

Optimal Transport Model Inversion for Biophysically-Grounded Virtual Brain Simulations

Marvin Koss
Universität Heidelberg
koss@cl.uni-heidelberg.de

PhD Research Proposal — Draft v3

January 2026

Abstract

We propose a computational framework for model inversion in whole-brain simulations that addresses the fundamental identifiability problem inherent to neuroimaging. Building on The Virtual Brain (TVB) platform and recent advances in region-specific mean-field modeling, we develop a principled inference scheme grounded in optimal transport theory. The key novelties are: (i) a rigorous identifiability framework based on monotone triangular transport maps that guarantees unique parameter recovery up to well-characterized equivalence classes, (ii) a thermodynamic interpretation connecting neural transfer functions to statistical mechanical partition functions, with entropy production in brain dynamics (measured via multivariate Ornstein-Uhlenbeck processes) determining identifiable information content, and (iii) a mathematical correspondence between mean-field neural dynamics and transformer attention mechanisms that enables cross-pollination of techniques. We prove that the Knothe-Rosenblatt transport provides the unique optimal coupling between parameter and observable spaces, enabling valid counterfactual reasoning about circuit parameters. The framework is general and applicable to any TVB-compatible neural mass model. We present a conservative exemplary application to schizophrenia, motivated by recent computational psychiatry findings on thalamocortical dysconnectivity.

Contents

1	Clinical Motivation and Context	4
1.1	The Promise of Computational Neuroimaging	4
1.2	What Can This Framework Offer Clinically?	4
1.3	Related Work	5
1.3.1	The Virtual Brain Platform	5
1.3.2	Region-Specific Mean-Field Models	5
1.3.3	Dynamic Causal Modeling	5
1.3.4	Computational Psychiatry and Schizophrenia	5
2	Mathematical Framework	5
2.1	The Hodgkin-Huxley Model: From Channels to Populations	5
2.1.1	Single Neuron Dynamics	6
2.1.2	Conductance-Based Synaptic Input	6
2.1.3	The Effective Membrane Equation	6

2.2	The Transfer Function	7
2.3	Mean-Field Population Dynamics	8
2.4	Whole-Brain Network Model	8
2.5	Hemodynamic Forward Model	8
2.6	The Inverse Problem	9
3	Identifiability via Brenier’s Theorem	9
3.1	The Identifiability Problem	9
3.2	Brenier’s Theorem: The Foundation	9
3.3	Knothe-Rosenblatt Transport and Triangular Maps	10
3.4	Structural Causal Models and TMI Maps	10
3.5	Application to Brain Model Inversion	11
3.5.1	The Forward Model as Transport	11
3.5.2	The Brenier Identifiability Criterion	11
3.6	Practical Implementation via Normalizing Flows	11
3.6.1	Causal Ordering on Effective Parameters	11
3.6.2	Triangular Transport via Normalizing Flows	12
3.6.3	Training Objective	12
3.7	Contrast with Dynamic Causal Modeling	12
3.7.1	Effective Connectivity in DCM	13
3.7.2	Key Differences from the Brenier Approach	13
3.7.3	When to Use Each Approach	14
3.8	Strengths of the Brenier Approach	14
3.9	Limitations and Alternatives	14
3.9.1	Alternative Approaches	14
4	Thermodynamic Interpretations	15
4.1	Transfer Functions as Partition Functions	16
4.2	Entropy Production in Brain Dynamics	16
4.2.1	Multivariate Ornstein-Uhlenbeck Process	16
4.2.2	Entropy Production Rate	16
4.2.3	Entropy Production and Consciousness	17
4.3	Identifiability and Entropy Production	17
4.4	Isomorphism with Transformer Architectures	17
4.4.1	The Correspondence	17
4.4.2	Dynamics-Level Correspondence	17
4.4.3	Entropy Production in Transformers	18
4.4.4	Training as Inversion	18
4.5	Hierarchical Free Energy Structure	19
5	Exemplary Application: Schizophrenia	19
5.1	Clinical Rationale	19
5.2	Hypotheses	19
5.3	Conservative Pilot Study Design	19
5.3.1	Participants	19
5.3.2	Assessments	20
5.3.3	Neuroimaging Protocol (3T)	20
5.3.4	Analysis Pipeline	20
5.3.5	Power Analysis	20
5.4	Expected Outcomes	20
5.5	Limitations	20

6	Validation Strategy	21
7	Timeline and Feasibility	21
8	Conclusion	21

1 Clinical Motivation and Context

1.1 The Promise of Computational Neuroimaging

Functional magnetic resonance imaging (fMRI) has transformed our understanding of brain function, yet its clinical translation remains limited (Voineskos et al., 2024). The blood-oxygen-level-dependent (BOLD) signal is an indirect, temporally blurred measurement of neural activity, with a typical repetition time (TR) of 2 seconds that severely limits temporal resolution. The signal chain from neural computation to measured BOLD involves multiple transformations:

$$\underbrace{\text{Synaptic activity}}_{\text{ms}} \rightarrow \underbrace{\text{Firing rates}}_{10\text{-}100\text{ ms}} \rightarrow \underbrace{\text{Metabolic demand}}_{100\text{ ms-}1\text{ s}} \rightarrow \underbrace{\text{Hemodynamics}}_{1\text{-}5\text{ s}} \rightarrow \underbrace{\text{BOLD}}_{T_R=2\text{ s}} \quad (1)$$

Each arrow represents information loss. The central clinical question is: *Can we extract mechanistically interpretable information about neural circuit function from BOLD data, and can such information inform diagnosis, prognosis, or treatment selection in psychiatric disorders?*

1.2 What Can This Framework Offer Clinically?

The proposed framework addresses several critical gaps in current neuroimaging-based approaches to psychiatry:

Beyond Correlation Matrices. Standard functional connectivity (FC) analyses report pairwise correlations between regional BOLD signals. While informative, these correlations lack mechanistic interpretation—they do not distinguish whether altered connectivity reflects changes in synaptic strength, neuronal excitability, or neuromodulatory tone. Our framework estimates *effective circuit parameters* that, while not directly measurable, are interpretable within a biophysical model.

Principled Uncertainty Quantification. Most machine learning approaches to neuroimaging biomarkers provide point predictions without uncertainty estimates. The optimal transport framework we develop provides full posterior distributions over parameters via normalizing flows, enabling clinicians to assess confidence in individual-level predictions.

Counterfactual Reasoning. A unique strength of the Brenier-based approach is that it enables valid counterfactual queries: “What would this patient’s functional connectivity have been if their E/I ratio had been different?” This capability is essential for treatment planning and mechanistic understanding.

Whole-Brain vs. Region-of-Interest Analyses. Traditional approaches force a choice between:

- **ROI-based:** High interpretability but ignores distributed network effects
- **Whole-brain:** Captures global dynamics but with reduced interpretability

The TVB framework naturally accommodates both scales: whole-brain structural connectivity constrains the network, while region-specific models capture local circuit properties.

Integration of Structural and Functional Data. The framework naturally incorporates:

- **Structural connectivity (SC)** from diffusion MRI
- **Functional connectivity (FC)** from resting-state fMRI
- **White matter properties** (myelination, conduction velocities)

1.3 Related Work

1.3.1 The Virtual Brain Platform

The Virtual Brain (TVB) is an open-source neuroinformatics platform for whole-brain network simulations (Sanz Leon et al., 2013; Ritter et al., 2013). The standard TVB pipeline:

1. Parcellate brain into R regions (typically 68–400)
2. Extract structural connectivity from diffusion MRI tractography
3. Place neural mass models at each node
4. Couple regions via structural connectivity with conduction delays
5. Simulate neural dynamics and generate synthetic BOLD via hemodynamic model
6. Fit model parameters to match empirical functional connectivity

Recent developments include: the Virtual Brain Twin project funded by the European Commission with 10 million EUR (Jirsa et al., 2023); TVB C++ providing 10× speedup for production simulations (Martín et al., 2025); the Bayesian Virtual Epileptic Patient (BVEP) for personalized epilepsy modeling (Hashemi et al., 2020); and amortized inference for privacy-preserving personalization.

1.3.2 Region-Specific Mean-Field Models

A critical recent advance is the development of region-specific neural mass models that respect local microcircuit architecture. Lorenzi et al. (2023) developed a cerebellar mean-field model incorporating four populations (granule cells, Golgi cells, molecular layer interneurons, Purkinje cells) with biophysically-constrained connectivity. When integrated into TVB (Lorenzi et al., 2025), this cerebellar mean-field model (cMF-TVB) achieved approximately 50% error reduction compared to generic oscillators, demonstrating that region-specific models significantly improve simulation fidelity.

1.3.3 Dynamic Causal Modeling

Dynamic Causal Modeling (DCM) (Friston et al., 2003, 2019) is the dominant framework for effective connectivity inference. DCM uses Bayesian inference with uncertainty quantification but has limitations:

- Phenomenological state equations (not biophysically grounded)
- Limited to small networks (typically 2–10 regions)
- Known identifiability concerns (Daunizeau et al., 2011)

Our framework addresses these limitations through biophysically-grounded models and principled identifiability analysis via optimal transport.

1.3.4 Computational Psychiatry and Schizophrenia

Computational approaches to schizophrenia have revealed consistent findings (Friston, 2022; Huys et al., 2016). Adams et al. (2022) demonstrated, using both EEG and fMRI paradigms with computational modeling, that schizophrenia is associated with a consistent loss of pyramidal cell synaptic gain across multiple experimental contexts. This finding motivates our exemplary application and suggests that effective E/I parameters may be sensitive biomarkers.

2 Mathematical Framework

2.1 The Hodgkin-Huxley Model: From Channels to Populations

To ground our framework in biophysics, we begin with the Hodgkin-Huxley (HH) model (Hodgkin & Huxley, 1952) and systematically derive mean-field population equations.

2.1.1 Single Neuron Dynamics

The HH model describes action potential generation via voltage-gated ion channels:

Definition 2.1 (Hodgkin-Huxley Equations). *The membrane potential $V(t)$ evolves according to:*

$$C_m \frac{dV}{dt} = -g_{Na}m^3h(V - E_{Na}) - g_Kn^4(V - E_K) - g_L(V - E_L) + I_{ext} \quad (2)$$

where the gating variables $(m, h, n) \in [0, 1]^3$ represent channel activation/inactivation probabilities:

$$\frac{dm}{dt} = \alpha_m(V)(1 - m) - \beta_m(V)m \quad (3)$$

$$\frac{dh}{dt} = \alpha_h(V)(1 - h) - \beta_h(V)h \quad (4)$$

$$\frac{dn}{dt} = \alpha_n(V)(1 - n) - \beta_n(V)n \quad (5)$$

Here $\alpha_x(V), \beta_x(V)$ are voltage-dependent rate functions, g_{Na}, g_K, g_L are maximal conductances, and E_{Na}, E_K, E_L are reversal potentials.

Remark 2.2 (Timescale Separation). *The gating variables equilibrate on timescales $\tau_m \sim 0.1$ ms, $\tau_h \sim 1$ ms, $\tau_n \sim 1$ ms, while action potentials occur on ~ 1 ms timescales. This separation enables adiabatic elimination.*

2.1.2 Conductance-Based Synaptic Input

In networks, neurons receive synaptic input that modulates membrane conductance:

Definition 2.3 (Synaptic Conductances). *For a neuron receiving input from populations $s = 1, \dots, K$:*

$$I_{syn}(t) = \sum_s g_s(t)(V(t) - E_s) \quad (6)$$

where $g_s(t)$ is the synaptic conductance from population s :

$$g_s(t) = \sum_{j \in s} \sum_k Q_s \cdot \alpha(t - t_j^{(k)}) \quad (7)$$

with Q_s the quantal conductance, $t_j^{(k)}$ the k -th spike time of neuron j , and $\alpha(t)$ the synaptic kernel (e.g., alpha function with time constant τ_s).

Under Poisson spiking assumptions, the total synaptic conductance has mean and variance:

$$\mu_{g_s} = K_s Q_s \tau_s \nu_s \quad (8)$$

$$\sigma_{g_s}^2 = K_s Q_s^2 \tau_s \nu_s / 2 \quad (9)$$

where K_s is the number of synapses from population s and ν_s is its firing rate.

2.1.3 The Effective Membrane Equation

For conductance-based neurons receiving fluctuating synaptic input, the membrane potential follows an effective stochastic differential equation (El Boustani & Destexhe, 2009):

Proposition 2.4 (Effective Membrane Dynamics). *Under the diffusion approximation (high rate, small quantal conductances), the membrane potential approximately follows:*

$$\tau_V \frac{dV}{dt} = -(V - \mu_V) + \sigma_V \eta(t) \quad (10)$$

where $\eta(t)$ is white noise and:

$$\mu_G = g_L + \sum_s K_s Q_s \tau_s \nu_s \quad (11)$$

$$\mu_V = \frac{g_L E_L + \sum_s K_s Q_s \tau_s \nu_s E_s}{\mu_G} \quad (12)$$

$$\sigma_V^2 = \sum_s \frac{K_s Q_s^2 \tau_s \nu_s}{2\mu_G^2} \cdot \frac{(E_s - \mu_V)^2 \tau_s}{\tau_V + \tau_s} \quad (13)$$

$$\tau_V = C_m / \mu_G \quad (14)$$

This is an **Ornstein-Uhlenbeck process** with mean μ_V , variance σ_V^2 , and time constant τ_V —all of which depend on the input firing rates $\{\nu_s\}$.

2.2 The Transfer Function

The key quantity linking input firing rates to output firing rate is the *transfer function*.

Definition 2.5 (Transfer Function). *The transfer function $\mathcal{T} : \mathbb{R}_+^K \times \Theta \rightarrow \mathbb{R}_+$ maps input firing rates to output firing rate:*

$$\nu_{out} = \mathcal{T}(\nu_1, \dots, \nu_K; \theta) \quad (15)$$

where θ contains biophysical parameters (conductances, time constants, reversal potentials, connectivity).

Theorem 2.6 (Semi-Analytical Transfer Function). *For a conductance-based neuron with effective membrane dynamics (10) and threshold V_{th} , the stationary firing rate is:*

$$\mathcal{T}(\nu) = \frac{1}{2\tau_V} \operatorname{erfc} \left(\frac{V_{th} - \mu_V(\nu)}{\sqrt{2}\sigma_V(\nu)} \right) \quad (16)$$

where erfc is the complementary error function.

Proof sketch. For the OU process (10), the stationary distribution is Gaussian:

$$p_{ss}(V) = \frac{1}{\sqrt{2\pi}\sigma_V} \exp \left(-\frac{(V - \mu_V)^2}{2\sigma_V^2} \right) \quad (17)$$

The firing rate equals the probability flux across threshold, which for a reflecting boundary at V_{reset} and absorbing boundary at V_{th} yields (16) in the limit $V_{reset} \rightarrow -\infty$ (see Zerlaut et al. (2018) for the full derivation with finite reset). \square

Remark 2.7 (Functional Form). *In practice, the transfer function is often fit to a more flexible form:*

$$\mathcal{T}(\nu) = F \left(\frac{\mu_V(\nu) - V_{th}^{eff}}{\sigma_V(\nu)} \right) \quad (18)$$

where F is a sigmoidal function and V_{th}^{eff} is an effective threshold capturing spike initiation dynamics. The key insight is that \mathcal{T} depends on input rates only through μ_V and σ_V .

2.3 Mean-Field Population Dynamics

Assumption 2.8 (Asynchronous Irregular Regime). *The network operates in the asynchronous irregular (AI) regime where:*

1. *Spike times across neurons are approximately independent (no synchrony)*
2. *Individual neurons fire irregularly (high coefficient of variation)*
3. *Population activity is approximately stationary over timescales $\gg \tau_{syn}$*

Under Assumption 2.8, the population firing rate evolves according to:

Definition 2.9 (Mean-Field Dynamics). *For a circuit with P interacting populations, the firing rates $\boldsymbol{\nu}(t) = (\nu_1(t), \dots, \nu_P(t))^\top$ evolve as:*

$$T_p \frac{d\nu_p}{dt} = \mathcal{T}_p(\boldsymbol{\nu}; \boldsymbol{\theta}) - \nu_p + \eta_p(t), \quad p = 1, \dots, P \quad (19)$$

where T_p is the effective time constant of population p and $\eta_p(t)$ captures finite-size fluctuations.

This is a **relaxation dynamics** toward a fixed point $\boldsymbol{\nu}^*$ satisfying $\nu_p^* = \mathcal{T}_p(\boldsymbol{\nu}^*)$.

2.4 Whole-Brain Network Model

Following the TVB framework, we couple R brain regions:

Definition 2.10 (Whole-Brain Model). *Let $\mathbf{C} \in \mathbb{R}^{R \times R}$ be the structural connectivity matrix and $\mathbf{D} \in \mathbb{R}^{R \times R}$ the distance matrix. The coupled system is:*

$$T_p^{(r)} \frac{d\nu_p^{(r)}}{dt} = \mathcal{T}_p^{(r)}(\boldsymbol{\nu}^{(r)}, \boldsymbol{\nu}^{aff,(r)}; \boldsymbol{\theta}^{(r)}) - \nu_p^{(r)} + \eta_p^{(r)}(t) \quad (20)$$

where the afferent input from other regions is:

$$\nu^{aff,(r)}(t) = G \sum_{s \neq r} C_{rs} \nu_{out}^{(s)}(t - D_{rs}/v) \quad (21)$$

with G the global coupling strength and v the conduction velocity.

2.5 Hemodynamic Forward Model

The Balloon-Windkessel model connects neural activity to BOLD (Buxton et al., 1998; Friston et al., 2000):

Definition 2.11 (Balloon-Windkessel Model). *The BOLD signal $y(t)$ is generated from neural activity $z(t) = \sum_p w_p \nu_p(t)$ via:*

$$\dot{s} = z - \kappa s - \gamma(f - 1) \quad (22)$$

$$\dot{f} = s \quad (23)$$

$$\tau_0 \dot{v} = f - v^{1/\alpha} \quad (24)$$

$$\tau_0 \dot{q} = \frac{f}{\rho} (1 - (1 - \rho)^{1/f}) - \frac{q}{v} v^{1/\alpha} \quad (25)$$

$$y = V_0 \left[k_1(1 - q) + k_2 \left(1 - \frac{q}{v} \right) + k_3(1 - v) \right] \quad (26)$$

where s is vasodilatory signal, f is blood flow, v is blood volume, q is deoxyhemoglobin content.

2.6 The Inverse Problem

Definition 2.12 (Model Inversion Problem). *Given empirical BOLD $\mathbf{Y}^{emp} \in \mathbb{R}^{T \times R}$, structural connectivity \mathbf{C} , forward model $\mathcal{M} : \Theta \rightarrow \mathbb{R}^{T \times R}$, and prior $\pi_0 \in \mathcal{P}(\Theta)$, find the posterior $\pi^* \in \mathcal{P}(\Theta)$.*

The forward model is the composition:

$$\mathcal{M}(\boldsymbol{\theta}) = \text{HRF} \circ \text{Network-MF}(\boldsymbol{\theta}) \quad (27)$$

3 Identifiability via Brenier's Theorem

This section develops our main theoretical contribution: using optimal transport theory, specifically Brenier's uniqueness theorem, to provide rigorous identifiability guarantees for neural model inversion.

3.1 The Identifiability Problem

Definition 3.1 (Parameter Degeneracy). *Multiple parameter configurations produce statistically indistinguishable BOLD dynamics:*

$$\exists \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \in \Theta : \quad \mathcal{M}(\boldsymbol{\theta}_1) \stackrel{d}{=} \mathcal{M}(\boldsymbol{\theta}_2) \quad (28)$$

where $\stackrel{d}{=}$ denotes equality in distribution.

Sources of degeneracy:

1. **Compensatory changes:** Increased g_E + increased g_I can maintain E/I balance
2. **Time constant trade-offs:** Faster dynamics + weaker coupling \approx slower dynamics + stronger coupling (after hemodynamic filtering)
3. **HRF uncertainty:** Neural parameter changes absorbed by HRF parameters

Rather than viewing degeneracy as a problem to overcome, we embrace it through optimal transport theory, which provides a principled way to handle equivalence classes.

3.2 Brenier's Theorem: The Foundation

We begin with the fundamental result from optimal transport theory:

Theorem 3.2 (Brenier, 1991). *Let μ, ν be probability measures on \mathbb{R}^n with μ absolutely continuous with respect to Lebesgue measure. For the quadratic cost $c(x, y) = \|x - y\|^2$, there exists a **unique** optimal transport map $T^* : \mathbb{R}^n \rightarrow \mathbb{R}^n$ solving:*

$$T^* = \arg \min_{T: T_{\#}\mu = \nu} \int_{\mathbb{R}^n} \|x - T(x)\|^2 d\mu(x) \quad (29)$$

Moreover, $T^* = \nabla \phi$ for some convex function $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$.

Proof sketch (see Villani (2003); McCann & Guillen (2011)). The key steps are:

1. Relax to the Kantorovich problem (transport plans instead of maps)
2. Establish duality: optimal plan concentrated on subdifferential of convex function
3. Use absolute continuity of μ to show subdifferential is single-valued a.e.
4. Conclude $T^* = \nabla \phi$ is unique μ -a.e.

□

Corollary 3.3 (Monotonicity). *The Brenier map $T^* = \nabla\phi$ is monotone:*

$$\langle T^*(x) - T^*(y), x - y \rangle \geq 0 \quad \forall x, y \in \mathbb{R}^n \quad (30)$$

This follows from $\nabla^2\phi \succeq 0$ (Hessian positive semi-definite) for convex ϕ .

3.3 Knothe-Rosenblatt Transport and Triangular Maps

For our purposes, a special case of Brenier transport is particularly relevant:

Definition 3.4 (Triangular Monotone Increasing Map). *A map $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is **triangular monotone increasing (TMI)** if:*

$$T(x) = \begin{pmatrix} T_1(x_1) \\ T_2(x_1, x_2) \\ \vdots \\ T_n(x_1, \dots, x_n) \end{pmatrix} \quad (31)$$

where each T_i is strictly monotone increasing in x_i .

Theorem 3.5 (Knothe-Rosenblatt Transport). *For any absolutely continuous measures μ, ν on \mathbb{R}^n , there exists a **unique** TMI map S such that $S_{\#}\mu = \nu$. This map is called the Knothe-Rosenblatt (KR) rearrangement.*

Remark 3.6 (Relation to Brenier). *The KR map coincides with the Brenier map when the target measure is a product measure. In general, they differ, but both provide **unique** transport maps.*

3.4 Structural Causal Models and TMI Maps

The connection to identifiability comes through structural causal models (SCMs):

Definition 3.7 (Structural Causal Model). *An SCM with causal ordering $\pi = (1, 2, \dots, n)$ has the form:*

$$X_1 = f_1(U_1) \quad (32)$$

$$X_2 = f_2(X_1, U_2) \quad (33)$$

$$\vdots \quad (34)$$

$$X_n = f_n(X_1, \dots, X_{n-1}, U_n) \quad (35)$$

where U_1, \dots, U_n are mutually independent exogenous variables.

Proposition 3.8 (SCM as TMI Map). *If each structural function f_i is strictly monotone in U_i , then the map $U \mapsto X$ is a TMI map.*

Theorem 3.9 (Identifiability via TMI (Javaloy et al., 2023; Xi & Bloem-Reddy, 2023)). *If the data-generating SCM has:*

1. *Mutually independent exogenous variables U*
2. *Structural functions that are TMI in U*

*Then the SCM is **identifiable up to component-wise invertible transformations** from observational data alone.*

3.5 Application to Brain Model Inversion

We now apply this framework to our inverse problem.

3.5.1 The Forward Model as Transport

Definition 3.10 (Effective Parameter Space). *Define the effective parameter space as the quotient:*

$$\Theta_{\text{eff}} = \Theta / \sim \quad (36)$$

where $\theta_1 \sim \theta_2$ iff $\mathcal{M}(\theta_1) \stackrel{d}{=} \mathcal{M}(\theta_2)$ (same FC distribution).

Proposition 3.11 (Forward Model Induces Transport). *The forward model $\theta \mapsto \text{FC}(\theta)$ induces a transport plan $\gamma \in \Gamma(\mu_\theta, \nu_{\text{FC}})$ between:*

- *Source: prior distribution μ_θ on parameters*
- *Target: distribution ν_{FC} of functional connectivity matrices*

3.5.2 The Brenier Identifiability Criterion

Theorem 3.12 (Brenier Identifiability for Brain Models). *The inverse map $\text{FC} \mapsto \theta_{\text{eff}}$ is unique if and only if:*

1. *The forward map restricted to Θ_{eff} is injective*
2. *There exists a convex potential $\phi : \mathcal{F} \rightarrow \mathbb{R}$ (where \mathcal{F} is FC space) such that the inverse map equals $\nabla \phi$*

*Under these conditions, Brenier’s theorem guarantees the inverse is the **unique optimal transport map**.*

Proof. By Brenier’s theorem (3.2), if the inverse $\text{FC} \mapsto \theta_{\text{eff}}$ can be written as $\nabla \phi$ for convex ϕ , then it is the unique optimal transport map from ν_{FC} to $\mu_{\theta_{\text{eff}}}$. The inverse $T^{-1} = \nabla \phi^*$ where ϕ^* is the Legendre transform (also convex). Uniqueness of Brenier map implies uniqueness of inverse. \square

Corollary 3.13 (Counterfactual Identifiability). *Under the conditions of Theorem 3.12, counterfactual queries “What would FC have been if θ_{eff} had been θ' ?” are identifiable.*

3.6 Practical Implementation via Normalizing Flows

Rather than computing Brenier maps directly (intractable), we use autoregressive normalizing flows that are TMI by construction (Khemakhem et al., 2021; Javaloy et al., 2023).

3.6.1 Causal Ordering on Effective Parameters

Assumption 3.14 (Effective Parameter Ordering). *We assume a causal ordering on effective parameters:*

$$\theta_{\text{eff}} = (\theta_{\text{eff}}^{(1)}, \theta_{\text{eff}}^{(2)}, \dots, \theta_{\text{eff}}^{(k)}) \quad (37)$$

where more identifiable parameters come first (e.g., global coupling before regional parameters).

Algorithm 1 Brenier-Based Model Inversion via Normalizing Flows

Require: Empirical FC distribution $\{\text{FC}_{\text{emp}}^{(j)}\}_{j=1}^N$, prior π_0 on Θ_{eff}

Require: Flow architecture T_ϕ (TMI by construction)

```

1: Initialize flow parameters  $\phi$ 
2: for epoch = 1, ...,  $E$  do
3:   for mini-batch  $\{\text{FC}^{(j)}\}$  do
4:     Compute inverse:  $z^{(j)} = T_\phi^{-1}(\text{FC}^{(j)})$  ▷ Tractable for TMI flows
5:     Evaluate log-likelihood:
6:      $\ell(\phi) = \sum_j \left[ \log p_z(z^{(j)}) + \log |\det \nabla T_\phi^{-1}(\text{FC}^{(j)})| \right]$ 
7:     Update  $\phi \leftarrow \phi + \eta \nabla_\phi \ell(\phi)$ 
8:   end for
9: end for
10: return Trained flow  $T_\phi$ 

```

3.6.2 Triangular Transport via Normalizing Flows

Definition 3.15 (Autoregressive Normalizing Flow). *An autoregressive normalizing flow $T_\phi : \mathbb{R}^k \rightarrow \mathbb{R}^m$ is a TMI map:*

$$T_\phi(z) = \begin{pmatrix} T_1(z_1; \phi_1) \\ T_2(z_1, z_2; \phi_2) \\ \vdots \\ T_k(z_1, \dots, z_k; \phi_k) \end{pmatrix} \quad (38)$$

where $z \sim \mathcal{N}(0, I)$ and each T_i is implemented as a monotonic neural network (e.g., via integration of positive functions).

Proposition 3.16 (Jacobian Structure). *The Jacobian of a TMI flow is lower triangular:*

$$\frac{\partial T_\phi}{\partial z} = \begin{pmatrix} \frac{\partial T_1}{\partial z_1} & 0 & \dots & 0 \\ \frac{\partial T_2}{\partial z_1} & \frac{\partial T_2}{\partial z_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial T_k}{\partial z_1} & \frac{\partial T_k}{\partial z_2} & \dots & \frac{\partial T_k}{\partial z_k} \end{pmatrix} \quad (39)$$

Thus $\det(\nabla T_\phi) = \prod_i \frac{\partial T_i}{\partial z_i}$ is $O(k)$ to compute (vs. $O(k^3)$ for general flows).

3.6.3 Training Objective

The trained flow provides:

- **Posterior samples:** $\theta_{\text{eff}} \sim T_\phi(z)$ with $z \sim \mathcal{N}(0, I)$
- **Density evaluation:** $p(\theta_{\text{eff}}|\text{FC}) = p_z(T_\phi^{-1}(\text{FC})) \cdot |\det \nabla T_\phi^{-1}|$
- **Counterfactuals:** Transport FC to different parameter values via T_ϕ

Theorem 3.17 (Convergence to KR Transport). *As flow capacity increases (more layers, wider networks) and training data grows, the learned TMI flow T_ϕ converges to the true KR transport map, which by Theorem 3.5 is unique.*

3.7 Contrast with Dynamic Causal Modeling

Dynamic Causal Modeling (DCM) (Friston et al., 2003, 2019) is the dominant framework for inferring effective connectivity from neuroimaging data. While our approach shares DCM’s goal of moving beyond correlational analyses, the mathematical foundations and identifiability guarantees differ substantially.

3.7.1 Effective Connectivity in DCM

DCM models neuronal dynamics via a bilinear state-space system:

$$\dot{\mathbf{z}} = \left(\mathbf{A} + \sum_j u_j \mathbf{B}^{(j)} \right) \mathbf{z} + \mathbf{C} \mathbf{u} \quad (40)$$

where \mathbf{z} represents hidden neuronal states, \mathbf{A} is the *endogenous* (intrinsic) connectivity, $\mathbf{B}^{(j)}$ captures *modulatory* effects of experimental inputs u_j , and \mathbf{C} represents *driving* inputs. The matrix \mathbf{A} defines “effective connectivity”—the directed causal influence that one region exerts over another (Friston, 2011).

Inference in DCM proceeds via variational Laplace, approximating the posterior as Gaussian:

$$p(\boldsymbol{\theta}|\mathbf{Y}) \approx \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \quad (41)$$

Model comparison uses the variational free energy \mathcal{F} as an approximation to log model evidence.

3.7.2 Key Differences from the Brenier Approach

Table 1: Comparison of DCM and Brenier-based model inversion

Aspect	DCM	Brenier/OT Approach
State equations	Phenomenological bilinear system (40)	Biophysically-derived mean-field (Hodgkin-Huxley based)
Posterior approximation	Gaussian (variational Laplace)	Exact via normalizing flows
Identifiability	Known issues; no formal guarantees (Daunizeau et al., 2011; Razi & Friston, 2016)	Brenier uniqueness theorem guarantees unique transport map
Network scale	Small (2–10 regions typical)	Whole-brain (~ 100 regions)
Likelihood	Gaussian on time series	Transport cost on FC distributions
Counterfactuals	Limited; requires re-estimation	Native support via unique transport map
Effective parameters	Connectivity matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$	Quotient space Θ_{eff} with entropy-based characterization

Identifiability Guarantees. A fundamental limitation of DCM is the lack of formal identifiability guarantees. Daunizeau et al. (2011) demonstrated that DCM parameters can be practically non-identifiable even with high SNR data, and Razi & Friston (2016) showed that different model parameterizations can yield equivalent fits. The Brenier approach addresses this directly: Theorem 3.2 guarantees that the optimal transport map is *unique*, and by constructing inference via TMI flows (Definition 3.15), we inherit this uniqueness. Non-identifiable parameter directions are explicitly characterized via the quotient $\Theta_{\text{eff}} = \Theta/\sim$ rather than manifesting as posterior correlations or flat directions in the free energy landscape.

Posterior Representation. DCM’s variational Laplace approximation assumes Gaussian posteriors, which can be severely misspecified for:

- Multi-modal posteriors (e.g., when multiple connectivity patterns explain data equally well)

- Bounded parameters (e.g., time constants must be positive)
- Parameters with complex dependencies

Normalizing flows represent arbitrary posterior shapes, with the TMI structure ensuring computational tractability ($O(k)$ Jacobian computation) while maintaining Brenier uniqueness.

Functional vs. Effective Connectivity. DCM distinguishes “functional connectivity” (observed correlations) from “effective connectivity” (model-based causal influences). Our framework offers a third perspective: *effective parameters* Θ_{eff} are not connectivity matrices per se, but the identifiable degrees of freedom in the full biophysical parameter space. These may include E/I ratios, time constants, and coupling strengths—quantities that determine connectivity but are more directly interpretable in terms of circuit mechanisms.

The Role of Optimal Transport. In DCM, the likelihood compares predicted and observed time series point-wise:

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \prod_t \mathcal{N}(y_t; g(\mathbf{z}_t(\boldsymbol{\theta})), \sigma^2) \quad (42)$$

This treats each time point independently, ignoring the distributional structure of FC dynamics. The Wasserstein distance, by contrast, compares *distributions* of FC states, capturing:

- Temporal variability in connectivity patterns
- Geometric structure of the FC manifold
- Distributional shifts rather than point-wise deviations

More importantly, Brenier’s theorem transforms the inference problem from likelihood maximization to finding a unique optimal coupling—a fundamentally different mathematical object with stronger guarantees.

3.7.3 When to Use Each Approach

DCM remains appropriate for:

- Task-based fMRI with well-defined experimental manipulations
- Small, hypothesis-driven networks (e.g., testing specific circuit models)
- Settings where Gaussian posteriors are adequate

The Brenier approach is preferred for:

- Resting-state fMRI where experimental inputs are absent
- Whole-brain analyses requiring scalability
- Applications requiring formal identifiability guarantees
- Counterfactual reasoning about parameter changes
- Non-Gaussian or multi-modal posteriors

Remark 3.18 (Complementary Perspectives). *The approaches are not mutually exclusive. DCM’s bilinear structure (40) could be embedded within our framework by using it as the forward model \mathcal{M} , with Brenier-based inference replacing variational Laplace. This would combine DCM’s interpretable effective connectivity parameterization with the identifiability guarantees of optimal transport.*

3.8 Strengths of the Brenier Approach

3.9 Limitations and Alternatives

3.9.1 Alternative Approaches

Gaussian Likelihood (MSE). The standard approach uses:

$$p(\text{FC}|\theta) \propto \exp\left(-\frac{\|\text{FC} - \text{FC}_{\text{sim}}(\theta)\|^2}{2\sigma^2}\right) \quad (43)$$

Table 2: Strengths of Brenier-based identifiability

Strength	Explanation
Theoretical foundation	Brenier uniqueness is a theorem, not an assumption
No likelihood specification	OT uses geometry; no need for $p(\text{FC} \theta)$
Handles non-Gaussian	Works for any absolutely continuous distributions
Natural uncertainty	Flow gives full posterior, not just point estimate
Counterfactual validity	Uniqueness enables valid causal inference
Computational tractability	TMI flows have $O(n)$ Jacobian computation

Table 3: Limitations of Brenier approach and mitigations

Limitation	Explanation	Mitigation
Absolute continuity	μ must have density w.r.t. Lebesgue	Add small noise / regularization
Ordering dependence	KR map depends on variable ordering	Choose ordering by identifiability
Approximation error	Finite-capacity flows may not exactly represent transport	Increase expressivity; universality results
Non-convex training	Flow training has local minima	Multiple initializations
FC manifold structure	FC matrices live on SPD manifold	Riemannian flows or tangent projection
Doesn't solve degeneracy	If forward model has symmetries, Brenier gives one inverse	Must define Θ_{eff} correctly

Advantage: Simple, well-understood. **Disadvantage:** Assumes Gaussian noise; doesn't capture distributional structure; no uniqueness guarantees.

Simulation-Based Inference (SBI). Methods like neural posterior estimation (Cranmer et al., 2020) learn $p(\theta|\text{FC})$ directly. **Advantage:** Flexible; handles complex likelihoods. **Disadvantage:** No uniqueness guarantees; posterior may be multi-modal without principled resolution.

Wasserstein Likelihood. Our previous formulation used:

$$p(\text{FC}^{\text{emp}}|\theta) \propto \exp\left(-\frac{W_2^2(\mu^{\text{emp}}, \mu^{\text{sim}}(\theta))}{2\sigma^2}\right) \quad (44)$$

Advantage: Captures distributional discrepancy. **Disadvantage:** Ad-hoc; no principled connection to identifiability.

Why Brenier is Preferable. The Brenier/KR approach provides **theoretical guarantees** (uniqueness via theorem, not assumption), **counterfactual validity** (essential for causal reasoning), and **computational tractability** (TMI flows). The cost is requiring careful definition of Θ_{eff} and choosing an appropriate ordering.

4 Thermodynamic Interpretations

This section develops deep connections between neural dynamics, statistical mechanics, and transformer architectures as rigorous mathematical correspondences.

4.1 Transfer Functions as Partition Functions

Theorem 4.1 (Transfer Function as Partition Function). *The mean-field transfer function (16) has the structure of a statistical mechanical partition function. Define:*

$$E(V) = -\log \Phi(V) + \frac{(V - \mu_V)^2}{2\sigma_V^2} \quad (45)$$

where $\Phi(V)$ is the instantaneous firing probability. Then:

$$\nu = \int_{\mathbb{R}} \Phi(V) p(V) dV \propto \int_{\mathbb{R}} e^{-\beta_{\text{eff}} E(V)} dV = Z(\beta_{\text{eff}}) \quad (46)$$

with **effective inverse temperature** $\beta_{\text{eff}} = 1/\sigma_V^2$.

Corollary 4.2 (Temperature Interpretation). • High input variance σ_V^2 (“high temperature”): diffuse neural responses
• Low input variance σ_V^2 (“low temperature”): sharp, deterministic responses
This parallels the softmax temperature in attention mechanisms.

4.2 Entropy Production in Brain Dynamics

Recent work by Gilson et al. (2023) establishes that entropy production rate in brain dynamics correlates with consciousness levels. We incorporate this into our framework.

4.2.1 Multivariate Ornstein-Uhlenbeck Process

At the mesoscale, brain dynamics can be approximated by a multivariate OU process:

Definition 4.3 (Multivariate OU Process). *The BOLD signal $\mathbf{x}(t) \in \mathbb{R}^R$ follows:*

$$d\mathbf{x} = -\mathbf{B}\mathbf{x} dt + \mathbf{D}^{1/2} d\mathbf{W} \quad (47)$$

where $\mathbf{B} \in \mathbb{R}^{R \times R}$ is the drift matrix (effective connectivity), \mathbf{D} is the noise covariance, and \mathbf{W} is a Wiener process.

The stationary covariance \mathbf{S} satisfies the Lyapunov equation:

$$\mathbf{B}\mathbf{S} + \mathbf{S}\mathbf{B}^\top = \mathbf{D} \quad (48)$$

4.2.2 Entropy Production Rate

Definition 4.4 (Onsager Matrix). *The Onsager matrix is:*

$$\mathbf{L} = \mathbf{B}\mathbf{S} \quad (49)$$

Its antisymmetric part encodes irreversibility:

$$\mathbf{Q} = \frac{\mathbf{L} - \mathbf{L}^\top}{2} \quad (50)$$

Theorem 4.5 (Entropy Production Rate (Godrèche & Luck, 2018; Gilson et al., 2023)). *The entropy production rate of the MOU process (47) is:*

$$\Phi = \text{tr}(\mathbf{Q}\mathbf{S}^{-1}\mathbf{Q}\mathbf{D}^{-1}) \geq 0 \quad (51)$$

with equality ($\Phi = 0$) if and only if the process is time-reversible ($\mathbf{Q} = 0$, i.e., $\mathbf{B}\mathbf{D} = \mathbf{D}\mathbf{B}^\top$).

Corollary 4.6 (Irreversibility from Asymmetric Connectivity). *Non-zero entropy production ($\Phi > 0$) occurs when effective connectivity \mathbf{B} is asymmetric. This asymmetry drives directed information flow and indicates far-from-equilibrium dynamics.*

4.2.3 Entropy Production and Consciousness

Theorem 4.7 (Gilson et al., 2023). *Fitting MOU processes to fMRI data during sleep stages reveals:*

$$\Phi_{\text{wake}} > \Phi_{N1} > \Phi_{N2} > \Phi_{N3} \quad (52)$$

Entropy production monotonically decreases with sleep depth, providing a thermodynamic signature of consciousness.

Remark 4.8 (Implications for Our Framework). 1. Φ is a measurable quantity from BOLD data (via MOU fitting)

2. Parameters affecting Φ are likely identifiable (they influence observable irreversibility)

3. Clinical populations (e.g., schizophrenia, disorders of consciousness) may show altered Φ

4.3 Identifiability and Entropy Production

We now connect entropy production to parameter identifiability:

Theorem 4.9 (Identifiability via Entropy Production). *Let θ be parameters and FC be observed functional connectivity. The identifiable information is bounded by:*

$$I(\theta; \text{FC}) \leq I(\theta; \nu) \leq H(\theta) \quad (53)$$

The “lost” information $\Delta S = H(\theta) - I(\theta; \text{FC})$ equals the entropy produced during forward model coarse-graining and determines the dimension of the non-identifiable manifold.

Proposition 4.10 (Entropy Production Determines Identifiability). *Parameters that affect Φ (entropy production rate) are more likely identifiable than those that don’t, because Φ is directly measurable from BOLD dynamics via MOU fitting.*

Definition 4.11 (Effective Parameters via Entropy). *The effective parameters Θ_{eff} can be characterized as those functions of θ that:*

1. Affect Φ (entropy production rate)
2. Affect stationary covariance \mathbf{S} (functional connectivity structure)
3. Are preserved under hemodynamic filtering

4.4 Isomorphism with Transformer Architectures

The mathematical structure of neural mean-field models is isomorphic to transformer attention. This is not an analogy—it is a precise mathematical correspondence.

4.4.1 The Correspondence

4.4.2 Dynamics-Level Correspondence

Theorem 4.12 (Isomorphism of Neural and Transformer Dynamics). *The neural mean-field update:*

$$\nu_r^{(t+1)} = \nu_r^{(t)} + \frac{dt}{T_r} \left[\mathcal{T}_r \left(\sum_s W_{rs} \nu_s^{(t)} \right) - \nu_r^{(t)} \right] \quad (54)$$

is isomorphic to the transformer attention update:

$$x_i^{(\ell+1)} = x_i^{(\ell)} + \sum_j \frac{\exp(q_i \cdot k_j / \sqrt{d})}{\sum_{j'} \exp(q_i \cdot k_{j'} / \sqrt{d})} v_j \quad (55)$$

under the correspondences in Table 4.

Table 4: Correspondence between neural mean-field and transformers

Concept	Neural Mean-Field	Transformer
State variable	Firing rate ν_r	Token embedding x_i
Interaction weights	Effective connectivity W_{rs}	Attention weights A_{ij}
Nonlinearity	Transfer function $\mathcal{T}(\cdot)$	Softmax attention
Partition function	$Z = \int e^{-\beta E(V)} dV$	$Z = \sum_j e^{q_i \cdot k_j / \sqrt{d}}$
Temperature	$\beta_{\text{eff}} = 1/\sigma_V^2$	$\beta = \sqrt{d_k}$
Dynamics	$\tau \dot{\nu} = \mathcal{T}(\nu) - \nu$	$x^{(\ell+1)} = x^{(\ell)} + \text{Attn}(x^{(\ell)})$
Asymmetry	$W_{rs} \neq W_{sr}$	$A_{ij} \neq A_{ji}$
Entropy production	$\Phi = \text{tr}(\mathbf{Q}\mathbf{S}^{-1}\mathbf{Q}\mathbf{D}^{-1})$	Layer-wise information loss

Proof sketch. Both systems implement relaxation dynamics toward energy minima:

1. **Neural:** Energy $E(\nu) = \sum_r U_r(\nu_r) + \frac{1}{2} \sum_{r,s} W_{rs} \nu_r \nu_s$
2. **Transformer:** Hopfield energy $E(\mathbf{x}) = -\frac{1}{\beta} \log \sum_j \exp(\beta q \cdot k_j) + \frac{1}{2} \|\mathbf{x}\|^2$

The transfer function \mathcal{T} and softmax attention both compute gradients of log-partition functions:

$$\mathcal{T}(\mu_V) = \frac{\partial}{\partial \mu_V} \log Z(\mu_V, \sigma_V), \quad \text{Attn}(q) = \frac{\partial}{\partial q} \log \sum_j e^{q \cdot k_j} \quad (56)$$

This establishes the isomorphism at the level of gradient flows on energy landscapes. □

4.4.3 Entropy Production in Transformers

Proposition 4.13 (Transformer Entropy Production). *The asymmetry of attention ($A_{ij} \neq A_{ji}$) creates irreversible dynamics analogous to asymmetric effective connectivity in brains. Define:*

$$\Phi_{\text{transformer}} = \sum_{i,j} (A_{ij} - A_{ji})^2 \quad (57)$$

This measures “directed attention flow” and is non-zero for typical attention patterns.

Remark 4.14 (Implications). 1. *Both brains and transformers operate far from equilibrium*

2. *Asymmetric interactions drive both neural computation and attention*

3. *Entropy production may be a universal signature of information processing*

4.4.4 Training as Inversion

Corollary 4.15 (Cross-Pollination). 1. **Brain** \rightarrow **Transformer**: *Entropy production analysis may diagnose training dynamics*

2. **Transformer** \rightarrow **Brain**: *Normalizing flow techniques (our Brenier approach) can be applied to both*

3. **Shared**: *Both benefit from understanding identifiability via optimal transport*

Table 5: Correspondence between brain model inversion and transformer training

Brain Model Inversion	Transformer Training
Observe BOLD/FC	Observe (input, output) pairs
Infer $\theta = (W, \tau, g, \dots)$	Learn weights (W_Q, W_K, W_V, \dots)
Likelihood $p(\text{FC} \theta)$	Loss $\mathcal{L}(\text{output}, \text{target})$
Prior $p(\theta)$	Regularization / architecture constraints
Posterior $p(\theta \text{FC})$	Trained weights
Non-identifiable manifold	Flat directions in loss landscape
Effective parameters	Parameters affecting output

4.5 Hierarchical Free Energy Structure

Both systems exhibit hierarchical timescale separation:

Definition 4.16 (Hierarchical Free Energies). 1. **Fast** ($\sim ms$): Gating variables / attention weights

$$F_{fast} = -\frac{1}{\beta_{fast}} \log Z_{fast} \quad (58)$$

2. **Medium** ($\sim 10-100 ms$): Firing rates / token embeddings

$$F_{medium} = \sum_r U_r(\nu_r) + \text{interaction terms} \quad (59)$$

3. **Slow** ($\sim 1-5 s$): Hemodynamics / layer aggregation

$$F_{slow} = \text{coarse-grained observable} \quad (60)$$

At each level, fast variables equilibrate conditional on slow variables, enabling adiabatic elimination and hierarchical inference.

5 Exemplary Application: Schizophrenia

5.1 Clinical Rationale

Schizophrenia is characterized by:

- Widespread functional dysconnectivity (Voineskos et al., 2024)
- Thalamocortical circuit abnormalities (Anticevic & Murray, 2016)
- Consistent computational signature: loss of pyramidal cell synaptic gain (Adams et al., 2022)

5.2 Hypotheses

H1 Schizophrenia patients show altered effective E/I ratio ($\theta_{E/I}^{\text{eff}}$) in thalamocortical circuits

H2 Effective parameters correlate with symptom dimensions (PANSS)

H3 Test-retest reliability of effective parameters exceeds ICC > 0.6

5.3 Conservative Pilot Study Design

5.3.1 Participants

- $N = 10$ schizophrenia patients (stable on atypical antipsychotics ≥ 4 weeks)
- $N = 10$ age/sex-matched healthy controls
- Exclusions: substance use disorder, intellectual disability, neurological conditions

Table 6: Validation strategy

Level	Method	Success Criterion	Timeline
Simulation	Parameter recovery	θ^* within 95% CI	Year 1
Reliability	Test-retest	ICC > 0.6	Year 2 H1
Pharmacological	GABA agonist probe	Expected $\uparrow \theta_{E/I}^{\text{eff}}$	Year 2 H2
Cross-modal	MRS correlation	Sig. corr. with Glx/GABA	Year 3

5.3.2 Assessments

- Positive and Negative Syndrome Scale (PANSS)
- Clinical Global Impression - Severity (CGI-S)
- Brief cognitive battery

5.3.3 Neuroimaging Protocol (3T)

- T1-MPRAGE: 1mm isotropic
- Multi-shell diffusion: $b = 1000/2000/3000$ s/mm², 64 directions/shell
- Resting-state fMRI: 15 min, eyes open, $T_R = 2$ s, 2.5 mm isotropic
- Test-retest: 5 controls scanned twice, 1–2 weeks apart

5.3.4 Analysis Pipeline

1. Construct subject-specific SC from diffusion MRI
2. Build whole-brain TVB model with generic neural masses
3. Apply Brenier-based inversion via normalizing flow (Section 3)
4. Extract effective parameters and entropy production rate Φ
5. Group comparisons via permutation tests
6. Symptom correlations via Spearman rank correlation
7. Test-retest reliability via intraclass correlation (ICC)

5.3.5 Power Analysis

With $N = 10$ per group, $\alpha = 0.05$, we have 80% power to detect large effect sizes ($d \geq 1.2$). This is appropriate for a pilot study aimed at feasibility and preliminary effect estimation.

5.4 Expected Outcomes

- Demonstration of effective parameter estimation from clinical fMRI
- Preliminary group differences in $\theta_{E/I}^{\text{eff}}$ and Φ
- Reliability estimates to inform larger studies

5.5 Limitations

- Small sample size (pilot)
- Medication effects not controlled
- Generic (not region-specific) neural masses initially
- Correlation \neq causation

Table 7: Three-year timeline

Period	Milestone	Deliverable
Year 1, H1	Literature review, TVB integration	Technical report
Year 1, H2	Brenier flow implementation, simulation validation	Software release
Year 2, H1	Test-retest reliability study	Methods paper
Year 2, H2	Pilot clinical data collection	Dataset
Year 3, H1	Clinical analysis, thermodynamic framework	Theory paper
Year 3, H2	Dissertation writing	Thesis

Table 8: Feasibility assessment

Component	Risk	Mitigation
Brenier flow implementation	Medium	POT/GeomLoss libraries; existing flow architectures
TVB integration	Low	Established platform; active community
Clinical recruitment	Medium	Existing collaborations; realistic N
Computational resources	Low	GPU cluster available

6 Validation Strategy

7 Timeline and Feasibility

8 Conclusion

We have proposed a principled framework for brain model inversion with the following key contributions:

1. **Brenier identifiability:** Using optimal transport theory, specifically Brenier’s uniqueness theorem and Knothe-Rosenblatt transport, we provide rigorous guarantees for parameter identifiability. The TMI structure of autoregressive normalizing flows ensures unique recovery of effective parameters.
2. **Thermodynamic grounding:** We establish that neural transfer functions are partition functions, entropy production rate (measurable from BOLD via MOU fitting) determines identifiable information content, and the forward model’s coarse-graining produces entropy equal to lost parameter information.
3. **Transformer isomorphism:** The mathematical correspondence between neural mean-field dynamics and transformer attention should be exact. Both implement gradient flows on energy landscapes with partition-function nonlinearities, and both exhibit entropy production from asymmetric interactions.
4. **Clinical applicability:** The conservative pilot design demonstrates feasibility while acknowledging limitations. If validated, the framework could contribute to precision psychiatry through mechanistically-interpretable, uncertainty-quantified biomarkers.

The framework is general: applicable to any TVB-compatible neural mass model and any clinical/cognitive question where circuit-level inference from neuroimaging is desired.

Acknowledgments

[To be added]

References

- Friston, K. J. (2011). Functional and effective connectivity: A review. *Brain Connectivity*, 1(1), 13–36.
- Razi, A., Kahan, J., Raine, G., Hanber, L., Rotshtein, P., Penny, W. D., & Friston, K. J. (2015). Construct validation of a DCM for resting state fMRI. *NeuroImage*, 106, 1–14.
- Adams, R. A., et al. (2022). Computational modeling of EEG and fMRI paradigms indicates a consistent loss of pyramidal cell synaptic gain in schizophrenia. *Biological Psychiatry*, 91(1), 59–72.
- Anticevic, A., & Murray, J. D. (2016). Toward understanding thalamocortical dysfunction in schizophrenia through computational models. *Schizophrenia Research*, 180, 70–77.
- Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4), 375–417.
- Buxton, R. B., Wong, E. C., & Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation. *Magnetic Resonance in Medicine*, 39(6), 855–864.
- Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48), 30055–30062.
- Daunizeau, J., David, O., & Stephan, K. E. (2011). Dynamic causal modelling: A critical review. *NeuroImage*, 58(2), 312–322.
- El Boustani, S., & Destexhe, A. (2009). A master equation formalism for macroscopic modeling. *Neural Computation*, 21(1), 46–100.
- Friston, K. J., Mechelli, A., Turner, R., & Price, C. J. (2000). Nonlinear responses in fMRI: The balloon model. *NeuroImage*, 12(4), 466–477.
- Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4), 1273–1302.
- Friston, K. J., et al. (2019). Dynamic causal modelling revisited. *NeuroImage*, 199, 730–744.
- Friston, K. J. (2022). Computational psychiatry: From synapses to sentience. *Molecular Psychiatry*, 28, 256–268.
- Gilson, M., Tagliazucchi, E., & Cofré, R. (2023). Entropy production of multivariate Ornstein-Uhlenbeck processes correlates with consciousness levels in the human brain. *Physical Review E*, 107(2), 024121.
- Godrèche, C., & Luck, J.-M. (2018). Characterising the nonequilibrium stationary states of Ornstein-Uhlenbeck processes. *Journal of Physics A*, 52, 035002.
- Hashemi, M., et al. (2020). The Bayesian Virtual Epileptic Patient. *NeuroImage*, 217, 116839.
- Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4), 500–544.
- Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge. *Nature Neuroscience*, 19(3), 404–413.
- Javaloy, A., Sánchez-Martín, P., & Valera, I. (2023). Causal normalizing flows: From theory to practice. *Advances in Neural Information Processing Systems*, 36.

- Jirsa, V. K., et al. (2023). Virtual brain twins: From basic neuroscience to clinical use. *The Lancet Psychiatry*, 10(12), 1003–1016.
- Khemakhem, I., Monti, R. P., Leech, R., & Hyvärinen, A. (2021). Causal autoregressive flows. *Proceedings of Machine Learning Research*, 130, 3520–3528.
- Lorenzi, R. M., et al. (2023). A multi-layer mean-field model of the cerebellum. *PLoS Computational Biology*, 19(9), e1011434.
- Lorenzi, R. M., et al. (2025). Region-specific mean field models enhance simulations of local and global brain dynamics. *npj Systems Biology and Applications*, 11, 66.
- Martín, D., et al. (2025). TVB C++: A fast and flexible back-end for The Virtual Brain. *Advanced Science*, 12, 2406440.
- McCann, R. J., & Guillen, N. (2011). Five lectures on optimal transportation. *Analysis and Geometry of Metric Measure Spaces*, 145–180.
- Ramsauer, H., et al. (2021). Hopfield networks is all you need. *International Conference on Learning Representations*.
- Ritter, P., Schirner, M., McIntosh, A. R., & Jirsa, V. K. (2013). The Virtual Brain integrates computational modeling and multimodal neuroimaging. *Brain Connectivity*, 3(2), 121–145.
- Sanz Leon, P., et al. (2013). The Virtual Brain: A simulator of primate brain network dynamics. *Frontiers in Neuroinformatics*, 7, 10.
- Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.
- Voineskos, A. N., et al. (2024). Functional magnetic resonance imaging in schizophrenia: Current evidence and future directions. *World Psychiatry*, 23(1), 26–51.
- Xi, Q., & Bloem-Reddy, B. (2023). Indeterminacy in generative models: Characterization and strong identifiability. *Proceedings of Machine Learning Research*, 206, 6912–6939.
- Zerlaut, Y., et al. (2018). Modeling mesoscopic cortical dynamics using a mean-field model. *Journal of Computational Neuroscience*, 44(1), 45–61.