1. Supplemental Information

1.1. Interpreting Model Parameters

As discussed in the main text, the neural modeling framework of MINDy is inherently phenomenological in that it is not directly derived from biophysical first-principals. The weight parameter ($W$), for instance, serves to measure effective connectivity and should not be confused with synaptic efficacy (or any other directly measurable anatomical metric, such as white matter integrity). The phenomenological nature of these equations gives tractability to the fitting problem. However, this fact does not preclude the model’s interpretable and predictive nature. The parametric form that we have chosen leads itself to interpretability by separating the dynamics into three distinct components, interregional-signaling, local decay, and a nonlinear mapping between local excitation and output, which parallel the components used in conventional neural mass models. In the following subsections, we present fuller descriptions of the potential relations between model parameters and underlying biological processes.

1.1.1. Interpreting Model Weights

In our model, the connectivity matrix defines the causal ability of mean regional activity in the sender region to monotonically change mean regional activity levels in the receiving region within a specific time window.
This causal influence has standardly been termed effective connectivity within the fMRI (and EEG) literature. More precisely, however, in the model the effects must begin within the duration of one TR (720ms in our case), and last long enough to invoke a metabolic response. As such, our definition is slightly more specific than the notion of effective connectivity, as we specify that these relations must be weakly monotone: all else being equal, increasing (decreasing) the activity of region A will never decrease (increase) the activity of region B (Fig. 1 A, B). We use the term weakly monotone as regions may exhibit saturated activity within our model and thus have little room to increase/decrease. In contrast with our definition, effective connectivity does not specify the nature of the relationship between regions. In this case, model-free methods such as transfer entropy ([1]) can be employed to study non-monotone relationships within a very small number of dimensions. We also define our temporal range of interactions to be between 500ms and 2s. We do not use a tighter temporal range such as 500ms-1s as temporal variations inherent in BOLD imaging, such as physiological changes in the hemodynamic response (e.g. under anesthesia; [2],[3]) lead to some uncertainty in timing. In addition, there are methodological limitations inherent in rapid acquisition methods such as multiband imaging ([4]), which have led some investigators to prefer TR’s closer to 2s. In either case, our definition limits the duration of interest to the order of a typical fMRI trial.

The monotone and temporal constraints can also differentiate our W matrix from structural connectivity, the latter of which does not necessarily reflect how regions interact. If two regions communicate in a very heterogeneous manner and/or these interactions only result in very transient changes, these regions would not be connected in our W matrix, even if a direct white matter tract linked them (Fig. 1B). Of course, this scenario also suggests that those portions of brain would also not meet the definition of a cortical parcel due to their heterogeneity. Finer parcellation schemes lead to correspondingly more homogeneous “regions”, so, with a sufficiently high resolution parcellation, we expect that most forms of structural connectivity would meet the monotone requirement, with a single cell as the theoretical limiting case. In summary, our form of connectivity in the W matrix describes not just the ability of regions to causally influence each other but to do so with easily predictable (monotone) consequences in a specific time scale. For ease of presentation, however, we use the term “effective connectivity” to refer to this matrix and also make connection with existing terminology.

1.1.2. Interpreting Model Curvature

In the original theory of neural mass models ([5], [6]), the decay-term and transfer function were meant to capture phenomenological components
Fig 1. Interpreting Phenomenological Model Parameters. A) The weight matrix largely captures monotone causal relationships. However, the sign of the causal relationship depends upon the sign of the actual inter-regional connection and which neurons are involved. Excitatory connections/cells are depicted with green arrows and inhibitory in red dots. B) When the sign of connections between regions is mixed, it is possible for indirect relationships to appear stronger than direct connections. Local network structure could influence transfer function and decay parameters. C) Networks with greater reciprocal inhibition (red lines) have a faster time-scale, hence greater decay than those with reciprocal excitation (green). D) A toy example of a network with near binary output due to reciprocal excitation in the output cells (triangles). E) A toy example with a more graded output rule due to inhomogeneities in the excitation of output cells.

of the individual population without corresponding to a singular biological feature. For instance, the transfer function of neural mass models is usually derived from the probability of neuronal spiking as a function of excitation. If cells within each population are assumed homogeneous, the population level activity is proportional to the individual spiking probability when refractory periods are negligible. Under this homogeneity assumption, inter-parcel variation in the transfer function slope would directly reflect variation in the cellular spiking probability between parcels. For cortical neurons with low-firing rates at rest, the spiking threshold is essentially constant (unlike bursting cells for instance), so a high slope might be interpreted as low noise. Since the ground-truth relation with excitation is binary for each cell at a given time (“all or none” spiking), all deviations from that relation must be due to variations in how much of the population-level excitation each neuron receives.

However, if we instead allow parcels to be internally heterogeneous, the transfer function slope parameter may indicate heterogeneities in either the spiking threshold or how excitation is distributed within the parcel (Fig. 1 D, E). For a simple leaky integrate-and-fire model of neurons the individual transfer functions are binary (infinite slope). However, as the variation in
firing thresholds between cells increases, the cumulative probability of population spike count becomes more graded corresponding to the sum of binary functions with different thresholds. Other sources of variation such as noise or inhomogeneities in projections to cells within the population would have a similar effect (Fig. 1E). Thus, although the exact source of variation (i.e., between regions or individuals) in the transfer function slope is unknown, a likely contributor is the degree of within-parcel variation, which may be due to inhomogeneities in internal/external inputs or neuronal dynamics.

However, there are at least three other potential physiological influences in transfer function slope. The first is the relationship between neuronal activity and the BOLD response. The neural components of the BOLD signal are more closely related to synaptic activity than neuronal spiking, so the likelihood of synaptic activity achieving a spike may also be a factor. For instance, for a given number of excitatory synaptic events, the likelihood of the post-synaptic cell firing generally increases with the synchrony of these events. Thus, the degree of synchronization could be another factor in the transfer function slope with parcels having greater synchronization of excitatory inputs having a higher slope. Alternatively, variation in neurovascular coupling between regions may affect the relationship. Regions with less predictable or less uniform hemodynamics would likely receive a lower transfer function slope similar to the case of neuronal variation. In this case, however, the lower slope results from uncertainty in observations rather than variation ("noise") within the generative system.

A final factor may be the intrinsic dynamics of each population. As the BOLD-based observations are temporally coarse (i.e., low resolution), the activity level of each population is more reflective of the average level of synaptic activity over hundreds of milliseconds. Thus, the transfer function seeks to relate the sum of parcel output over hundreds of milliseconds to the sum of parcel input (internal and external) over hundreds of milliseconds. Populations with more temporal integration (better “memory”) are less sensitive to variation in input timing so transfer function slope might also increase with parcel memory. However, results actually indicated the opposite: parcels with greater slopes consistently had parameters reflecting less temporal integration (larger decay; see Results). Temporal integration within our model is reflected by the decay parameter, with high decay indicating less temporal integration.

1.1.3. Interpreting Model Decay

For neural mass models, the decay term describes how quickly a homogeneous population returns to its baseline level of activity. It is assumed that, in the absence of external inputs, the time course will be exponential, leading to the linear term $-Dx$. Many cellular models also contain
a linear decay term corresponding to the leak current, with $D$ equal to the membrane time constant. At the population level, however, the decay term cannot be easily related to any biophysically comparable parameter, e.g., leak potassium conductance. Instead, the decay parameter should be considered as a phenomenological fit to the general pattern of homogeneous populations returning to some baseline rest level. In the current model, however, we relax the assumption of linear decay by also allowing “self-connections” in the connectivity matrix. That is not to say that the individual population members (neurons) contain autoconnections, but that by allowing both a nonlinear term and a linear term we allow a greater range of possible intrinsic dynamics including self-excitation at the population level (Fig. 1 C). When the model is fit to the HCP data, all individuals were found to display nonnegative values for the nonlinear self-interaction. The resultant intrinsic dynamics for each isolated parcel consist of a nonlinear self-excitation and a linear self-inhibition which can lead to either a single stable equilibrium (near the mean BOLD signal) or bistability wherein initial conditions sufficiently above the mean will all converge to one equilibrium and those sufficiently below the mean converge to another. The bistable case generally results when the maximal slope of the self-excitatory component is larger than the decay term (see Supp. for precise conditions). In general, we expect that the decay parameter is related to the relative proportions of local excitation/inhibition within each parcel (Fig. 1 C). The anatomical distribution of decay terms across parcels was largely consistent across subjects (Fig. 4B).
1.2. Derivation of the MINDy Transfer Function

The use of a new transfer function is motivated by the desire to unify the three main classes of transfer function employed in both artificial neural networks and biological neural models: the rectified linear unit (ReLU), softplus, and the logistic sigmoidal function. These functions differ in their curvature (ReLU is piecewise linear, while the others are smooth) and their boundedness (ReLU and softplus are unbounded, linear in the positive limit). Rather than specifying one of these functions explicitly, as is usually done, we chose to create a more general functional form and let the data select the function’s shape on a person x region basis. This form consists of a generalized class of sigmoidal functions which can be varied from smooth to piecewise-linear. We’ll later show that this property enables approximation of the other two classes (ReLU/softplus) over bounded domains. Our function is generated by integrating the difference of two shifted sigmoidal functions (denoted $\sigma(y)$).

$$\hat{\psi}_\alpha(x) := \int_{-\infty}^{x} \sigma(\alpha(y + .5)) - \sigma(\alpha(y - .5)) dy$$ (1)

By Proposition 1 (below), the shift by $\pm .5$ guarantees that $\hat{\psi}$ will have the same limits as the original function $\sigma(x)$ and also retain any of the original function’s reflection symmetries about $x = 0$. Moreover, $\hat{\psi}$ reduces to the definite integral (Proposition 1):

$$\hat{\psi}_\alpha(x) = \int_{x-5}^{x+.5} \sigma(\alpha y) dy$$ (2)

When $\alpha$ is small, this formulation generates smooth sigmoidal functions with curvature dependent upon the choice of $\sigma$. However, in the limiting case of large $\alpha$, the function approaches a shifted ReLu function for sigmoids ($\sigma$) that have a lower-limit of zero and appropriate rescaling ($b$):

$$x \leq b \implies \lim_{\alpha \to \infty} b\hat{\psi}_\alpha(b^{-1}x - .5) = ReLu(x) := max(0, x).$$ (3)

For finite values of $\alpha, b$ the function is smoothed and can be rescaled/shifted to behave like a soft-plus function over desired intervals. An analogous means of generating functions was previously used for modeling dendritic saturation ([10]) starting from the logistic sigmoid function:

$$\sigma(x) := \frac{1}{1 + e^{-x}} \implies \hat{\psi}_\alpha(x) = ln\left(\frac{1 + e^{\alpha x+.5}}{1 + e^{\alpha x-.5}}\right)$$ (4)

We chose to use the sigmoidal function:

$$\sigma(x) := \frac{x}{\sqrt{1 + x^2}} \implies \psi_\alpha = \sqrt{\alpha^{-2} + (x + .5)^2} - \sqrt{\alpha^{-2} + (x - .5)^2}$$ (5)
which takes values on \((-1, 1)\) similar to the hyperbolic-tangent (tanh). We favored the chosen sigmoidal basis over tanh/logistic (as was used in \cite{10}), because the resultant transfer function \(\psi\) involved slightly faster and more stable operations (i.e. avoided using log). Additional terms to further customize the slope/intercept of the transfer function were initially considered, e.g.:

\[
\hat{f}(x) = W\psi_\alpha(b \circ (x + s)) - Dx + c
\]  

(6)

However, we observed in early tests that the scaling term could be reduced to a scalar constant \((b = 20/3)\) and fitted values for \(s\) were effectively zero (for z-scored data). We also observed that when transfer functions were bounded over \([-1, 1]\) the \(c\) term became effectively zero which was not the case when we tested transfer functions bounded over \([0, 1]\). Thus, we chose to use functions bounded over \([-1, 1]\) so that the \(s\) and \(c\) terms could be removed. Of course, MINDy models can always be rewritten in an equivalent form featuring a non-negative transfer function and constant drive

\[
c_i := -\sum_j W_{i,j},
\]  

(7)

since \(\psi(x) + 1\) is non-negative and

\[
f(x) = W\psi(x) - Dx = W(\psi(x) + 1) - Dx + c.
\]  

(8)

Proposition 1. Define the operator \(\Phi : C^0 \rightarrow C^1:\)

\[
[\Phi \circ \sigma](x) := \int_{-\infty}^{x} \sigma(y + .5) - \sigma(y - .5)dy.
\]  

(9)

Suppose that \(\sigma\) is non-decreasing and bounded. Then \(\lim \inf[\sigma] = \lim \inf[\Phi \circ \sigma]\) and \(\lim \sup[\sigma] = \lim \sup[\Phi \circ \sigma]\). Moreover,

\[
[\Phi \circ \sigma](x) = \int_{x - .5}^{x + .5} \sigma(y)dy.
\]  

(10)

Proof. By Lebesgue’s Theorem for Monotone Functions, \(\sigma\) is differentiable almost everywhere and we can write a function \(D\sigma\) equal to the derivative of \(\sigma\) at differentiable points and zero otherwise which satisfies:

\[
\sigma(x + .5) - \sigma(x - .5) = \int_{x - .5}^{x + .5} D\sigma(y)dy.
\]  

(11)

By rearranging limits of integration we produce:

\[
[\Phi \circ \sigma](x) = \int_{x - .5}^{x + .5} \int_{-\infty}^{z} D\sigma(y)dydz = \int_{x - .5}^{x + .5} \sigma(y)dy.
\]  

(12)
By monotonicity we have:

\[ \sigma(x - .5) \leq \int_{x-.5}^{x+.5} \sigma(y) dy \leq \sigma(x + .5). \]  

(13)

By the monotone convergence theorem, \( \sigma \) converges to its infimum/supremum. Taking the negative limits for \( x \):

\[ \lim \inf [\sigma] \leq \lim \inf [\Phi \circ \sigma] \leq \lim \inf [\sigma], \]  

(14)

and similarly for the positive limit (lim sup). Applying the squeeze theorem completes the proof.

1.3. Accelerated Stochastic Gradients through NADAM

To fit the models, we use a variant of the stochastic-gradient descent (SGD) method: NADAM (Nesterov-accelerated adaptive moment estimation [9]) which builds upon the earlier ADAM algorithm ([11]). Gradient descent methods are algorithms that attempt to minimize a cost function, by updating parameters based upon the cost function’s current slope (gradient). For an error function \( E \) and a parameter \( \theta \), the original gradient descent algorithm updates the estimate of the parameter (denoted \( \theta_k \)) at each iteration \( (k) \) of the algorithm according to:

\[ \theta_{k+1} = \theta_k - \eta \frac{\partial E}{\partial \theta_k} \]  

(15)

In which \( \eta \) is the user-chosen learning rate parameter. Although highly efficient, gradient descent algorithms are not guaranteed to reach a global minimum for non-convex problems; further, the original gradient-descent method is prone to getting “trapped” in local minima. Additionally, global-minima of highly non-convex problems may not be desirable as they sometimes poorly generalize [3]. Since the development of first-generation gradient-descent algorithms, substantial progress has been made in generalizing the method to handle non-convex surfaces, often by adding a “momentum” term. Momentum in SGD makes the system’s evolution a function of not only the current gradient, but also past gradients. Like physical momentum, this memory allows the algorithm to “roll past” small dips in the error surface. The NADAM algorithm is one of the most recent advances in momentum-based SGD ([9]). Rather than just updating the parameter estimate \( (\theta_k) \) at each time step, NADAM also updates a moving average of the gradient \( (m_k) \) and the squared gradient \( (n_k) \). The moving average of the gradient adds momentum, while the moving average of the squared gradient is used to adaptively scale updates according to the mean square error. The memory of the moving average gradients and squared gradients are controlled by the hyperparameters \( \mu \) and \( \nu \), respectively. A
Fig 3. Schematic of NADAM benefits with illustrative error surfaces (y-axis) for fitting a parameter (x-axis values) on the first scanning session (left) and the second (right). The NADAM algorithm uses momentum to avoid shallow local minima (green). This feature also prevents convergence to overly sharp minima (even if they are global) because such error surfaces can often correspond to overfitting (blue) and hence do not generalize across sessions. Rather, NADAM emphasizes solutions to deep basins (purple) which may prove the most robust.

“regularization” hyperparameter ($\varepsilon$) stabilizes the learning rate and prevents division by zero. The NADAM algorithm thus updates as follows:

\[
m_{k+1} = \mu m_k + (1 - \mu) \frac{\partial E}{\partial \theta_k} \\
n_{k+1} = \nu n_k + (1 - \nu) \frac{\partial E^2}{\partial \theta_k} \\
\theta_{k+1} = \theta_k - \eta \frac{\frac{1 - \mu}{1 - \mu^{k+1}} \frac{\partial E}{\partial \theta_k} + \frac{\mu}{1 - \mu^{k+1}} m_{k+1}}{\sqrt{\frac{n_{k+1}}{1 - \mu^{k+1}}} + \varepsilon}
\]

Like its predecessor, the ADAM algorithm, NADAM makes use of momentum to avoid converging to shallow minima and also incorporates estimates of the error surface curvature. However, like all SGD methods, the NADAM algorithm is still only guaranteed to converge to a local minimum. The advantage, however, is that the NADAM algorithm improves the depth and breadth of that local minimum. Due to the limited amount of data per subject we prioritize robustness over goodness-of-fit so the global-minimum is not necessarily desirable and might actually correspond to over-fitting. There are thus two main advantages to using modified SGD over a global-optimizer: 1) computational efficiency, which enables us to fit very large networks, and 2) emphasis on robust solutions, which improves cross-validation and prevents over-fitting.

1.4. Hyperparameters in Model Fitting

In deconstructing the connectivity matrix, we produce three terms: one $n \times n$ sparse component ($W_S$) and one $n \times m$ rectangular matrix for each of the two diffuse components ($W_1, W_2^T$) in which $n$ denotes the number of
parcels and \( m < n \) denotes the chosen dimensionality of the diffuse matrix. Hence, \( W_{\text{full}} := W_S + W_1 W_2^T \). The sparsity of \( W_S \) is achieved with \( L_1 \) regularization with penalty \( \lambda_1 \) and both of the diffuse components are also \( L_1 \) penalized with the same coefficient (\( \lambda_2 \)) for both halves. The full diffuse matrix \( W_1 W_2^T \) also receives \( L_2 \) regularization.

The full integrated cost function which includes the regularization penalty is thus:

\[
J = \frac{1}{2} E_T \left[ \| (X_{T+1} - X_T) - [(W_S + W_L) \psi_A (X_T) - DX_T] \|_2^2 \right] \\
+ \lambda_1 \| W_S \|_1 + \lambda_2 \text{Tr}(|W_S|) + \lambda_3 (\| W_1 \|_1 + \| W_2 \|_1) + \frac{\lambda_4}{2} \| W_L \|_2^2 
\]  

(19)

with the notation \( E_T \) denoting the expected value over all time points within the minibatch. The NADAM algorithm itself involves four parameters: an update rate parameter, two decay parameters for computing moving averages, and one “regularization” parameter (\( \beta \)). Unlike the regularization parameters for the weight matrices, which factor into the error and steer the model towards sparse solutions, the NADAM regularization parameter simply serves to stabilize the speed of updates and prevent division by zero. We chose parameters for each variable: \( W_S, W_L, \alpha, D \). As with the regularization terms, we used the same parameters for the two halves of the diffuse component: \( W_1 \) and \( W_2 \). We found that the least impactful hyperparameters are the NADAM decay rate hyperparameters, which only need to be slightly less than one. The most impactful hyperparameters are the \( L_1 \) regularization penalties for the weight matrices which control the balance between over-fitting and under-fitting.

1.5. Interpretations of the Weight-Decomposition and Well-Posedness of the Problem

From a Bayesian perspective, this penalty function is equivalent to maximum a posteriori (MAP) estimation with fixed Laplace distribution (symmetric exponential) priors for each of the individual weight matrices and a normal prior on the combined low-rank component. The Laplace (distribution) prior is unrelated to Laplace approximation as used in Bayesian estimation. Since we assume that process noise is iid. between parcels, its influence (scaling the prediction error term) gets absorbed in the regularization coefficients (by multiplying all terms of the log-likelihood by the noise variance). From a linear-algebra perspective, the regularization prioritizes matrices which can be minimally perturbed to produce a skewed eigenvalue spectrum with sparse eigenvectors. It is important to note that the sum of sparse and low-rank matrices (e.g. \( W \)) need not be sparse nor low-rank so this decomposition is quite flexible. The values of each \( \lambda_i \) and
The primary function of this decomposition is to prevent over-fitting. In all contexts, the potential for over-fitting is related to the difference of model and data degrees of freedom (parameters vs. measurements). While the scale of the current MINDy model might induce initial skepticism, the current problem is similarly well-posed as several recent attempts at fitting a smaller number of parameters. For instance, a recent approach by Wang and colleagues ([12]) to fit just local parameters (2 per node) in modeling the functional connectivity matrix results in 138 parameters estimated from 2,278 data points collapsed across the whole brain (approximately 16.5 measurements per parameter). Although the current approach estimates far more parameters (n+2 per node) it also utilizes many more data points: the whole multivariate time series is used for estimation rather than just the functional connectivity matrix. This results in a ratio of roughly 12 measurements per parameter using the full HCP resting-state data for subjects with MMP cortical and Freesurfer subcortical parcellation (11 for the gwMRF-400). The HCP dataset also contains double this quantity for a subset of subjects who also participated in a later retest session. Of course, these “back-of-the-envelope” calculations assume the worst-case scenario of no parameter covariation. In reality, we expect the set of underlying effective connectivity matrices to be much more constrained—a fact that we exploit via our weight decomposition. Moreover, the sparse regularization priors result in many weights becoming negligible (i.e. very near zero), so even fewer non-trivial estimates are made (Fig. 5A,B). The fitting process is also tractable due to the use of the NADAM algorithm ([9]) which is optimized for simultaneously fitting very large numbers of parameters in a highly efficient manner (approximately one minute per model on a laptop).

1.6. Reliability and Individual Differences in the Weight Decomposition

In the main text (Sec. 2.2), we introduced a linear decomposition of the weight matrix into sparse and low rank components:

$$ W := W_S + W_L = W_S + W_1 W_2^T. $$ (20)

This decomposition was motivated by the dual influences of sparse, long-distance connections between “hub” regions (which motivates $W_S$) and the propagation of these signals along subnetworks (which motivates $W_L$). This formulation was developed as a fitting heuristic by which we could approach the high-dimensional model-estimation inherent in MINDy (Eq. 4). All previous analyses have focused upon the final weight matrix $W$, since the dynamical systems models in MINDy do not require explicit consideration of the components ($W_S$ and $W_L$). In this section, we present
preliminary analyses which suggests reliable individual differences in this decomposition. We do not separately analyze the two rectangular matrices $W_1$ and $W_2$ which define $W_L$ since they are not unique (e.g. their column indices are arbitrary). As with the full weight-matrix, we measured the degree to which estimates were similar within-subject (different sessions) vs. between-subject. We only consider non-diagonal elements of the matrices since the recurrent elements (diagonals) are distinguished by a separate regularization term in the cost function (Eq. 4). The results for recurrent connections in isolation were: within-subject: $r = .56 \pm .06$, between-subject: $r = .42 \pm .07$. We found that both components had greater similarity within-subject ($W_S: r = .53 \pm .06$ and $W_L: r = .67 \pm .04$) than between-subjects ($W_S: r = .35 \pm .03$ and $W_L: r = .39 \pm .04$). Thus, the component matrices had significant reliability and exhibited individual differences although the reliabilities were lower than that of the full weight matrix (see Sec. 3.3).

We also considered the degree to which sparsity vs. low-dimensionality of a subject’s weight matrix was a reliable trait. We quantified this value by comparing the (log) relative magnitude of the sparse and low-rank matrices ($U$):

$$U := \ln\left(\frac{\|W_S\|_F^2}{\|W_L\|_F^2}\right)$$

(21)

With $\|W\|_F^2$ denoting the squared Frobenius-norm (sum of squared matrix elements). Each model produces a single (scalar) value for $U$. We tested whether individual differences in this quantity were reliable either with or without including recurrent connections. We found that results were reliable in either case (with: ICC = .763, without: ICC = .710) and that individual subject’s values were highly correlated for the two cases ($r(51) = .919$ for the mean across sessions). When recurrent connections are included, the sparse component is larger (i.e. $U > 0$) on average: $U = .265 \pm .261$ while the low-rank component otherwise dominates: $U = -.528 \pm .215$. Thus, results indicate that the relative magnitude of sparse vs. low-rank components is a reliable marker of individual differences in MINDy models. However, the degree to which the overall weight matrix is dominated by either component will depend upon whether the recurrent (sparse) connections are considered. Moreover, we expect that the average value of this ratio will depend upon the particular choice of regularization parameters which will favor either component. Therefore, while results are promising in terms of individual differences, we do not recommend using the weight decomposition to quantify the general sparseness/dimensionality of brain networks without considering the influence of regularization hyperparameters.
Fig 4. Local MINDy parameters (curvature and decay) exhibit consist anatomical structure within and between networks. A) Distribution of curvature parameter values for each brain parcel grouped according to network (17-network [13]). B) Curvature parameters reordered according to mean demonstrate that within-network variability is also consistent. C) Anatomical profile of group-mean curvature for the MMP atlas ([14]). D) Profile for the gwMRF ([13]) parcellation. E-H) same as A-D but for the decay parameter. I) Hierarchical Heterogeneity map by Demirtas and colleagues ([15]) using group T1/T2 ratio. J,K) same as G,H but for the first principal component of decay across subjects.

1.7. The Influence of Hyperparameter Choices on Sparsity

The relative value of the hyperparameters $\lambda_1$ vs. $\lambda_2$ can influence the sparsity of the MINDy connection matrix. Overly small values of $\lambda_1$ will not generate sparsity in the “sparse” matrix ($W_S$) while overly large values of $\lambda_1$ relative $\lambda_2$ will also decrease sparsity by biasing solution coefficients toward the low-rank component ($W_L$). However, in practice, even the low-rank matrix is substantially sparser than the rsFC which generates overdispersion (“heavy-tailedness”). We quantify this property via kurtosis:

$$Kurt[X] := E \left( \frac{X - E[X]}{\sigma[X]} \right)^4$$

which, for a normal distribution, is 3. The kurtosis for each subject’s low-rank component (35.6 ± 8.1) substantially exceeds that of the rsFC matrix (7.2 ± 0.9) although both are dwarfed by the sparse component (188.7 ± 27.8). Thus, both components of the weight matrix are more sparse than rsFC so hyperparameter choices which bias towards either term will still result in sparser solutions than rsFC. Lastly we considered the case in which all regularization terms are equal zero which forms a lower-bound case for sparsity (i.e. all other regularization values should produce more sparse estimates). Individual model estimates without regularization are extremely noisy (hence the need for regularization) and less sparse ($Kurt = 5.9 ± 0.7$) than rsFC. However, the group-mean of these noisy
estimates ($Kurt = 24.4$) is also more sparse than the group-mean for rsFC ($Kurt = 9.0$). Moreover, the group-mean without regularization was highly correlated ($r(175559) = .917$) with the group-mean for the full MINDy model. We conclude that individual weight estimates are sparse for a range of hyperparameter values and the group mean of estimated weights for our dataset is more sparse than rsFC, irrespective of hyperparameter choices.

1.8. Comparing MINDy and Spectral Dynamic Causal Modeling

We view the primary contributions of MINDy in its scalability, biological interpretability, and the ability to predict nonstationary resting-state dynamics. However, one non-unique benefit of MINDy is the data-driven characterization of effective connectivity via the weight parameter. Other methods, such as stochastic DCM and spectral DCM have also used (linear) dynamical systems models to estimate effective connectivity. By converting problems into the frequency domain, spectral DCM (spDCM; [16]) has been applied to brain models consisting of 36 regions, but still has significantly higher computational cost than MINDy. One question, therefore, is whether the scalability of MINDy comes at the cost of accuracy. We tested this question in a series of ground-truth simulations. To be clear, we are not seeking to demonstrate that MINDy is necessarily a better estimator of low-dimensional effective connectivity, but rather that the scalability of MINDy does not significantly impair accuracy (i.e. MINDy is at least as good as DCM). One inherent advantage of spectral DCM is the ability to estimate region-specific hemodynamic kernels which is not part of the currently proposed MINDy model (although extensions for HRF estimation are being developed [17]). Thus, we consider two features when comparing MINDy and spectral DCM (spDCM): scalability and robustness to spatial variation in the HRF.

1.8.1. Benchmarking with Unbiased Ground-truths

The main difficulty in comparing MINDy and DCM is the different underlying assumptions—spDCM, for instance, has only been validated using a linear ground-truth ([16]). We took a number of steps to prevent bias based upon differing assumptions (SI Tab.1) and when bias was inevitable, we made choices that favored spDCM. First, we used two ground-truths which were not based upon either technique: a neural mass + balloon-Windkessel ground-truth and a continuous asymmetric Hopfield-model ground truth with either a random global HRF or spatially variable hemodynamics (the same models as for model-mismatch analyses in Sec. 3.7.5). Random HRFs were generated by sampling $\alpha_1 \sim \mathcal{N}(6, \sigma^2)$ and $\beta_1 \sim \mathcal{N}(1, (\sigma/6)^2)$ in Eq. 5. For the global HRF simulations, $\sigma = .5$ and the same values of $\alpha_1$ and $\beta_1$ were used for each node (for a given simulation). In the spatially-variable HRF simulations, values were independently drawn for each node. We implemented spectral-DCM using the MATLAB code provided with SPM-12.
Fig 5. Comparison of accuracy and run-time for MINDy and spDCM. A) Accuracy in estimating ground-truth connectivity from Hopfield-network simulations by network size. “Reduced MINDy” indicates that all regularization terms were removed from MINDy to avoid bias (analogous results for the full MINDy model are in Sec. 3.7.5). Lines indicate mean and bars indicate first/third quartiles. B) Accuracy in estimating ground-truth connectivity from Neural Mass simulations by network size. Results for the full MINDy model are in Sec. 3.7.5. C) Model performance as a function of HRF spatial variability for two network sizes: 6 and 8 nodes. Note that MINDy performance decreases with HRF spatial variability, whereas the effect for spDCM is minor. D) Full run times for MINDy and spDCM for each simulation as a function of simulation type/size. E) Run time per EM iteration (spDCM) and for 10,000 mini-batches in MINDy. We chose to compare with 10,000 mini-batches so that the run-times would be comparable for the smallest network size ($n_{pop}=6$). F) Model performance as a function of HRF spatial variability for two network sizes: 6 and 8 nodes. Note that MINDy performance decreases with HRF spatial variability, whereas the effect for spDCM is minor.
(function name: “spm_dcm_fmri_csd”) to compute the expected value for each connection weight based upon cross-spectral density (spectral DCM). Simulations and model-fitting were performed single-core on Intel Xeon E5-2630v3 CPUs. Since linear models like DCM do not separate recurrent connections and decay, we only compared accuracy for non-recurrent (off-diagonal) elements of the connectivity matrix for each technique. Both ground-truth simulations were integrated at time-scales faster than the sampling rate (dt=.025 vs. 725ms TR for the neural mass and dt=100ms vs. 700ms TR for Hopfield to mirror HCP acquisitions). This feature ensures that results are not biased against spDCM due to simulation time-scale, since MINDy discretizes the model in terms of TR, while spDCM maintains a continuous-time estimation framework.

We used the same hyper-distributions to parameterize neural-mass simulations and Hopfield networks as in Section 3.7.5, but for smaller network sizes. We made three further adjustments to reduce bias: first, we set all regularization terms from MINDy equal to zero, so there would be no inherent advantage to MINDy based upon the hyper-distributions of network structure. The regularization-based calculations were still performed (preserved run time), but they had no effect on the solution (were always zero). We also increased the simulation length of the neural-mass simulation from 3000 to 5000 TRs and removed the nearest-neighbor smoothing from the Hopfield Network simulations which had been included to mirror the empirical processing-pipeline. These adjustments did not affect MINDy but were found to increase the performance of spDCM.

In addition to MINDy and spDCM, we tested the accuracy of functional connectivity and one-step prediction by multiple-regression (solving $\Delta x = Mx_t$). The latter case provides an additional control by providing an alternative method to parameterize linear dynamical systems. This control is important because it can indicate that cases in which spDCM underperforms are due to the estimation technique rather than linearity per se. (i.e. cases in which multiple-regression is accurate but spDCM is not). When all regularization terms are set equal to zero and the transfer function is linear, the MINDy and multiple-regression models are equivalent. The regression approach differs from spDCM, however, in the strength of the linearity assumption. Whereas spDCM seeks a linear model that best explains statistics drawn from the full time-course (i.e. assumes global linear dynamics), the regression approach (like MINDy) considers the changes at each TR (i.e. the collection of local dynamics).

1.8.2. MINDy Performs Competitively with DCM

Results indicated that MINDy scaled-well in terms of performance and run-time. Moreover, this scalability did not generally come at a cost to per-
formance relative contemporary higher-complexity techniques (i.e. spectral DCM). In all situations tested (model x size x HRF), the reduced MINDy model (no regularization) performed at least as well as all competitors in retrieving ground-truth connection weights (SI Fig. 5 A-C). All methods performed poorer for the neural-mass ground-truth (SI Fig. 5B; SI Tab. 2) than for the Hopfield network (SI Fig. 5A; SI Tab. 3 SI Tab. 4) which was expected due to the greater difficulty of the problem (much faster timescales and more complicated models). We note that the full MINDy model performs substantially better in neural mass simulations than the reduced version (see Sec. 3.7.5). However, even when regularization was removed to prevent bias due to assumptions on network structure (“reduced” MINDy), performance remained competitive with spDCM. We also observed that spDCM performance decreased with network size for the Hopfield simulation (SI Fig. 5B; SI Tab. 3), but the regression-based model did not. Thus, this feature cannot be explained solely in terms of functional form (linearity) although the reduced MINDy did consistently outperform regression.

Moreover, the computational complexity of MINDy is substantially lower than that of spDCM. The theoretical limiting computational complexity of MINDy is a second-degree polynomial in the number of nodes since the highest-complexity operation in terms of \( n_{\text{pop}} \) is analytically calculating the error-gradient with respect to the weight-matrix \( (O(n_{\text{pop}}^2)) \) although the empirical complexity was substantially lower for these simulations (the quadratic term only dominates for much larger \( n_{\text{pop}} \); SI Fig. 5D). By contrast, the spectral-DCM code packaged with SPM has at least fourth-order complexity in terms of the population size for a fully-connected model (SI Fig. 5D). Each Expectation-Maximization (EM) iteration is dominated by \( O(n_{\text{pop}}^3) \) (SI Fig. 5E), but the total complexity can be even greater if the number of iterations until convergence also increases with \( n_{\text{pop}} \) (SI Fig. 5F). For instance, the median total runtime for the neural mass and Hopfield network simulations scaled with \( O(n^{5.66\pm.44}) \) and \( O(n^{5.85\pm.38}) \), respectively (95% confidence estimated using the “fit” function in MATLAB 2020a for linear power functions: “power1”). The empirical complexity of each EM iteration was roughly the theoretical limit of 4: \( O(n^{3.84\pm.20}) \) and \( O(n^{3.97\pm.16}) \) for the neural mass and Hopfield simulations, respectively.

However, even the minimal case of fourth-order complexity can severely limit scalability. For context, a recent “large-scale” spectral DCM paper ([16]) employed 36 brain ROIs with a run-time of between 1,280 and 2,560 minutes per model (between 64 and 128 iterations at 20 minutes each). Increasing resolution from 36 ROIs to the 419-node parcellation we employed (19 subcortical + 400 cortical [13]) would increase CPU time by a factor of over 18,000 (roughly 44 to 89 years per model for the same data and
Table 1. Types of assumptions (“Difference”) made by MINDy vs. spDCM and our controls to mitigate these differences in simulated comparisons (SI Sec. 1.8).

| Difference          | MINDy          | spDCM          | Adjustment/Control                      |
|---------------------|----------------|----------------|----------------------------------------|
| Model Form          | Nonlinear      | Linear         | Model Mismatch; Time-Invariant          |
|                     | Network        | Time-Invariant | Regression Control                      |
| Regularization      | Yes            | No             | Set MINDy regularization equal 0        |
| Time-Scale          | Discrete (TR)  | Continuous     | Simulation dt<<TR                       |
| Local Components    | Recurrent +    | Decay          | Only tested non-recurrent connections   |
|                     | Decay          |                |                                        |
| HRF                 | Fixed          | Local Estimates| Parametrically varied spatial HRF       |

hardware as [16]). By contrast, the current MINDy models for HCP data were locally fit on a laptop in less than one minute each (Intel i7-8750H CPU, 2.2GHz, 6 cores).

One area in which spectral DCM proved advantageous, however, was in robustness to spatial variability in the hemodynamic response (SI Fig. 5.C; SI Tab. 4). As with the analogous simulations in Section 3.7.5, MINDy performance decreased with the underlying HRF’s spatial variability (see also [17]). This pattern is expected since the currently proposed MINDy assumes a fixed HRF which these simulations violate. The performance benefit of MINDy over spDCM, likewise decreased with HRF spatial variability. For the smallest network (6 nodes) and highest level of HRF variability considered, the difference between models’ accuracy was negligible (although statistically significant; SI Tab. 4). Thus, there may be cases of extreme HRF spatial variability in which spectral DCM outperforms MINDy for sufficiently small networks. Although MINDy scales significantly better than competing approaches like spectral DCM, the current version is less robust to spatial variability in the HRF (although see [17] for upcoming extensions). Nonetheless, for all simulations considered, MINDy performed competitively with spDCM within the latter’s scope. We conclude that MINDy performs at least as well as spDCM, while scaling far better.

1.9. Anatomical Distribution of Individual Differences

In concert with the previous analyses of individual differences (Sec. 3.5) across parameter-types (weights, curvature, decay) we also investigated the anatomical distribution of individual differences within each parameter-type. These analyses are post-hoc (exploratory) so we report results as a potential launching pad for future investigations and as a means to understand how MINDy models encode individual differences. We do not perform hypothesis-testing and we caution against interpreting these analyses as
Table 2. Accuracy ($r$) for each method (Regr.=regression) for the neural-mass simulation (SI Sec. 1.8) with variable population sizes (left side). Test-statistics (paired t-test; 2-tailed) for MINDy-spDCM are provided on the right side. The number of nodes per simulation is listed under “nodes” whereas the number of simulation instances is listed under “N”. The paired difference in accuracy between MINDy and spDCM is denoted $\Delta r$. We denote $*=p<.01$ and $**=p<.001$.

| Nodes | rsFC  | Regr.  | MINDy | spDCM | N   | $\Delta r$ | t    |
|-------|-------|--------|-------|-------|-----|-----------|-----|
| 6     | .46(.14) | .34(.20) | .52(.15) | .28(.24) | 240 | .25(.26) | 14.8** |
| 8     | .44(.10) | .35(.16) | .52(.11) | .27(.20) | 320 | .24(.21) | 20.4** |
| 10    | .45(.08) | .35(.12) | .52(.09) | .38(.10) | 130 | .15(.10) | 17.3** |
| 12    | .43(.07) | .35(.11) | .50(.09) | .36(.09) | 80  | .15(.08) | 16.6** |
| 14    | .45(.06) | .35(.09) | .53(.07) | .38(.07) | 40  | .15(.06) | 14.5** |
| 16    | .43(.05) | .32(.09) | .50(.08) | .36(.06) | 20  | .14(.03) | 22.8** |

Table 3. Accuracy ($r$) for each method (Regr.=regression) for the Hopfield-network simulation (SI Sec. 1.8) with variable population sizes (left side) and a random global HRF ($\sigma=.5$). For this simulation we added a new population size (9 Nodes) post-hoc to see whether the unexpected decrease in spDCM accuracy between 8 and 10 nodes was continuous (it was). Since this special case contained an odd number of nodes the hyperdistribution parameter $q$ in Sec. 2.6.2 was always equal to one (instead of one and two with equal probability). Test-statistics (paired t-test; 2-tailed) for MINDy-spDCM are provided on the right side. The number of nodes per simulation is listed under “nodes” whereas the number of simulation instances is listed under “N”. The paired difference in accuracy between MINDy and spDCM is denoted $\Delta r$. We denote $*=p<.01$ and $**=p<.001$.

| Nodes | rsFC  | Regr.  | MINDy | spDCM | N   | $\Delta r$ | t    |
|-------|-------|--------|-------|-------|-----|-----------|-----|
| 6     | .49(.11) | .83(.08) | .86(.07) | .76(.20) | 200 | .10(.19) | 7.0** |
| 8     | .46(.08) | .83(.06) | .86(.05) | .65(.23) | 200 | .22(.22) | 14.3** |
| 9     | .44(.06) | .83(.06) | .85(.05) | .38(.19) | 105 | .48(.19) | 25.5** |
| 10    | .43(.06) | .84(.05) | .87(.04) | .26(.12) | 80  | .60(.12) | 45.1** |
| 12    | .40(.06) | .83(.05) | .86(.04) | .24(.10) | 50  | .62(.11) | 39.7** |
| 14    | .39(.05) | .82(.04) | .86(.03) | .20(.07) | 38  | .67(.08) | 48.9** |
| 16    | .37(.04) | .83(.03) | .86(.03) | .17(.08) | 13  | .69(.08) | 30.0** |

Table 4. Accuracy ($r$) for each method (Regr.=regression) for the Hopfield-network simulation (SI Sec. 1.8) with spatially variable HRF and 6 nodes (left side). Test-statistics (paired t-test; 2-tailed) for MINDy-spDCM are provided on the right side. The standard-deviation of the HRF parameters is listed under “$\sigma$” whereas the number of simulation instances is listed under “N”. The paired difference in accuracy between MINDy and spDCM is denoted $\Delta r$. We denote $*=p<.01$ and $**=p<.001$.

| $\sigma$-HRF | rsFC  | Regr.  | MINDy | spDCM | N   | $\Delta r$ | t    |
|---------------|-------|--------|-------|-------|-----|-----------|-----|
| .1            | .48(.10) | .87(.06) | .89(.04) | .77(.18) | 280 | .12(.17) | 11.8** |
| .2            | .48(.11) | .86(.06) | .89(.05) | .77(.19) | 280 | .12(.18) | 11.1** |
| .5            | .49(.12) | .83(.08) | .86(.07) | .76(.18) | 280 | .10(.18) | 9.3**  |
| .75           | .47(.12) | .80(.10) | .82(.10) | .74(.20) | 280 | .08(.21) | 6.5**  |
| 1.0           | .47(.11) | .72(.16) | .75(.15) | .71(.21) | 280 | .03(.21) | 2.7*   |
**Fig 6.** Anatomical distribution of inter-individual variation in MINDy. A) QCD of the curvature parameter with top-20% threshold. B) Same as A) but for the decay parameter. C) QCD for MINDy connection weights. Weights in which the sign was inconsistent across subjects (<75% agreement) or low reliability (Fisher ICC<.5) were censored (grey). D) Mean weight QCD within each network combinations. If over 95% of parcel-wise connections were censored, the network-level connection was also censored (white).
stand-alone findings due to their exploratory nature and relatively low sample-size. We also note that these analyses are performed upon a biased sample of subjects—those that had no high-motion scans (>1/3 frames censored) so these results may also fail to describe variability in the full HCP subject pool (see [7] for cognitive covariates of motion) and its target population (American young adults). These caveats aside, we considered the degree of individual variation for each parcel/connection. We used the quartile coefficient of dispersion (QCD) to quantify the degree of variability within parcels/connections. Conceptually, the QCD is a robust analogue of the more commonly used Coefficient of Variation and is defined in terms of the first \(Q_1\) and third \(Q_3\) data quartiles as:

\[ QCD := \frac{Q_3 - Q_1}{Q_3 + Q_1} \]  

The reason that we apply QCD instead of Coefficient of Variation is that, while both assume the true population follows a ratio-scale (e.g. are one-sided) the QCD is robust to extreme values which, due to measurement error violate the ratio-scale assumption. We censored connections which were unreliable (Fisher’s ICC ≤ .5) or incompatible with QCD (\(Q_1\) and \(Q_3\) differed in sign). We then transformed the other variables onto admissible distributions (non-negative ratio scales) by shifting the curvature and decay parameters so that they had a minimum value of zero. Interestingly, the anatomical distribution of QCD appeared to differ between the curvature and decay parameters. The curvature had the highest QCD in parcels of inferior frontal gyrus, early visual cortex, and a large posterior section of frontal cortex (SI Fig. 6A). For the decay parameter, QCD was highest in visual regions and the bilateral portion of somatosensory cortex traditionally associated with hands (SI Fig. 6B). By contrast, connection weights had the lowest QCD for connections within the visual networks (especially peripheral visual; SI Fig. 6C,D). Some of the highest QCD connections involved the temporal-parietal network (inter and intra-network) and connections to the limbic system. Of course, as previously mentioned, these analyses are purely exploratory and should only be interpreted as an example of how MINDy separates the sources of individual differences (weights, curvature, decay) rather than as a basic neuroscientific result.

1.10. Directed Connectivity Identified by MINDy

The simplest way to characterize connection asymmetries is in terms of regions being sinks (input weights greater than output weights in absolute value) vs. sources (output weights greater than input weights in absolute value). For now we focus upon sources and do so separately for positive (SI Fig. 7A) and negative connections (SI Fig. 7B). For positive sources MINDy most strongly identifies inferior frontal gyrus (IFG),
Fig 7. Connection asymmetries identified by MINDy: A) Difference of total input weights minus output weights for positive connections only (normalized units). B) Same as (A) but for negative connections only (using difference in magnitude of input/output weights). C) Difference of output and input weights for positive output-biased connections in the parcel with the greatest positive output-bias. D) Same as C), but for the parcel with greatest negative output bias. E) The parcel with the second-greatest negative output bias is the contralateral analogue to the parcel in (D). Parcel numbers are labeled for the 17-network gwMRF parcellation ([13]).

bilateral parieto-occipital sulcus, and dorsal prefrontal cortex (SI Fig. 7 A). MINDy identifies the strongest excitatory source as a region of left IFG (see main text; SI Fig. 7 C) and identified bilateral IFG as the strongest negative sources with negative outward-biased connections primarily to components of the Default Mode Network (IPL and medial PFC) with a general contralateral bias (SI Fig. 7 D,E). The role of right IFG in inhibition is well-documented within neuroimaging (e.g. [18]) and lesion studies suggest an inhibitory role for left IFG as well ([19]). These results indicate that the asymmetries within MINDy weights are functionally interpretable. However, these initial findings only scratched the surface of possible analyses.

1.11. Nonlinear Dynamics in MINDy

As discussed in the main-text (Sec. 3.6.2), the fact that MINDy is nonlinear does not inherently imply that the model behaves qualitatively different from models relying upon a linear approximation (e.g. DCM). This distinction is critical to understanding how MINDy models resting-state dynamics: as random fluctuations about a single equilibrium (like DCM) or as topologically significant (nontrivial) dynamics. In general, proving the existence of global behavior in MINDy models constitutes a nontrivial endeavor given their high-dimensionality. However, we can easily rule out trivial dynamics (a Lyapunov-stable global attractor) by
Table 5. Ground-truth validation performance of MINDy and rsFC in recovering the weight matrix of a single subject and the arithmetic difference of weight matrices between subjects. We denote significance with \( ** = p < .001 \), 2-tailed for the contrast Weights minus rsFC.

|                     | Weights | FC     | paired-t (df=33) |
|---------------------|---------|--------|------------------|
| Single Subject      | .800 (.025) | .436 (.065) | 45.629**        |
| Indiv. Differences  | .544 (.049) | .293 (.046) | 58.618**        |

examining eigenvalues of the Jacobian at zero. Since all subjects contain at least one positive (real part) eigenvalue (SI Fig. 8), we conclude (by Proposition 2) that no empirically-parameterized MINDy model is globally Lyapunov-stable.

**Proposition 2.** Consider a continuous-time dynamical system evolving according to \( \dot{x} = f(x) \), with \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n = W \psi(x) - Dx \) with \( \psi : \mathbb{R}^n \rightarrow \mathbb{R}^n \) an odd function \((\psi(-x) = -\psi(x))\) and \( n \times n \) matrices \( W, D \). Suppose that the Jacobian at the origin \((F'(0))\) has at least one eigenvalue with positive real part and none with zero real part. Then for any fixed point \( x_s : f(x_s) = 0 \), at least one of the following hold:

1. There exists a set \( U \) of non-zero measure, whose positive limit-set \( \omega^+(U) \) does not contain \( x_s \).
2. \( x_s \) is not Lyapunov stable—i.e. there exists \( \epsilon > 0 \) for which there is no \( \delta > 0 \) satisfying \( \|x(0) - x_s\| < \delta \implies \forall t > 0, \|x(t) - x_s\| < \epsilon \).

**Proof.** We consider two cases depending upon whether \( x_s \neq 0 \). If \( x_s \neq 0 \) is a fixed point then so is its reflection \( -x_s \neq x_s \) since \( f(-x) = -f(x) \). Suppose that \( x_s \) is locally Lyapunov stable (violating implication 2) and thus possesses an attractive basin \( V \) of non-zero measure. Then \( -V \) is an attractive basin of \( -x_s \) hence \( x_s \notin \omega^+(-V) \) which confirms the first implication. Thus, consider the alternative case: \( x_s = 0 \). By the hypothesis, \( x_s \) is a hyperbolic fixed-point so there exists an open neighborhood \( N \) containing \( x_s \) for which the dynamics on \( N \) are topologically conjugate those of the linearization \( \dot{x} \approx F'(0)x \) (Hartman-Grobman Theorem). The linearization \( F'(0) \) is unstable (at least one eigenvalue with positive real part) which implies that any sufficiently small ball about \( x_s = 0 \) will also be an unstable set. This contradicts Lyapunov stability (confirming implication 2).

1.12. MINDy Optimization: Under the Hood

To stabilize MINDy’s fitting procedure we use two changes of variable during the fitting process. Instead of directly fitting the term \( D \) we fit \( D_2 \) satisfying the relation \( D := D_{\text{min}} + D_2^2 \). This change keeps the estimated parameters away from the pathological conditions in which \( D \) is either
Fig 8. Eigenvalue analyses indicate that empirical MINDy models do not possess a global, Lyapunov stable attractor. A) Distribution of eigenvalues (real-part) for local-linearization about the origin. Shading indicates ±SD and black lines give the data’s maximum/minimum. Note the presence of positive eigenvalues which indicate nontrivial dynamics (Proposition 2). B) Scatterplot of these eigenvalues (all subjects) in the complex plane suggests that the greatest “spin” (complex components) occurs along the unstable subspaces (corresponding to eigenvalues of positive real-part).

Fig 9. MINDy strongly predicts individual and group FC. A) Correlation between empirical FC and simulated FC from the same scanning day for either the same subject or a different subject (mean across other subjects). Blue line indicates group mean. Correlations are averaged across scanning sessions. B) Same as A) but for with all data combined across sessions (simulations from each session’s model were combined) C) Group-mean of empirical FC vs. group-mean of simulated FC (both combined across sessions). Notice that while the correlation is high, the magnitude of simulated FC is smaller than empirical. D) Slightly decreasing noise produces a predict group-mean FC that is nearly indistinguishable from that observed empirically. E) Group-average empirical FC combined across sessions. F) Group-average FC of simulated data combined across sessions. The identity line (perfect match) is indicated in red G) Same as F) but with a slightly decreased noise term (85% of original).
**Fig 10.** MINDy differentiates model parameters between individuals and identifies the source of individual differences. A) Changing only the curvature (left) or decay (right) parameters of a simulated subject has more impact on the simulated rsFC (blue) than on MINDy weight estimates refit to the new simulated data (red). B) Weight matrices are individualized: weight matrices derived from different scans of the same subject are universally more similar than weights fit to another subject. C,D) same as B) but for the curvature and decay parameters, respectively. E) The rsFC matrix is more similar for different scans of the same subject than between subjects. F) Individualized models better generalize to new data from the same subject than to a new subject.

**Table 6.** Sensitivity of MINDy weights and rsFC to changes in non-connectivity parameters in a ground-truth simulation. Performance is measured in terms of false positives—the percentage of connections that change (thresholded by \( p < .05 \)) due to a change in the ground truth model’s curvature/decay. Thus, lower values indicate less false positives (less sensitivity) due to non-connectivity variables. We denote significance with \( ** = p < .001 \), 2-tailed for the contrast Weights minus rsFC.

|                        | Weights     | FC          | paired-t (df=339) |
|------------------------|-------------|-------------|-------------------|
| Changing Curvature     | .063 (.002) | .079 (.027) | -10.639**         |
| Changing Decay         | .076 (.006) | .2749 (.120) | -31.753**         |

**Table 7.** Test-retest correlation across scanning sessions for MINDy parameters. Statistics are present in mean correlation (SD) form. Group-level permutation testing (100,000 each) produced \( p's \approx 0 \) for all parameters vs. chance. Accuracy is in correct assignment for subjects based upon maximal similarity between sessions (e.g. how often is the subject most similar to themselves?). Statistical tests are for the contrast weight vs. FC. W/in=within subject, Btwn=between subject, Diff=w/in subject minus between, Acc.=accuracy. We denote significance with \( * = p < .05 \) and \( ** = p < .001 \), 2-tailed and Bonferroni corrected.

|               | FC       | Weight(W) | \( t(\text{mean}) \) | Curv(\( \alpha \)) | Decay(\( D \)) |
|---------------|----------|-----------|----------------------|---------------------|----------------|
| W/in.         | .757 (.052) | .802 (.018) | 7.883**              | .772 (.029)         | .798 (.062)    |
| Btwn.         | .511 (.044) | .635 (.021) | 38.143**             | .637 (.041)         | .600 (.069)    |
| Diff.         | .246 (.038) | .167 (.019) | -23.076**            | .135 (.023)         | .198 (.052)    |
| Acc.          | 100      | 100       | 0                    | 100                 | 94.3           |
**Fig 11.** MINDy reproduces reliable, accurate estimates of dynamic functional connectivity (DFC). We use $\sigma$-DFC to disambiguate the standard-deviation measure for DFC (e.g. [20], [21]) from other uses of standard-deviation. A) Data simulated from test-retest models (models fit to separate sessions) has at least as high reliability on average as the original data for $\sigma$-DFC. B) Same as A) but for the excursion metric of DFC. C) Correlation between observed and simulated excursion across subjects by region-pair (combining across scanning sessions). D) Predicted group average $\sigma$-DFC for the model simulations (left) and recorded data (right) combining across scanning sessions. E) Same as D) but for excursion. F) Correlation between observed and predicted group-average excursion across region-pairs. The $\sigma$-DFC analogues of C and F are reported in the main text (Fig. 6).

**Table 8.** Comparing the test-retest reliability and pre-processing sensitivity of the MINDy connectivity parameter and the resting state functional connectivity. The pipelines correspond to using motion without CompCor or GSR correction, using motion + CompCor or using motion + CompCor + GSR (default). Results are presented in mean(SD) form for the group distribution of individual test-retest correlations or correlations between different levels of preprocessing applied to the same session. Statistical tests consisted of paired t-tests for the mean correlation, and F-tests for testing heterogeneity of variance. Results generally favored the MINDy connectivity matrix over the FC matrix (greater mean reliability and less variation) but the absolute differences, although highly statistically significant, are not profound. We denote significance with $* = p < .05$ and $** = p < .001$, 2-tailed and Bonferroni corrected.

|                | Weights  | FC       | t (mean) | F (var)       |
|----------------|----------|----------|----------|---------------|
| Motion         | .8096 (.0268) | .7909 (.0619) | 3.1507*   | 5.3219**      |
| CompCor        | .8039 (.0185) | .7481 (.0504) | 10.5096** | 7.378**       |
| GSR            | .8021 (.0180) | .7571 (.0523) | 7.8834**  | 8.4200**      |
| M vs. C        | .9323 (.0178) | .8512 (.0582) | 13.6332** | 10.7247**     |
| M vs. G        | .9116 (.0191) | .7910 (.0776) | 14.1496** | 16.4823**     |
| C vs. G        | .9736 (.0037) | .9669 (.0143) | 4.2162**  | 14.6271**     |
After pre-processing, MINDy fits are robust to motion. Fitting performance was measured by the cross-validated goodness of fit (A,E) and the reliability for each parameter (B-D,F-H). Individual differences in motion were quantified by either membership in median-split high vs. low motion groups (first two rows) or as a continuous variable (bottom row). Groups were assigned for each combination of motion measurement (number of TRs censored, median Framewise Displacement, or Median Absolute Deviation (MAD) of DVARS) and motion type: either the total motion of a subject averaged across scanning sessions (A-D, I,K) or the absolute difference in motion artifact between sessions (E-H,J,L). There was no significant relationship with motion as a discrete characteristic (e.g. high vs. low: A-H) or as a continuous characteristic: group level correlations between motion measures and fitting performance in (I,J) and the associated (uncorrected) inverse p-values in (K,L).

MINDy performance in inverting the weight matrix and its asymmetries in cases of model mismatch. Ground truth models were either a tanh rate-model downscaled to the fMRI TR, a rate model with spatially heterogeneous hemodynamic response functions, or a neural mass model using the nonlinear Balloon-Windkessel model of hemodynamics.

| Model                  | Full $W$   | $W - W^T$ | N  |
|------------------------|------------|------------|----|
| Rate-Model             | 0.949 (0.009) | 0.971 (0.007) | 1700 |
| Rate+spatial HRF       | 0.754 (0.026) | 0.770 (0.027) | 1680 |
| Neural Mass + B-W      | 0.642 (0.032) | 0.567 (0.036) | 1480 |
Fig 13. Sensitivity to various forms of noise in data (Same as Fig. A,B, but with 17 levels of noise). A-C) MINDy estimates all parameters of ground-truth models accurately even in the presence of additive measurement noise. D) Examples of how increasing the variability of hemodynamic parameters changes the shape of randomly drawn hemodynamic response functions (HRF). E-G) Hemodynamic variability does not alter the mean performance of MINDy estimates. H) Hemodynamic variability decreases the consistency of MINDy performance (more variable correlations with ground truth).

Fig 14. MINDy estimates are robust to secondary preprocessing choices: Motion-only (Red), Motion+CompCor (Green), Motion+CompCor+GSR (Red). A) Test-retest reliability of rsFC and the MINDy weights (W), curvature (α) and Decay (D) parameters by secondary preprocessing pipeline. B) Correlation between parameter estimates for different preprocessing pipelines applied to the same scanning data (single session). Color of the bar indicates which pipelines are being compared: R+G indicates similarity between Motion-only and CompCor estimates, R+B indicates Motion-only vs. GSR, G+B indicates CompCor vs. GSR (results are displayed in this order left to right within each column).
Table 10. Definition of variables ("Name") used in MINDy gradient calculations and their interpretation ("Meaning"). The term \(n_{\text{Batch}}\) denotes the number of samples (time-points) included in each minibatch (the training data for a given NADAM iteration).

| Name \(x, \varepsilon\) | Equation | Meaning |
|----------------------|----------|---------|
| \(W\) \(x\) \(W_S + W_L\) | Full weights |
| \(P_1(x)\) \(\sqrt{\xi^{-2} + x \circ (x + b^{-1})}\) | Part of \(\psi\) |
| \(P_2(x)\) \(\sqrt{\xi^{-2} + x \circ (x - b^{-1})}\) | Part of \(\psi\) |
| \(\psi(x, \varepsilon)\) \(b(P_1 - P_2)\) | Transfer Function (\(\psi\)) |
| \(D\) \(D_2^2 + D_{\text{min}}\) | Full Decay |
| \(R\) \(dX - W_K \psi(x) + Dx\) | Residual Error |
| \(Q\) \(W_K^T R\) | 2 \(\frac{\partial R}{\partial \psi}\) |
| \(Z\) \(R \psi(x)^T / n_{\text{Batch}}\) | \(E_T \left[ \frac{\partial R}{\partial W_L} \right]\) |
| \(Y\) \(Z - \lambda_4 W_L\) | \(\frac{\partial J}{\partial W_L}\) |

Table 11. Equations used to efficiently calculate MINDy parameter gradients. These equations leverage the additional variables defined in SI Tab. 10. Note that the decay parameter is updated in terms of its square-root \(D_2\) and the curvature parameter is updated in terms of the linearized form \(\xi\).

| Parameter | Negative Error Gradient \(-\frac{\partial J}{\partial \omega}\) |
|-----------|--------------------------------------------------|
| \(W_S\)  | \(Z - \lambda_1 \text{sgn}(W_S) - \lambda_4 \text{diag}(\text{sgn}(W_S))\) |
| \(W_1\)  | \(YW_1^T - \lambda_3 \text{sgn}(W_1)\) |
| \(W_2\)  | \(W_1 Y - \lambda_3 \text{sgn}(W_2)\) |
| \(\xi\)   | \(-\xi^{-3} b E_T[Q \circ (1/P_1 - 1/P_2)]\) |
| \(D_2\)  | \(-2D_2 \circ E_T[R \circ X]\) |

negative or very small which can cause models to explode in the long term. The term \(D_{\text{min}}\) is a constant, positive hyperparameter. This step does not significantly alter computational complexity and we found that it did not alter results for our current initialization setting of \(D\) as our estimates never approached the pathological regions. However, we included this change of variable in the code as a safeguard should it become relevant for future users.

The second change of variable served to linearize the effects of the nonlinear curvature parameter \(\alpha\). Rather than explicitly fitting \(\alpha\) we fit the variable \(\xi := b/(\sqrt{\alpha^2 + 0.25})\) which satisfies \(\xi = \max_x (\psi'_\alpha(x))\). This transformation smooths the relation between the nonlinear parameter (\(\xi\) instead of \(\alpha\)) and its effects on the model’s vector field. The new parameter \(\xi\) is constrained to be smaller than or equal to \(2b\) so that \(\alpha^2 \geq 0\). Efficient gradient calculations were performed by first calculating the variables in Table 10 and then calculating gradients as in Table 11. In all cases MINDy was run for 5000 iterations with batch size 300 (300 time-points used in each iteration).
Table 12. NADAM hyperparameters for each MINDy parameter and the distributions used to initialize each parameter. NADAM hyperparameters consist of the update rate ("Rate"), decay rate of gradients (µ), decay rate of squared-gradients (ν), and regularization term ε.

| Variable | Rate (η × 10^5) | µ   | ν   | ε   | Initialization   |
|----------|-----------------|-----|-----|-----|-----------------|
| Ws       | 2.5             | .9  | .95 | .15 | N(0, .01)       |
| W1,2     | 6.25            | .9  | .95 | .15 | N(0, .01)       |
| α        | 12.5            | .9  | .95 | .2  | .1 + √.25 + |N(0, .25)| + b^2/4 |
| D2       | 1750            | .9  | .95 | 200 | 1.75 + √|N(0, .25)| |

Table 13. Chosen values for other hyperparameters used in MINDy. These (non-NADAM) hyperparameters consist of the four regularization terms (λ_i) in the cost function (Eq. 4), the minimum allowable value for D (D_{min}), and the scaling factor of the transfer-function (b).

| Variable | Value   |
|----------|---------|
| λ_1      | .075    |
| λ_2      | .2      |
| λ_3      | .05     |
| λ_4      | .05     |
| D_{min}  | .1      |
| b        | 20/3    |

Acknowledgments

MS was funded by NSF-DGE-1143954 from the US National Science Foundation. TB acknowledges R37 MH066078 from the US National Institute of Health. SC holds a Career Award at the Scientific Interface from the Burroughs-Wellcome Fund. Portions of this work were supported by AFOSR 15RT0189, NSF ECCS 1509342 and NSF CMMI 1537015, NSF NCS-FO 1835209 and NIMH Administrative Supplement MH066078-15S1 from the US Air Force Office of Scientific Research, US National Science Foundation, and US National Institute of Mental Health, respectively.

References

[1] T. Schreiber, Measuring information transfer, Phys. Rev. Lett. 85 (2000) 461–464. doi:10.1103/PhysRevLett.85.461

[2] A. Schroeter, F. Schlegel, A. Seuwen, J. Grandjean, M. Rudin, Specificity of stimulus-evoked fMRI responses in the mouse: The influence of systemic physiological changes associated with innocuous stimulation under four different anesthetics, NeuroImage 94 (2014) 372 – 384. doi:10.1016/j.neuroimage.2014.01.046

[3] F. Schlegel, A. Schroeter, M. Rudin, The hemodynamic response to somatosensory stimulation in mice depends on the anesthetic used:
Implications on analysis of mouse fmri data, NeuroImage 116 (2015) 40 – 49. doi:10.1016/j.neuroimage.2015.05.013

[4] N. Todd, S. Moeller, E. J. Auerbach, E. Yacoub, G. Flandin, N. Weiskopf, Evaluation of 2d multiband epi imaging for high-resolution, whole-brain, task-based fmri studies at 3t: Sensitivity and slice leakage artifacts, NeuroImage 124 (2016) 32 – 42. doi:10.1016/j.neuroimage.2015.08.056

[5] H. R. Wilson, J. D. Cowan, Excitatory and inhibitory interactions in localized populations of model neurons, Biophysical Journal 12 (1) (1972) 1–24. doi:10.1016/S0006-3495(72)86068-5

[6] G. Deco, V. K. Jirsa, P. A. Robinson, M. Breakspear, K. Friston, The dynamic brain: From spiking neurons to neural masses and cortical fields, PLOS Computational Biology 4 (8) (2008) 1–35. doi:10.1371/journal.pcbi.1000092

[7] J. S. Siegel, A. Mitra, T. O. Laumann, B. A. Seitzman, M. Raichle, M. Corbetta, A. Z. Snyder, Data quality influences observed links between functional connectivity and behavior, Cerebral Cortex 27 (9) (2017) 4492–4502. doi:10.1093/cercor/bhw253

[8] N. Weiner, Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications, MIT Press, New York City, 1949.

[9] T. Dozat, Incorporating nesterov momentum into adam, Proceedings of 4th International Conference on Learning Representations, Workshop Track, 2016 (2016).

[10] M. Singh, D. Zald, A simple transfer function for nonlinear dendritic integration, Frontiers in Computational Neuroscience 9 (2015) 98. doi:10.3389/fncom.2015.00098.

[11] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980 (2014). arXiv:1412.6980 URL http://arxiv.org/abs/1412.6980

[12] P. Wang, R. Kong, X. Kong, R. Liégeois, C. Orban, G. Deco, M. P. van den Heuvel, B. Thomas Yeo, Inversion of a large-scale circuit model reveals a cortical hierarchy in the dynamic resting human brain, Science Advances 5 (1) (2019). doi:10.1126/sciadv.aat7854

[13] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, B. T. Yeo, Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri, Cerebral Cortex (2017) 1–20 doi:10.1093/cercor/bhx179
[14] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, E. Harwell, John and Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, D. C. Van Essen, A multi-modal parcelation of human cerebral cortex, Nature 536 (2016) 171–178. doi:10.1038/nature18933.

[15] M. Demirtaş, J. B. Burt, M. Helmer, J. L. Ji, B. D. Adkinson, M. F. Glasser, D. C. Van Essen, S. N. Sotiropoulos, A. Anticevic, J. D. Murray, Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics, Neuron 101 (6) (2019) 1181–1194. doi:10.1016/j.neuron.2019.01.017.

[16] A. Razi, M. L. Seghier, Y. Zhou, P. McColgan, Large-scale dcm for resting-state fmri, Network Neuroscience 1 (3) (2017). doi:10.1162/NETN_a_00015.

[17] M. F. Singh, A. Wang, T. S. Braver, S. Ching, Scalable surrogate deconvolution for identification of partially-observable systems and brain modeling, bioRxiv (2020). doi:10.1101/2020.03.20.000661. URL https://www.biorxiv.org/content/early/2020/03/23/2020.03.20.000661

[18] A. R. Aron, T. W. Robbins, R. A. Poldrack, Inhibition and the right inferior frontal cortex, Trends in Cognitive Sciences 8 (4) (2004) 170 – 177. doi:10.1016/j.tics.2004.02.010.

[19] D. Swick, V. Ashley, A. U. Turken, Left inferior frontal gyrus is critical for response inhibition, BMC Neuroscience 9 (1) (2008) 102. doi:10.1186/1471-2202-9-102.

[20] A. S. Choe, M. B. Nebel, A. D. Barber, J. R. Cohen, Y. Xu, J. J. Pekar, B. Caffo, M. A. Lindquist, Comparing test-retest reliability of dynamic functional connectivity methods, NeuroImage 158 (2017) 155 – 175. doi:10.1016/j.neuroimage.2017.07.005.

[21] C. Zhang, S. A. Baum, V. R. Adduru, B. B. Biswal, A. M. Michael, Test-retest reliability of dynamic functional connectivity in resting state fmri, NeuroImage 183 (2018) 907 – 918. doi:10.1016/j.neuroimage.2018.08.021.