = 3$ networks with only one up-regulatory influence from network 1 to network 2. Similarity values were high for both types of coefficients ($S_\Gamma = 0.97$, $S_B = 0.71$), network assignment could be perfectly retrieved and so could the network-wise directional graph. Similar observations were made when instead, 5 of the 6 possible cross-network couplings were included (figure 3(B)), demonstrating the flexibility of our SLR framework. Note that the median of optimal λ values across regions and transition types was larger in the former case ($\lambda_{\text{med}} = 120$ as opposed to $\lambda_{\text{med}} = 60$) and so was the median of ξ values ($\xi_{\text{med}} = 0.75$ as opposed to $\xi_{\text{med}} = 0.5$). This is unsurprising, since the former case included less coefficients to retrieve overall and a balance more in favour of co-activations.

Figures 3(C) and (D) depict the results from conceptually similar simulations when conducted with $N = 7$ networks instead. Given the elevated similarities between ground truth and SLR output matrices and the perfect network assignments, it can be seen that our framework graciously handles changes in the underlying network structure. There was only one disagreement between ground truth and extracted values at the level of the directional network-wise graph representation in figure 3(D): while all true edges were correctly retrieved, an erroneous one

depicted a down-regulatory influence of network 5 onto network 6 (note that this can already be seen from **B**, where some negative-valued probabilistic influences populate the associated patch of the matrix, as labelled in orange). However, this false positive edge was also the weakest of all the retrieved ones.

3.2. Comparison of performance to other approaches

In figure 4, we compare the performance of the SLR framework to other approaches: for the quality of co-activation coefficients, we consider the graphical lasso (GLasso) and a point process analysis (PPA). Regarding causal modulations, we consider a cross-spectral density (CSD)-based approach and an order-1 multivariate autoregressive model (MAR). The ground truth for these simulations is depicted in figure 4(A): it is the same as that probed in figure 2, but here, we consider the evolution of quality metrics as noise level and/or the number of training subjects change(s).
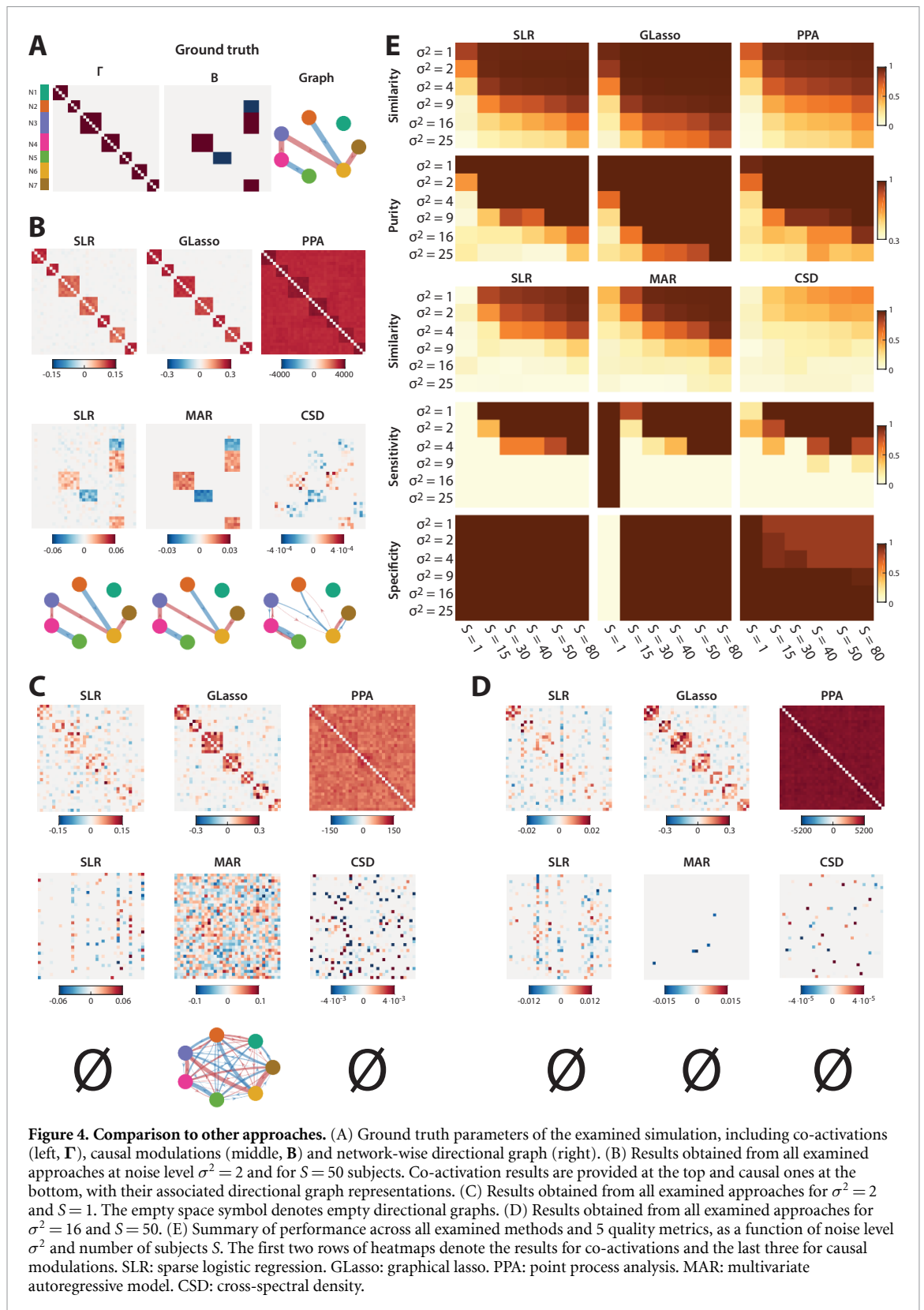
Figure 4(B) presents the results obtained by all approaches at $\sigma^2 = 2$ with $S = 50$ subjects (i.e., same parameters as in figure 2). Using GLasso, the co-activation structure is clearly retrieved, as in the SLR case. With PPA, it is also possible to resolve the different networks, but background intensity is larger. Regarding causal modulations, the use of MAR results in excellent outputs and in a perfect directional graph representation, as with the SLR framework. However, a CSD approach instead yields a sparser matrix of coefficients; while true modulations are indeed pinpointed, the anti-symmetrical nature of the output matrix prevents from inferring if network i up-regulates network j, or if instead, network j down-regulates network i. This is also seen in the directional graph representation, where edges always appear in pairs. Furthermore, the more prominent of the two edges is not always the correct one: while network 3 up-regulates network 6, the CSD approach instead yields a larger edge for a down-regulation from network 6 to network 3.

The outputs provided by all approaches are examined under more challenging settings in figures 4(C) ($\sigma^2 = 2$ and $S = 1$) and 4(D) ($\sigma^2 = 16$ and $S = 50$). While GLasso still enables to retrieve the majority of ground truth co-activations, PPA becomes almost incapable to do so (indeed, only very faint network-like patterns are seen in the associated matrices). The outputs from the SLR framework are intermediate: less coefficients are retrieved than in the GLasso case, but an underlying structure can still be discerned.

As for causal modulations, none of the outputs at such challenging noise settings are truly satisfying. Note that the MAR results in the single-subject case (figure 4(C)) are not sparse, because our cross-subject null strategy would not be applicable in that setting. Accordingly, it is the only case for which a non-empty
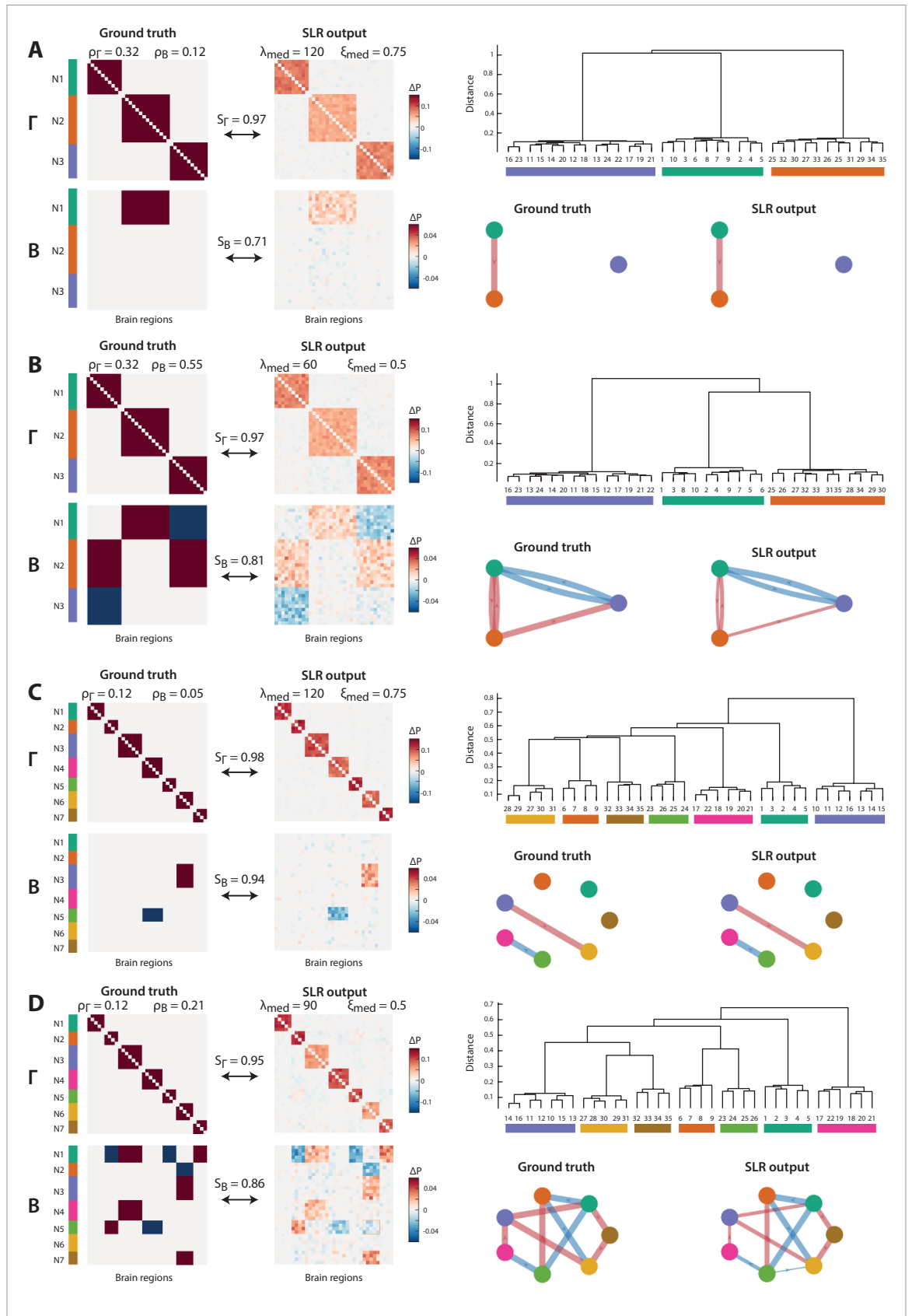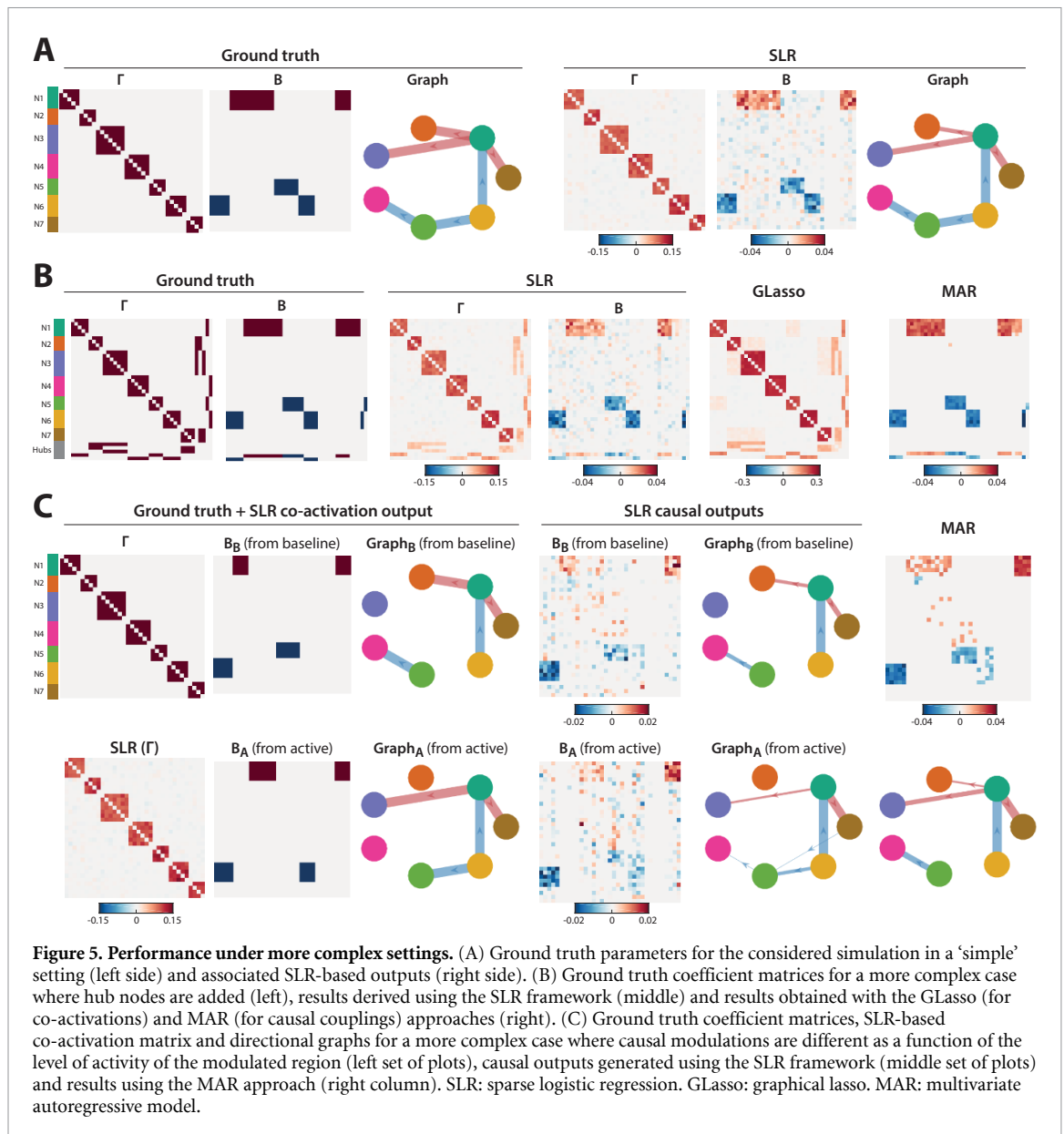
Figure 3. Flexibility of the framework to changes in network structure and causal modulation density. (A) For $N = 3$ networks and a weak extent of causal modulations ($\rho_B = 0.12$ compared to $\rho_\Gamma = 0.32$), ground truth co-activation and causal coefficients (left column), outputs from the SLR framework (middle column), dendrogram obtained upon hierarchical clustering from Γ (top right) and network-wise directional graphs for the ground truth and the SLR output cases (bottom right). λ_{med} and ξ_{med} are the median optimal λ and ξ values across all regions and transition types. S_Γ and S_B are the similarities between ground truth and output matrices for the co-activation and causal modulation cases, respectively. (B) Obtained results for a larger amount of cross-network causal modulations ($\rho_B = 0.55$) and $N = 3$ networks. (C) Obtained results for $N = 7$ networks and a low amount of cross-network causal modulations ($\rho_B = 0.05$ compared to $\rho_\Gamma = 0.12$). (D) Obtained results for $N = 7$ networks and a greater amount of cross-network causal modulations ($\rho_B = 0.21$). The orange contour in the output **B** matrix shows the patch yielding the false positive edge found in the directional graph representation. SLR: sparse logistic regression.

Figure 4. Comparison to other approaches. (A) Ground truth parameters of the examined simulation, including co-activations (left, **Γ**), causal modulations (middle, **B**) and network-wise directional graph (right). (B) Results obtained from all examined approaches at noise level $\sigma^2 = 2$ and for $S = 50$ subjects. Co-activation results are provided at the top and causal ones at the bottom, with their associated directional graph representations. (C) Results obtained from all examined approaches for $\sigma^2 = 2$ and $S = 1$. The empty space symbol denotes empty directional graphs. (D) Results obtained from all examined approaches for $\sigma^2 = 16$ and $S = 50$. (E) Summary of performance across all examined methods and 5 quality metrics, as a function of noise level σ^2 and number of subjects S. The first two rows of heatmaps denote the results for co-activations and the last three for causal modulations. SLR: sparse logistic regression. GLasso: graphical lasso. PPA: point process analysis. MAR: multivariate autoregressive model. CSD: cross-spectral density.

directional graph is retrieved: while containing many erroneous edges, the strongest ones nicely match the ground truth.

The full results of our comparative assessment are summarised in figure 4(E). It can be seen that when noise is increased (going from top to bottom in a given heatmap), or when less subjects are available for

parameter estimation (going from right to left), performance degrades as quantified by almost all metrics. The only exception is specificity, because causal modulation outputs will become fully sparse under more challenging simulation circumstances, thus preventing the detection of false positives (except for MAR at $S = 1$, as mentioned above).

Figure 5. Performance under more complex settings. (A) Ground truth parameters for the considered simulation in a 'simple' setting (left side) and associated SLR-based outputs (right side). (B) Ground truth coefficient matrices for a more complex case where hub nodes are added (left), results derived using the SLR framework (middle) and results obtained with the GLasso (for co-activations) and MAR (for causal couplings) approaches (right). (C) Ground truth coefficient matrices, SLR-based co-activation matrix and directional graphs for a more complex case where causal modulations are different as a function of the level of activity of the modulated region (left set of plots), causal outputs generated using the SLR framework (middle set of plots) and results using the MAR approach (right column). SLR: sparse logistic regression. GLasso: graphical lasso. MAR: multivariate autoregressive model.
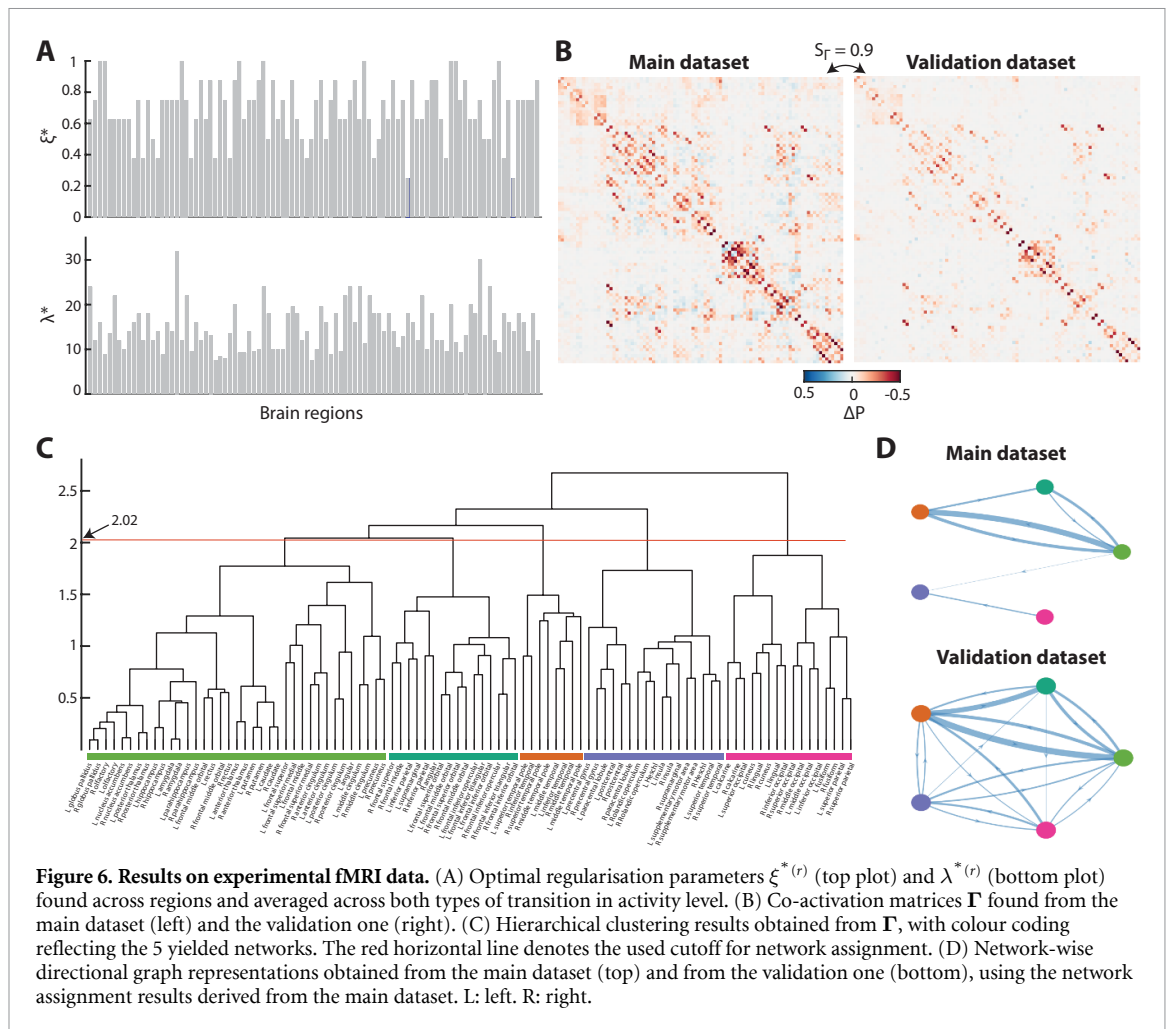
While qualitatively similar, it can also clearly be seen that performance degrades most rapidly in the PPA case (for co-activation metrics) and in the CSD case (for causal metrics). GLasso and MAR are the most precise approaches, shortly followed by our SLR framework, which performs slightly worse under more challenging settings. However, we should emphasise that co-activations and causal modulations are then jointly derived, instead of only one of the two sets with other competing approaches.

To complement the above, figure 5 provides evidence that our SLR framework may also be of use in more complex (and possibly more realistic) settings. We first consider a new simple simulation case, as depicted in figure 5(A) (noise settings are similar to those employed elsewhere, with $\sigma^2 = 2$ and $S = 50$): unsurprisingly, the retrieved coefficient maps convincingly reveal the true underlying co-activation and causal structures of the system.

In figure 5(B), we consider a first increase in complexity by adding in 5 hub regions that jointly co-activate with two separate networks. Since we now deal with $R = 40$ regions, the number of subjects used for the estimates was increased to $S = 65$ (see section 2.4). Recovery of the ground truth remained excellent despite this additional layer of complexity, both for $\boldsymbol{\Gamma}$ and for $\mathbf{B}$. Interestingly, while with MAR the causal structure was still very cleanly recovered, GLasso started mistakenly revealing some cross-network co-activations between networks 2 and 3, 1 and 5, 2 and 7 and 3 and 7 (see the off-diagonal patches in the associated matrix).

In figure 5(C), we instead consider a ground truth scenario in which the causal modulations differ from one type of transition in activity level to the other. In more details, network 1 only up-regulates network 2 when the latter is at a baseline level of activity; mechanistically speaking, this could reflect the

Figure 6. Results on experimental fMRI data. (A) Optimal regularisation parameters $\xi^{*(r)}$ (top plot) and $\lambda^{*(r)}$ (bottom plot) found across regions and averaged across both types of transition in activity level. (B) Co-activation matrices $\mathbf{\Gamma}$ found from the main dataset (left) and the validation one (right). (C) Hierarchical clustering results obtained from $\mathbf{\Gamma}$, with colour coding reflecting the 5 yielded networks. The red horizontal line denotes the used cutoff for network assignment. (D) Network-wise directional graph representations obtained from the main dataset (top) and from the validation one (bottom), using the network assignment results derived from the main dataset. L: left. R: right.

fact that network 2 then starts self-sustaining itself, becoming immune to external modulations when it is active. Conversely, network 1 only up-regulates network 3 when the latter is active. Similar state-specific modulations are also introduced for down-regulatory influences from network 5 to 4 (only when network 4 is at baseline level of activity) and from network 6 to 5 (only when network 5 is active). Logically, a MAR approach cannot disentangle both ground truths and provides a trade-off solution that mixes the different types of edges, while even discarding one (the down-regulatory modulation from network 6 to 5). Using our SLR framework, while the results remain somehow noisy, both scenarios can be disentangled, as seen from the network-wise directional graph representations, and this is so despite still resorting to only $S = 50$ subjects in the estimations.

3.3. Application of the framework to experimental fMRI data

Figure 6 shows the results when applying our framework to experimental fMRI data ($R = 94$ regions). A main dataset of $S = 350$ subjects was used to derive the SLR outputs and the framework was then applied to a separate validation set of $S_{\text{VAL}} = 350$ subjects at the

extracted region-specific optimal regularisation parameters.

Computationally speaking, on an Intel Xeon Platinum 8160 CPU at 2.1 GHz with 24 cores, 512 GB RAM and Ubuntu 18.04, z-scoring and binarisation of the time courses were always achieved in a few seconds, while the selection of time points featuring both types of transition took in the order of half an hour per dataset. As for SLR framework steps, average computational time values across regions and regularisation levels were 0.13±0.03 s for the computation of $z_t^{(r)}$ and $\omega_t^{(r)}$ and 153.51±45.6 s for the computation of the $\alpha^{(r)}$, $\beta^{(r)}$ and $\gamma^{(r)}$ coefficients. Finally, the evaluation of the log-likelihood took 0.13±0.02 s. Thus, the two most time-consuming factors were the selection of time points and most importantly, the computation of the coefficients.

In figure 6(A), we can visualise optimal regularisation parameters extracted across all the regions at hand. Values for $\xi^{*(r)}$ fluctuated across areas, highlighting how some brain regions may highlight more co-activations, while others undergo more causal modulations. In general, the values were however closer to 1 (0.69±0.24, with a median of 0.75), denoting that co-activation is globally more influential than causal interplays. Regarding $\lambda^{*(r)}$, a variable range of

values could also be seen across areas, denoting that some are more heavily interconnected with the rest of the brain circuitry than others.

Figure 6(B) depicts the co-activation matrices Γ on the main and on the validation datasets. Visual agreement between both outputs is evident and this is quantitatively confirmed by a large similarity value of $S_\Gamma = 0.9$. Co-activations resulting from the main dataset then underwent hierarchical clustering, revealing a complex multi-scale organisation (figure 6(C)). An extensive analysis of such data would require to investigate the results at various possible numbers of clusters, gradually cutting the dendrogram at lower distance cutoffs. However, since our purpose only was the preliminary experimental application of our SLR framework, here, we solely considered one such cutoff value ($d_{cut} = 2.02$, as depicted by the horizontal red line; see section 2.7 for details).

With this partitioning, 5 distinct networks were extracted: network 1 (in green) included all subcortical regions as well as medial frontal, posterior cingulate and angular areas characteristic from the default mode network (DMN) [54]. Note that subcortical and DMN regions would be segmented into two separate networks at a lower cutoff value. Network 2 (in turquoise) primarily featured frontal areas reminiscent of executive control, while network 3 (in orange) included temporal regions, likely representing an auditory network. Network 4 (in purple) included precentral, paracentral, postcentral and supplementary motor areas typically associated to somatomotor function and also comprised the bilateral insula. Finally, network 5 (in pink) exclusively consisted in occipital regions characteristic of the visual system; it would be further split into primary and secondary sub-systems at a lower cutoff. Overall, the obtained network assignments are thus in line with RS neurophysiological knowledge. In addition, when regional assignments were extracted from the validation dataset, purity as computed between both clustering outcomes showed a fair value of 0.64, highlighting somehow generalisable subdivision of the regional data into networks.

Causal modulations considered across the extracted 5 networks are displayed in figure 6(D) for the main dataset and for the validation one (using similar regional assignments). Several observations can be made: first, more causal modulations are retrieved in the validation graph, possibly owing to the fact that the SLR algorithm was rerun only at optimal regularisation values, thus yielding slightly less tailored estimates to the data at hand. Second, the overlap with the main dataset results is nonetheless quite good: all the edges found from the main dataset are indeed present in the validation one and are also those with the strongest values. This provides confidence that the directional cross-network couplings seen in the main dataset are generalisable and can thus soundly

be discussed. Third, all these retrieved causal modulations are negative-valued: this means that when the modulating network is active, it will down-regulate the activity of the modulated network (either by making it more likely to transit from the active to the baseline state, or by making it less likely to become active). In particular, the subcortical/DMN, executive and temporal networks primarily inhibit each other by this mean, while visual and somatomotor networks remain more independent in their activity.

4. Discussion

In this work, we introduced a novel mathematical framework enabling to jointly derive the patterns of co-activation between brain regions, reflective of the brain's functional organisation as a set of RSNs [4, 6], and additional cross-regional causal modulations that enable to go beyond this network-level characterisation and also model more subtle cross-regional interplays. One can conceive our strategy as a joint recovery of FC (embedded in the Γ matrix) and EC (in $\mathbf{B}$).

Our strategy is an improvement over previous work that also used a logistic regression characterisation to describe causal interactions between functional brain networks [41]: in this former methodology, network maps had to be computed in a separate analytical step, prior to the establishment of their causal interplays. As such and much like the majority of other prominent dynamic FC approaches—see for instance [16, 19, 22, 55], more subtle relationships at a smaller spatial scale than that of RSNs are then lost.

On simulated data, both co-activation and causal coefficient sets could accurately be retrieved by our framework despite marked noise, and this held true in various configurations regarding the number of simulated networks and the balance between co-activations and causal influences (figures 2 and 3). In all the assessed cases, clear maxima could be observed in the log-likelihood curves of the simulated regions, confirming the efficiency of our cross-validation strategy in selecting meaningful regularisation parameters tailored to each area.

In the majority of our simulations, we considered enough data points for accurate estimation of the full model, as around 12 data points were available per parameter. Upon the investigation of more challenging cases, either due to increased noise or to a lower available amount of subjects for estimation (figure 4), only a restricted subset of ground truth entries were recovered, owing to the ℓ_1 norm properties [56]. These correctly retrieved coefficients were co-activations, not causal couplings, indicating that the former could more easily be extracted from our simulations. This is not so surprising given the used simulation strategy, where co-activation was modelled by simulating two identical time courses before noise addition, while causal couplings only changed

the probability to transit across activity levels. It will be interesting to consider alternative simulation schemes in future work, to more comprehensively evaluate the ease with which each coefficient type can be extracted.

Across the assessed noise and dataset size settings, our SLR framework was, on the whole, competitive in comparison to other existing methods. It globally outperformed PPA for the estimation of co-activations, and CSD-based retrieval of causal couplings. In addition, it came a close second to the widely applied GLasso and MAR in the respective recovery of co-activation and causal coefficients, only providing worse performance in the most challenging investigated settings.

Importantly, the worse outcomes of the PPA and CSD-based methods in our analyses do not imply that such tools are useless: in fact, one of the major assets of PPA is its computational speed compared to classical FC estimation [46] and indeed, it was the fastest of the examined pipelines. As for the use of CSD information to estimate causal modulations, results from such a family of approaches show an anti-symmetrical structure [57], which does not accommodate our underlying simulation assumptions as well as for other methods. In sum, which tools perform the best always depends on the considered metrics and simulation specificities.

In any case, our framework showed promising potential from the examined angles, especially given that it is the only of the assessed approaches that jointly retrieves co-activation and causal information at once. Theoretically speaking, it also enables to go even further, as two separate maps are obtained: one for the baseline-to-active transitions and one for the active-to-baseline ones. In most of the presented content, we treated the $0 \rightarrow +1$ and $+1 \rightarrow 0$ transitions as mirrors of each other, subtracting both sets of probabilistic couplings to obtain the analysed outputs. However, more complex information may lie within the individual coefficient matrices. Figure 5 showed that such activity state-specific modulatory influences can indeed be disentangled, although we leave more detailed investigations for future work.

On experimental fMRI data (figure 6), the optimal balance between co-activations and causal modulations—rendered by the $\xi^{*(r)}$ parameter—fluctuated across regions, evidencing the fact that both types of cross-regional interactions are required to accurately describe functional brain dynamics, in a way that is not spatially trivial. While the obtained median value of 0.75 indicates that on the whole, co-activations play a somehow dominating role, these results nonetheless highlight the importance of developing methodological approaches that do not only focus on one viewpoint, but instead attempt to jointly capture co-activations and causal interplays.

An important aspect to keep in mind—and a limitation of the present work—is the fact that although the SLR framework goes beyond the network-level spatial scale by revealing region-wise interactions, it still considers a set of spatially fixed parcels in doing so. The resolution of the used atlas can then be expected to influence obtained results and here, we only considered $R = 94$ separate areas, which remains a modest amount compared to the most state-of-the-art parcellations [50, 58]. This was, however, necessary to ensure the presence of enough data points for sound estimation.

Several technical developments may be envisioned to further improve our approach. First, the purely ℓ_1 regularisation strategy could be turned into an *elastic net* mix between ℓ_1 and ℓ_2 norms [59], but it would then come at the cost of an extra free parameter to specify. Second, neurobiologically relevant additional assumptions could be introduced to the model formulation, such as symmetry and non-negativity in the co-activation matrix $\boldsymbol{\Gamma}$, or the fact that co-activations and causal influences should be mutually exclusive.

Third, instead of the probability to transit from a given state of activity to another, one could consider the likelihood to show an innovation [19] (that is, go up or down in activity regardless of the exact starting point). By this mean, the current framework could seamlessly be generalised to more than only 2 states of activity, which may better represent the dynamics of some brain regions. This information is already available (by comparison to phase-randomised null data) from the *total activation* pipeline used in the deconvolution of the analysed fMRI data [19, 52, 53]. An additional interest would then be the easier comparison of results obtained from datasets acquired at various TRs, so that the increasingly understood specificities of fast TR datasets [60, 61] can be better disentangled from more general effects. To do so, one could determine whether a transient has just occurred prior to the assessed time point by jointly examining a span of a few samples ($t - 1$, $t - 2$, *etc*).

Fourth, the current framework enables to go from networks to regions, but one could push the same reasoning further by attempting to further separate this regional categorisation into smaller individual units—finer-grained parcels, or voxels. Such a multi-scale analysis would enable to dig into important aspects that may for now be blurred, such as the notable idiosyncrasy in FC patterns and network identities known to exist across subjects [62, 63].

Finally, a few promising practical applications of our framework can be foreseen: first, it will be exciting to compare co-activation and causal coefficients across different subject populations (*e.g.*, a set of healthy volunteers as opposed to a diseased population). To do so, bootstrapping could be conducted on each population and statistical testing could then be applied for each coefficient of interest. The examination of subject-specific properties will, however, be more challenging to address, as typically available

amounts of data only permit sound population-wise inference. Second, another possible application could be in *hyperscanning* [64], where two subjects are scanned in parallel while they interact. Co-activations, or causal modulations, could be quantified across both subjects as a way to shed light on the functional underpinnings of cooperative processing. Third, the specificities of our framework may be even better suited to the analysis of other data modalities for which temporal resolution enables to more closely track neuronal activity; applications to magnetoencephalography, electroencephalography or electrocorticography datasets are thus interesting avenues to explore.

Acknowledgments

ORCID iDs

Thomas A W Bolton ⬤ https://orcid.org/0000-0002-2081-4031
Eneko Uruñuela ⬤ https://orcid.org/0000-0001-6849-9088
Ye Tian ⬤ https://orcid.org/0000-0003-3107-5550
Andrew Zalesky ⬤ https://orcid.org/0000-0003-2298-9908
César Caballero-Gaudes ⬤ https://orcid.org/0000-0002-9068-5810
Dimitri Van De Ville ⬤ https://orcid.org/0000-0002-2879-3861

References

[1] Mišić B and Sporns O 2016 From regions to connections and networks: new bridges between brain and behavior *Current Opinion Neurobiol.* **40** 1–7
[2] Friston K J 1994 Functional and effective connectivity in neuroimaging: a synthesis *Human Brain Mapping* **2** 56–78
[3] Smith S *et al* 2011 Network modelling methods for fMRI *Neuroimage* **54** 875–91
[4] Damoiseaux J S, Rombouts S A R, Barkhof F, Scheltens P, Stam C J, Smith S M and Beckmann C F 2006 Consistent resting-state networks across healthy subjects *Proc. Natl Acad. Sci.* **103** 13848–53
[5] Power J D, Fair D A, Schlaggar B L and Petersen S E 2010 The development of human functional brain networks *Neuron* **67** 735–48
[6] Yeo B T T *et al* 2011 The organization of the human cerebral cortex estimated by intrinsic functional connectivity *J. Neurophysiol.* **106** 1125–65
[7] Greicius M 2008 Resting-state functional connectivity in neuropsychiatric disorders *Current Opinion Neurol.* **21** 424–30
[8] Bressler S L and Menon V 2010 Large-scale brain networks in cognition: emerging methods and principles *Trends Cognitive Sci.* **14** 277–90
[9] van den Heuvel M P and Hulshoff Pol H E 2010 Exploring the brain network: a review on resting-state fMRI functional connectivity *Eur. Neuropsychopharmacol.* **20** 519–34
[10] Chang C and Glover G H 2010 Time-frequency dynamics of resting-state brain connectivity measured with fMRI *Neuroimage* **50** 81–98
[11] Keilholz S D, Caballero-Gaudes C, Bandettini P, Deco G and Calhoun V D 2017 Time-resolved resting state fMRI analysis: current status, challenges and new directions *Brain Connectivity* **7** 465–81
[12] Karahanoğlu F I and Van De Ville D 2017 Dynamics of large-scale fMRI networks: deconstruct brain activity to build better models of brain function *Current Opinion Biomed. Eng.* **3** 28–36
[13] Preti M G, Bolton T A W and Van De Ville D 2017 The dynamic functional connectome: state-of-the-art and perspectives *Neuroimage* **160** 41–54
[14] Lurie D J *et al* 2020 Questions and controversies in the study of time-varying functional connectivity in resting fMRI *Network Neurosci.* **4** 30–69
[15] Leonardi N, Richiardi J, Gschwind M, Simioni S, Annoni J, Schluep M, Vuilleumier P and Van De Ville D 2013 Principal components of functional connectivity: a new approach to study dynamic brain connectivity during rest *Neuroimage* **83** 937–50
[16] Allen E A, Damaraju E, Plis S M, Erhardt E B, Eichele T and Calhoun V D 2014 Tracking whole-brain connectivity dynamics in the resting state *Cerebral Cortex* **24** 663–76
[17] Yaesoubi M, Miller R L and Calhoun V D 2015 Mutually temporally independent connectivity patterns: a new framework to study the dynamics of brain connectivity at rest with application to explain group difference based on gender *Neuroimage* **107** 85–94
[18] Liu X, Chang C and Duyn J H 2013 Decomposition of spontaneous brain activity into distinct fMRI co-activation patterns *Front. Syst. Neurosci.* **7** 1–11
[19] Karahanoğlu F I and Van De Ville D 2015 Transient brain activity disentangles fMRI resting-state dynamics in terms of spatially and temporally overlapping networks *Nat. Commun.* **6** 7751
[20] Smith S M *et al* 2012 Temporally-independent functional modes of spontaneous brain activity *Proc. Natl Acad. Sci.* **109** 3131–6
[21] Eavani H, Satterthwaite T D, Gur R E, Gur R C and Davatzikos C 2013 Unsupervised learning of functional network dynamics in resting state fMRI *Lecture Notes in Computer Science* **7917** 426–37
[22] Vidaurre D, Smith S M and Woolrich M W 2017 Brain network dynamics are hierarchically organized in time *Proc. Natl Acad. Sci.* **114** 201705120
[23] Zhang G, Cai B, Zhang A, Stephen J M, Wilson T W, Calhoun V D and Wang Y W 2019 Estimating dynamic functional brain connectivity with a sparse hidden markov model *IEEE Trans. Med. Imaging* **39** 488–98
[24] Chen S, Langely J, Chen X and Hu X 2016 Spatiotemporal modeling of brain dynamics using resting-state functional magnetic resonance imaging with Gaussian hidden Markov model *Brain Topography* **6** 326–34

[25] Chen T, Cai W, Ryali S, Supekar K and Menon V 2016 Distinct global brain dynamics and spatiotemporal organization of the salience network *PLOS Biology* **14** 1–21

[26] Pedersen M, Zalesky A, Omidvarnia A and Jackson G D 2018 Multilayer network switching rate predicts brain performance *Proc. Natl Acad. Sci.* **115** 13376–81

[27] Kottaram A, Johnston L, Ganella E, Pantelis C, Kotagiri R and Zalesky A 2018 Spatio-temporal dynamics of resting-state brain networks improve single-subject prediction of schizophrenia diagnosis *Human Brain Mapping* **39** 3663–81

[28] Iraji A, Miller R, Adali T and Calhoun V D 2020 Space: A missing piece of the dynamic puzzle *Trends Cognitive Sci.* **24** 135–49

[29] Majeed W, Magnuson M, Hasenkamp W, Schwarb H, Schumacher E H, Barsalou L and Keilholz S D 2011 Spatiotemporal dynamics of low frequency BOLD fluctuations in rats and humans *Neuroimage* **54** 1140–50

[30] Mitra A, Snyder A Z, Blazey T and Raichle M E 2015 Lag threads organize the brain's intrinsic activity *Proc. Natl Acad. Sci.* **112** E2235–E2244

[31] Takeda Y, Hiroe N, Yamashita O and Sato M 2016 Estimating repetitive spatiotemporal patterns from resting-state brain activity data *Neuroimage* **133** 251–65

[32] Razi A, Seghier M L, Zhou Y, McColgan P, Zeidman P, Park H, Sporns O, Rees G and Friston K J 2017 Large-scale DCMs for resting-state fMRI *Network Neurosci.* **1** 222–41

[33] Prando G, Zorzi M, Bertoldo A, Corbetta M, Zorzi M and Chiuso A 2020 Sparse DCM for whole-brain effective connectivity from resting-state fMRI data *Neuroimage* **208** 116367

[34] Lennartz C, Schiefer J, Rotter S, Hennig J and LeVan P 2018 Sparse estimation of resting-state effective connectivity from fMRI cross-spectra *Front. Neurosci.* **12** 287

[35] Gilson M, Moreno-Bote R, Ponce-Alvarez A, Ritter P and Deco G 2016 Estimation of directed effective connectivity from fMRI functional connectivity hints at asymmetries of cortical connectome *PLOS Computat. Biol.* **12** e1004762

[36] Gilson M *et al* 2020 Model-based whole-brain effective connectivity to study distributed cognition in health and disease *Network Neurosci.* **4** 338–73

[37] Liégeois R, Laumann T O, Snyder A Z, Zhou J and Yeo B T T 2017 Interpreting temporal fluctuations in resting-state functional connectivity MRI *Neuroimage* **163** 437–55

[38] Liégeois R, Li J, Kong R, Orban C, Van De Ville D, Ge T, Sabuncu M R and Yeo B T T 2019 Resting brain dynamics at different timescales capture distinct aspects of human behavior *Nat. Commun.* **10** 2317

[39] Friedman J, Hastie T and Tibshirani R 2010 Regularization paths for generalized linear models via coordinate descent *J. Statistical Software* **33** 1–22

[40] Christoff K, Irving Z C, Fox K C R, Spreng R N and Andrews-Hanna J R 2016 Mind-wandering as spontaneous thought: a dynamic framework *Nat. Rev. Neurosci.* **17** 718–31

[41] Bolton T A W, Tarun A, Sterpenich V, Schwartz S and Van De Ville D 2017 Interactions between large-scale functional brain networks are captured by sparse coupled HMMs *IEEE Trans. Med. Imaging* **37** 230–40

[42] Kang J, Pae C and Park H 2019 Graph-theoretical analysis for energy landscape reveals the organization of state transitions in the resting-state human cerebral cortex *PLOS ONE* **14** 0222161

[43] Friedman J, Hastie T, Höfling H and Tibshirani R *et al* 2007 Pathwise coordinate optimization *Ann. Appl. Stat.* **1** 302–32

[44] Yang Z, Hao T, Dikmen O, Chen X and Oja E 2012 Clustering by nonnegative matrix factorization

using graph random walk *Adv. Neural Inf. Proc. Syst.* pp 1079–87

[45] Friedman J, Hastie T and Tibshirani R 2008 Sparse inverse covariance estimation with the graphical lasso *Biostatistics* **9** 432–41

[46] Tagliazucchi E, Siniatchkin M, Laufs H and Chialvo D R 2016 The voxel-wise functional connectome can be efficiently derived from co-activations in a sparse spatio-temporal point-process *Front. Neurosci.* **10** 1–13

[47] Van Essen D C, Smith S M, Barch D M, Behrens T E J, Yacoub E and Ugurbil K 2013 The WU-Minn human connectome project: an overview *Neuroimage* **80** 62–79

[48] Smith S M *et al* 2013 Resting-state fMRI in the human connectome project *Neuroimage* **80** 144–68

[49] Tian Y, Margulies D S, Breakspear M and Zalesky A 2020 Hierarchical organization of the human subcortex unveiled with functional connectivity gradients *Nat. Neurosci.* **23** 1421–32

[50] Glasser M F *et al* 2016 A multi-modal parcellation of human cerebral cortex *Nature* **536** 171–8

[51] Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B and Loliot M 2002 Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain *Neuroimage* **15** 273–89

[52] Karahanoğlu F I, Caballero-Gaudes C, Lazeyras F and Van De Ville D 2013 Total activation: fMRI deconvolution through spatio-temporal regularization *Neuroimage* **73** 121–34

[53] Bolton T A W, Farouj Y, Inan M and Van De Ville D 2019 Structurally-informed deconvolution of functional magnetic resonance imaging data *16th Int. Symp. on Biomedical Imaging (ISBI)* 1545–9

[54] Buckner R L, Andrews-Hanna J R and Schacter D L 2008 The brain's default network *Ann. New York Acad. Sci.* **1124** 1–38

[55] Liu X and Duyn J H 2013 Time-varying functional network information extracted from brief instances of spontaneous brain activity *Proc. Natl Acad. Sci.* **110** 4392–7

[56] Tibshirani R 1994 Regression shrinkage and selection via the LASSO *J. R. Stat. Soc.* B **58** 267–88

[57] Friston K J, Kahan J, Biswal B and Razi A 2014 A DCM for resting state fMRI *Neuroimage* **94** 396–407

[58] Schaefer A, Kong R, Gordon E M, Laumann T O, Zuo X, Holmes A J, Eickhoff S B and Yeo B T T 2017 Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI *Cerebral Cortex* **28** 3095–114

[59] Zou H and Hastie T 2005 Regularization and variable selection via the elastic net *J. R. Stat. Soc.* B **67** 301–20

[60] Chen J E, Polimeni J R, Bollmann S and Glover G H 2019 On the analysis of rapidly sampled fMRI data *Neuroimage* **188** 807–20

[61] Power J D, Lynch C J, Silver B M, Dubin M J, Martin A and Jones R M 2019 Distinctions among real and apparent respiratory motions in human fMRI data *Neuroimage* **201** 116041

[62] Gordon E M *et al* 2017 Precision functional mapping of individual human brains *Neuron* **95** 791–807

[63] Kong R *et al* 2019 Spatial topography of individual-specific cortical networks predicts human cognition, personality and emotion *Cerebral Cortex* **29** 2533–51

[64] Montague P R *et al* 2002 Hyperscanning: simultaneous fMRI during linked social interactions *Neuroimage* **16** 1159–64