Region-Specific Mean-Field Hodgkin-Huxley Models with Optimal Transport Model Inversion

A Framework for Computational Biomarkers in Psychiatric Neuroimaging

PhD Research Proposal

Marvin Koss

Universität Heidelberg

koss@cl.uni-heidelberg.de

January 22, 2026

Abstract

We propose a computational framework combining **region-specific mean-field Hodgkin-Huxley (MF-HH) models** with **optimal transport (OT) based model inversion** for analyzing functional MRI data in psychiatric populations. Unlike generic neural mass models used in The Virtual Brain (TVB) or phenomenological approaches like Dynamic Causal Modeling (DCM), region-specific MF-HH models preserve biophysically interpretable parameters—synaptic conductances, membrane time constants, and population-specific coupling strengths. We develop an OT-based inversion scheme that estimates *effective circuit parameters* from BOLD functional connectivity, with explicit treatment of the identifiability problem inherent to indirect neuroimaging measurements. We propose applications to vmPFC/OFC-basal ganglia circuits (relevant to depression, addiction) and temporoparietal junction networks (relevant to schizophrenia, autism). Rather than claiming to recover true biophysical parameters, we frame the approach as extracting *model-derived signatures* of circuit function that may differ systematically between patient and control populations, generating mechanistic hypotheses for validation with complementary methods.

Contents

1 Introduction

1.1 The Challenge of Mechanistic Neuroimaging

Functional MRI has transformed cognitive neuroscience by revealing which brain regions activate during various tasks. However, fMRI faces a fundamental limitation: the blood-oxygen-level-dependent (BOLD) signal is an indirect, temporally blurred measurement of neural activity. The signal chain from neural computation to measured BOLD involves multiple transformations:

$$\underbrace{\text{Synaptic activity}}_{\text{ms timescale}} \rightarrow \underbrace{\text{Firing rates}}_{\text{10-100 ms}} \rightarrow \underbrace{\text{Metabolic demand}}_{\text{100 ms - 1 s}} \rightarrow \underbrace{\text{Hemodynamics}}_{\text{1-5 s}} \rightarrow \underbrace{\text{BOLD}}_{\text{TR: 0.5-2 s}} \quad (1)$$

Each arrow represents information loss. The central question motivating this proposal is: *Can we extract mechanistically interpretable information about neural circuit function from BOLD data, and can such information reveal what differs in psychiatric disorders?*

1.2 Current Approaches and Their Limitations

Two dominant frameworks attempt to bridge neural dynamics and fMRI:

The Virtual Brain (TVB) (Sanz Leon et al., 2013) places neural mass models at each brain region, coupled via structural connectivity from diffusion MRI. However, TVB typically uses *generic* oscillator models (Wong-Wang, Wilson-Cowan) that are identical across all regions, ignoring the substantial heterogeneity in cortical microcircuitry.

Dynamic Causal Modeling (DCM) (Friston et al., 2003) estimates effective connectivity by inverting a generative model of BOLD. However, DCM uses phenomenological neural state equations without direct biophysical interpretation; parameters like "self-inhibition" do not map onto quantities measurable by other methods.

Recent work by Lorenzi et al. (2025) demonstrates that *region-specific* mean-field models—which preserve the local microcircuit architecture—achieve substantially better fits to empirical data (approximately 50% error reduction) compared to generic models. This motivates extending the region-specific approach to circuits implicated in psychiatric disorders.

1.3 Proposal Overview

This proposal develops:

(i) A **paradigmatic mean-field Hodgkin-Huxley model** that can be adapted to specific brain regions (Section 3)

(ii) An **optimal transport framework for model inversion** with explicit treatment of parameter identifiability (Section 4)

(iii) **Clinical applications** to depression, addiction, schizophrenia, and autism, with concrete study designs (Section 5)

We adopt an epistemically honest framing: we do not claim to recover true biophysical parameters from BOLD. Instead, we aim to extract *effective parameters* and *circuit signatures* that may systematically differ between populations and generate testable mechanistic hypotheses.

2 Related Work

2.1 The Virtual Brain Project

The Virtual Brain (TVB) (Sanz Leon et al., 2013; Ritter et al., 2013) is a neuroinformatics platform for simulating whole-brain network dynamics. The standard TVB pipeline:

1. Parcellates the brain into N regions (typically 68-400)
2. Derives structural connectivity matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ from diffusion MRI tractography
3. Places identical neural mass models at each node
4. Simulates coupled dynamics and generates synthetic BOLD via hemodynamic forward model
5. Fits global parameters (coupling strength, conduction velocity) to match empirical functional connectivity

Limitation: The one-model-fits-all approach ignores regional heterogeneity. Visual cortex, prefrontal cortex, and basal ganglia have fundamentally different microcircuit architectures, cell type compositions, and computational roles.

Lorenzi et al. (2025) addressed this by developing a *cerebellar-specific* mean-field model (CRBL MF) with four populations matching cerebellar anatomy: granule cells, Golgi cells, molecular layer interneurons, and Purkinje cells. When integrated into TVB, this region-specific model significantly outperformed generic alternatives.

2.2 Dynamic Causal Modeling

DCM (Friston et al., 2003, 2019) inverts a generative model:

$$\dot{\boldsymbol{z}} = f(\boldsymbol{z}, \mathbf{u}, \boldsymbol{\theta}), \qquad \boldsymbol{y} = g(\boldsymbol{z}) + \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{z}$ are hidden neural states, $\mathbf{u}$ are inputs, $\boldsymbol{\theta}$ are connectivity parameters, g is the hemodynamic forward model, and $\boldsymbol{y}$ is observed BOLD.

DCM strengths:

- Bayesian inference with uncertainty quantification
- Model comparison via free energy / Bayes factors
- Extensive validation literature

DCM limitations:

- Neural state equations are phenomenological (e.g., $\dot{z}_i = \sum_j A_{ij} z_j$)
- Parameters lack direct biophysical interpretation
- Typically applied to small networks (2-10 regions)
- Identifiability concerns with complex models (Daunizeau et al., 2011)

2.3 Mean-Field Neural Models

The mean-field approach derives macroscopic population equations from microscopic spiking network dynamics (Wilson & Cowan, 1972; El Boustani & Destexhe, 2009; Zerlaut et al., 2018). Key developments:

Transfer function formalism (El Boustani & Destexhe, 2009): For a population of spiking neurons receiving Poisson-distributed synaptic input, the stationary firing rate can be expressed as a function of input rates:

$$\nu_{out} = \mathcal{T}(\nu_{in,1}, \ldots, \nu_{in,K}; \boldsymbol{\theta}) \tag{3}$$

where $\mathcal{T}$ is the *transfer function* determined by single-neuron properties and network statistics.

AdEx mean-field (Zerlaut et al., 2018): Derived semi-analytic transfer functions for adaptive exponential integrate-and-fire neurons, enabling efficient simulation of large-scale networks.

Hodgkin-Huxley mean-field (Carlu et al., 2020): Extended mean-field techniques to conductance-based Hodgkin-Huxley neurons, preserving biophysical parameters (ionic conductances, reversal potentials).

2.4 Computational Psychiatry

Computational psychiatry applies formal models to understand psychiatric disorders (Huys et al., 2016; Friston et al., 2016). Relevant findings:

Depression: Disrupted reinforcement learning, particularly reduced reward prediction errors in ventral striatum and altered model-based/model-free arbitration (Park et al., 2021; Kumar et al., 2008).

Addiction: vmPFC/OFC-striatal hypoconnectivity, shift from goal-directed to habitual control (Ersche et al., 2020; Everitt & Robbins, 2016).

Schizophrenia: Theory of mind deficits associated with reduced TPJ activation, aberrant precision weighting in predictive coding frameworks (Lee et al., 2011; Adams et al., 2013).

Gap: These findings typically use behavioral models (e.g., Q-learning) regressed against BOLD, or phenomenological connectivity measures. A biophysically grounded circuit model could link behavioral-level disruptions to cellular-level mechanisms.

3 Mean-Field Hodgkin-Huxley Models

This section develops the mathematical framework for region-specific mean-field models, starting from single-neuron Hodgkin-Huxley dynamics and deriving population-level equations.

3.1 Single-Neuron Hodgkin-Huxley Dynamics

The Hodgkin-Huxley model (Hodgkin & Huxley, 1952) describes action potential generation through voltage-gated ion channels:

Definition 3.1 (Hodgkin-Huxley Equations)**.** The membrane potential $V(t)$ of a single neuron evolves according to:

$$C_m \frac{dV}{dt} = -I_{ion}(V, m, h, n) - I_{syn}(V, t) + I_{ext} \tag{4}$$

where C_m is membrane capacitance (typically $\sim 1\,\mu\text{F/cm}^2$), and the ionic current is:

$$I_{ion}(V, m, h, n) = g_L(V - E_L) + g_{Na} m^3 h (V - E_{Na}) + g_K n^4 (V - E_K) \tag{5}$$

The terms in (5) represent:

- **Leak current**: $g_L(V - E_L)$ with conductance g_L and reversal potential $E_L \approx -65$ mV
- **Sodium current**: $g_{Na} m^3 h(V - E_{Na})$ with $g_{Na} \approx 120$ mS/cm^2, $E_{Na} \approx +50$ mV
- **Potassium current**: $g_K n^4 (V - E_K)$ with $g_K \approx 36$ mS/cm^2, $E_K \approx -77$ mV

The gating variables $m, h, n \in [0, 1]$ follow first-order kinetics:

Definition 3.2 (Gating Variable Dynamics)**.** Each gating variable $x \in \{m, h, n\}$ satisfies:

$$\frac{dx}{dt} = \alpha_x(V)(1 - x) - \beta_x(V)x = \frac{x_\infty(V) - x}{\tau_x(V)} \tag{6}$$

where the voltage-dependent rate functions $\alpha_x(V), \beta_x(V)$ determine the steady-state activation $x_\infty(V) = \alpha_x/(\alpha_x + \beta_x)$ and time constant $\tau_x(V) = 1/(\alpha_x + \beta_x)$.

Example 3.1 (Standard HH Rate Functions)**.** For the original squid axon parameters:

$$\alpha_m(V) = \frac{0.1(V+40)}{1-e^{-(V+40)/10}}, \qquad \beta_m(V) = 4e^{-(V+65)/18} \tag{7}$$

$$\alpha_h(V) = 0.07e^{-(V+65)/20}, \qquad \beta_h(V) = \frac{1}{1+e^{-(V+35)/10}} \tag{8}$$

$$\alpha_n(V) = \frac{0.01(V+55)}{1-e^{-(V+55)/10}}, \qquad \beta_n(V) = 0.125e^{-(V+65)/80} \tag{9}$$

3.2 Synaptic Currents

Neurons receive synaptic input from other neurons. For a neuron receiving input from K presynaptic populations:

Definition 3.3 (Synaptic Current)**.** The total synaptic current is:

$$I_{syn}(V,t) = \sum_{s=1}^{K} g_s(t)(V - E_s) \tag{10}$$

where $g_s(t)$ is the time-varying conductance from synapse type s (e.g., AMPA, NMDA, GABA$_A$), and E_s is the corresponding reversal potential:

- Excitatory (AMPA/NMDA): $E_{exc} \approx 0$ mV
- Inhibitory (GABA$_A$): $E_{inh} \approx -80$ mV

The synaptic conductance evolves as:

$$\frac{dg_s}{dt} = -\frac{g_s}{\tau_s} + Q_s \sum_{j \in \text{pre}_s} \sum_k \delta(t - t_j^k) \tag{11}$$

where τ_s is the synaptic time constant (AMPA: $\sim$2 ms, GABA$_A$: $\sim$10 ms, NMDA: $\sim$100 ms), Q_s is the quantal conductance, and t_j^k are presynaptic spike times.

3.3 From Spiking Networks to Mean-Field Equations

Consider a network of N Hodgkin-Huxley neurons, potentially organized into P populations (e.g., pyramidal cells, fast-spiking interneurons). Direct simulation scales as $O(N^2)$ for all-to-all connectivity, becoming intractable for large networks.

The mean-field approach replaces individual neuron dynamics with population-averaged statistics under the following:

Assumption 3.1 (Asynchronous Irregular Regime)**.** In the large-N limit, neurons fire asynchronously with approximately Poisson statistics. Correlations between individual spike trains vanish: $\langle \delta\nu_i(t)\delta\nu_j(t')\rangle \to 0$ for $i \neq j$.

Under Assumption 3.1, the input to each neuron can be characterized by its mean and variance, determined by presynaptic firing rates.

Definition 3.4 (Population Firing Rate)**.** The instantaneous firing rate of population p is:

$$\nu_p(t) = \frac{1}{N_p} \sum_{i \in \text{pop}_p} \sum_k \delta(t - t_i^k) \tag{12}$$

In the mean-field limit, $\nu_p(t)$ becomes a deterministic function satisfying population-level dynamics.

3.4 Transfer Function Formalism

The key quantity linking microscopic and macroscopic descriptions is the *transfer function*:

Definition 3.5 (Transfer Function)**.** The transfer function $\mathcal{T}_p : \mathbb{R}^K \times \Theta \to \mathbb{R}_+$ maps input firing rates to output firing rate:

$$\nu_p = \mathcal{T}_p(\nu_1, \ldots, \nu_K; \boldsymbol{\theta}_p) \tag{13}$$

where $\{\nu_s\}_{s=1}^K$ are firing rates of afferent populations, and $\boldsymbol{\theta}_p$ contains biophysical parameters:

$$\boldsymbol{\theta}_p = (g_{Na}, g_K, g_L, E_L, \{Q_{s\to p}\}, \{K_{s\to p}\}, \{\tau_s\}, \ldots) \tag{14}$$

For Hodgkin-Huxley neurons, the transfer function can be computed semi-analytically (Carlu et al., 2020):

Proposition 3.1 (HH Transfer Function)**.** *Under Assumption 3.1, the stationary firing rate of an HH population receiving fluctuating synaptic input is approximately:*

$$\mathcal{T}_p(\boldsymbol{\nu}) \approx \frac{1}{2\tau_V} \operatorname{erfc}\left(\frac{V_{th}^{eff} - \mu_V}{\sqrt{2}\sigma_V}\right) \tag{15}$$

where:

- μ_V *is the mean membrane potential (subthreshold)*
- σ_V *is the standard deviation of membrane potential fluctuations*
- τ_V *is the effective membrane time constant*
- V_{th}^{eff} *is an effective threshold incorporating HH nonlinearities*

These quantities depend on input statistics and biophysical parameters through:

$$\mu_G = g_L + \sum_s K_{s\to p} Q_s \tau_s \nu_s \tag{16}$$

$$\mu_V = \frac{g_L E_L + \sum_s K_{s\to p} Q_s \tau_s \nu_s E_s}{\mu_G} \tag{17}$$

$$\sigma_V^2 = \sum_s \frac{K_{s\to p} Q_s^2 \tau_s \nu_s}{2\mu_G^2} \cdot \frac{(E_s - \mu_V)^2 \tau_s}{\tau_m^{eff} + \tau_s} \tag{18}$$

$$\tau_m^{eff} = C_m/\mu_G \tag{19}$$

where $K_{s\to p}$ is the number of synapses from population s to a neuron in population p.

Remark 3.1 (Biophysical Interpretability). Unlike phenomenological models, the transfer function (15) depends explicitly on measurable quantities: membrane capacitance C_m, ionic conductances g_X, synaptic parameters $Q_s, \tau_s, K_{s\to p}$. Changes in these parameters have interpretable effects on circuit function.

3.5 Multi-Population Mean-Field Dynamics

For a circuit with P interacting populations, the mean-field dynamics form a coupled system:

Definition 3.6 (Mean-Field Hodgkin-Huxley System)**.** The firing rates $\boldsymbol{\nu}(t) = (\nu_1(t), \ldots, \nu_P(t))^\top$ evolve according to:

$$\boxed{T_p \frac{d\nu_p}{dt} = \mathcal{T}_p(\boldsymbol{\nu}; \boldsymbol{\theta}) - \nu_p + \eta_p(t), \quad p = 1, \ldots, P} \tag{20}$$

where:

- T_p is the effective time constant of population p
- $\mathcal{T}_p$ is the transfer function (Definition 3.5)
- $\eta_p(t)$ is a noise term capturing finite-size fluctuations

Remark 3.2 (Relation to Wilson-Cowan). Equation (20) generalizes the Wilson-Cowan equations (Wilson & Cowan, 1972):

$$T_E \dot{r}_E = -r_E + S_E(w_{EE} r_E - w_{EI} r_I + I_E) \tag{21}$$

The key difference is that our transfer function $\mathcal{T}_p$ is derived from HH dynamics with explicit biophysical parameters, rather than being a phenomenological sigmoid S_E.

3.6 Paradigmatic Circuit Model

We now define a paradigmatic three-population circuit that can be adapted to specific brain regions:

Definition 3.7 (Paradigmatic E-I-M Circuit)**.** The paradigmatic circuit comprises:

- **E**: Excitatory pyramidal neurons
- **I**: Fast-spiking inhibitory interneurons (PV+)
- **M**: Modulatory population (region-specific: D1/D2 MSNs in striatum, SOM+ interneurons in cortex, etc.)

The connectivity matrix is:

$$\mathbf{W} = \begin{pmatrix} w_{EE} & -w_{EI} & w_{EM} \\ w_{IE} & -w_{II} & w_{IM} \\ w_{ME} & -w_{MI} & w_{MM} \end{pmatrix} \tag{22}$$

where $w_{XY} = K_{Y \to X} Q_{Y \to X} \tau_Y$ is the effective coupling strength.

Region-specific adaptations:

- **vmPFC/OFC**: M = layer-specific projection neurons; add dopaminergic modulation of Q_{exc}
- **Striatum**: Replace E/I with $D1$-MSN/$D2$-MSN; M = cholinergic interneurons
- **TPJ**: Standard E/I with M = SOM+ interneurons; long-range coupling to mPFC

3.7 Hemodynamic Forward Model

To connect neural dynamics to BOLD, we use the Balloon-Windkessel model (Buxton et al., 1998; Friston et al., 2000):

Definition 3.8 (Balloon-Windkessel Model)**.** The BOLD signal $y(t)$ is generated from neural activity $z(t) = \sum_p \nu_p(t)$ via:

$$\dot{s} = z - \kappa s - \gamma(f - 1) \tag{23}$$

$$\dot{f} = s \tag{24}$$

$$\tau_0 \dot{v} = f - v^{1/\alpha} \tag{25}$$

$$\tau_0 \dot{q} = \frac{f}{\rho}(1 - (1-\rho)^{1/f}) - \frac{q}{v} v^{1/\alpha} \tag{26}$$

$$y = V_0 \left(k_1(1-q) + k_2 \left(1 - \frac{q}{v}\right) + k_3(1-v) \right) \tag{27}$$

where s is vasodilatory signal, f is blood flow, v is blood volume, q is deoxyhemoglobin content, and $\{V_0, k_1, k_2, k_3, \kappa, \gamma, \tau_0, \alpha, \rho\}$ are hemodynamic parameters.

Limitation 3.1 (Hemodynamic Blur). The Balloon-Windkessel model acts as a low-pass filter with time constant ∼5-6 seconds. Neural dynamics faster than ∼0.1 Hz are substantially attenuated, fundamentally limiting what can be recovered from BOLD.

4 Optimal Transport Model Inversion

This section develops the mathematical framework for estimating model parameters from fMRI data using optimal transport, with explicit treatment of the identifiability problem.

4.1 The Inverse Problem

Definition 4.1 (Model Inversion Problem)**.** Given:

- Empirical BOLD time series $\boldsymbol{y}^{emp} \in \mathbb{R}^{T \times R}$ (T timepoints, R regions)
- Forward model $\mathcal{M} : \Theta \to \mathbb{R}^{T \times R}$ mapping parameters to simulated BOLD
- Prior distribution $\pi_0 \in \mathcal{P}(\Theta)$ over parameters

Find the posterior distribution $\pi^* \in \mathcal{P}(\Theta)$ over parameters consistent with the data.

The forward model $\mathcal{M}$ comprises:

$$\mathcal{M}(\boldsymbol{\theta}) = \text{HRF} \circ \text{MF-HH}(\boldsymbol{\theta}) \tag{28}$$

where MF-HH generates neural dynamics via (20) and HRF applies the hemodynamic forward model (23)-(27).

4.2 The Identifiability Problem

Before developing the inversion method, we must confront the fundamental identifiability challenge:

Limitation 4.1 (Parameter Degeneracy). Multiple parameter configurations can produce statistically indistinguishable BOLD dynamics:

$$\exists\, \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \in \Theta : \quad d(\mathcal{M}(\boldsymbol{\theta}_1), \mathcal{M}(\boldsymbol{\theta}_2)) < \epsilon \tag{29}$$

for any practical threshold ϵ determined by measurement noise and finite data.

Sources of degeneracy:

1. **Compensatory parameter changes**: Increased g_{exc} + increased g_{inh} can yield similar E-I balance
2. **Time constant trade-offs**: Faster dynamics + weaker coupling ≈ slower dynamics + stronger coupling (after hemodynamic filtering)
3. **HRF uncertainty**: Neural parameter changes can be absorbed by HRF parameter changes

Remark 4.1 (Implications for Interpretation). We cannot claim to recover "true" biophysical parameters. Instead, we estimate *effective parameters*—model-derived quantities that summarize circuit properties at the resolution accessible from BOLD. Differences between populations should be interpreted as differences in *effective circuit signatures*, not literal synaptic conductances.

4.3 Functional Connectivity as Observable

Rather than comparing raw BOLD time series, we compare *functional connectivity* (FC) dynamics, which capture inter-regional coordination:

Definition 4.2 (Sliding-Window Functional Connectivity)**.** For window size W and step size Δ, the FC matrix at window k is:

$$FC_{ij}^{(k)} = \text{corr}\left(y_i(t), y_j(t)\right)_{t\in[k\Delta, k\Delta+W]} \tag{30}$$

The sequence $\{FC^{(k)}\}_{k=1}^{N_w}$ captures FC dynamics over the session.

Definition 4.3 (FC Distribution)**.** We represent FC dynamics as an empirical probability measure over the space of correlation matrices:

$$\mu^{emp} = \frac{1}{N_w}\sum_{k=1}^{N_w} \delta_{FC^{(k),emp}} \in \mathcal{P}(\mathcal{S}_R) \tag{31}$$

where $\mathcal{S}_R$ is the space of $R \times R$ correlation matrices. Similarly, for simulated data: $\mu^{sim}(\boldsymbol{\theta}) = \frac{1}{N_w}\sum_{k=1}^{N_w} \delta_{FC^{(k),sim}(\boldsymbol{\theta})}$.

4.4 Optimal Transport Distance

We compare FC distributions using the Wasserstein distance, which provides a geometrically meaningful metric on probability measures.

Definition 4.4 (2-Wasserstein Distance)**.** For probability measures $\mu, \nu \in \mathcal{P}_2()$ on a metric space $(, d)$, the 2-Wasserstein distance is:

$$\mathcal{W}_2(\mu,\nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\times} d(x,y)^2 \, d\gamma(x,y)\right)^{1/2} \tag{32}$$

where $\Gamma(\mu,\nu)$ is the set of couplings (joint distributions) with marginals μ and ν.

Proposition 4.1 (Properties of $\mathcal{W}_2$)**.** *The Wasserstein distance satisfies:*

1. ***Metric**: $\mathcal{W}_2(\mu,\nu) = 0 \Leftrightarrow \mu = \nu$; triangle inequality holds*
2. ***Geometry-aware**: Accounts for the metric structure of , not just overlap*
3. ***Interpolation**: Geodesics in $(\mathcal{P}_2(), \mathcal{W}_2)$ correspond to optimal mass transport*

For FC matrices, we use the Frobenius metric on $\mathcal{S}_R$:

$$d(FC_1, FC_2) = \|FC_1 - FC_2\|_F = \sqrt{\sum_{i,j}(FC_{1,ij} - FC_{2,ij})^2} \tag{33}$$

Why Wasserstein over alternatives?

- **vs. KL divergence**: $\mathcal{W}_2$ is defined even when supports don't overlap; KL diverges
- **vs. point-wise MSE**: $\mathcal{W}_2$ captures distributional shape, not just marginal statistics
- **vs. KS statistic**: $\mathcal{W}_2$ uses full geometry, not just maximum deviation

4.5 Entropic Regularization and Sinkhorn Algorithm

The optimal transport problem (32) is computationally expensive ($O(n^3 \log n)$ for n samples). Entropic regularization enables efficient approximation:

Definition 4.5 (Entropic OT)**.** The entropically regularized transport cost is:

$$\mathcal{W}_{2,\varepsilon}^2(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \left\{ \int d(x,y)^2 \, d\gamma + \varepsilon \mathrm{KL}(\gamma \| \mu\otimes\nu) \right\} \tag{34}$$

where $\varepsilon > 0$ is the regularization strength and KL is Kullback-Leibler divergence.

Proposition 4.2 (Sinkhorn Algorithm (Cuturi, 2013))**.** *For discrete measures $\mu = \sum_i a_i \delta_{x_i}$, $\nu = \sum_j b_j \delta_{y_j}$, the solution to* (34) *is:*

$$\gamma_{ij}^* = u_i K_{ij} v_j, \quad K_{ij} = \exp(-d(x_i,y_j)^2/\varepsilon) \tag{35}$$

where (u,v) are found by alternating projections:

$$u \leftarrow a \oslash (Kv), \qquad v \leftarrow b \oslash (K^\top u) \tag{36}$$

Convergence is linear with rate depending on ε.

4.6 Parameter-Space Optimization via Wasserstein Gradient Flow

We now formulate parameter estimation as optimization in the space of probability distributions over parameters.

Definition 4.6 (Inversion Objective)**.** The parameter posterior minimizes:

$$\mathcal{F}(\pi) = \mathbb{E}_{\boldsymbol{\theta}\sim\pi}\left[\mathcal{W}_{2,\varepsilon}^2(\mu^{emp}, \mu^{sim}(\boldsymbol{\theta}))\right] + \lambda \mathrm{KL}(\pi\|\pi_0) \tag{37}$$

where $\pi \in \mathcal{P}(\Theta)$ is a distribution over parameters, π_0 is the prior, and $\lambda > 0$ controls regularization.

Interpretation:

- First term: Expected Wasserstein distance between empirical and simulated FC (data fit)
- Second term: KL divergence from prior (regularization toward physiologically plausible parameters)

Definition 4.7 (Wasserstein Gradient Flow in Parameter Space)**.** The optimization proceeds via gradient flow in $(\mathcal{P}_2(\Theta), \mathcal{W}_2)$:

$$\frac{\partial \pi_t}{\partial t} = \nabla \cdot \left(\pi_t \nabla \frac{\delta \mathcal{F}}{\delta \pi}(\pi_t) \right) \tag{38}$$

where $\frac{\delta \mathcal{F}}{\delta \pi}$ is the first variation (functional derivative) of $\mathcal{F}$.

Theorem 4.1 (JKO Discretization (Jordan et al., 1998))**.** *The gradient flow* (38) *can be discretized via the proximal scheme:*

$$\boxed{\pi_{k+1} = \arg\min_{\pi\in\mathcal{P}_2(\Theta)} \left\{ \mathcal{F}(\pi) + \frac{1}{2\tau}\mathcal{W}_2^2(\pi,\pi_k) \right\}} \tag{39}$$

As $\tau \to 0$ and $k \to \infty$, iterates converge to a minimizer of $\mathcal{F}$ (under appropriate convexity conditions).

Remark 4.2 (Connection to Bayesian Inference). The JKO scheme (39) can be viewed as a geometric formulation of Bayesian inference:

- $\mathcal{F}(\pi)$ plays the role of negative log-posterior
- $\mathcal{W}_2^2(\pi,\pi_k)$ provides inertia preventing drastic updates
- The Wasserstein geometry respects the manifold structure of Θ

4.7 Computational Implementation

For practical computation, we use a particle approximation:

Algorithm 1 Wasserstein Gradient Flow Model Inversion

Require: Empirical BOLD $\boldsymbol{y}^{emp}$, prior π_0, particles J, iterations K, step size τ
Ensure: Particle approximation to posterior $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{J}$

Initialize particles $\{\boldsymbol{\theta}_0^{(j)}\}_{j=1}^{J} \sim \pi_0$
Compute μ^{emp} from sliding-window FC of $\boldsymbol{y}^{emp}$
for $k = 0, \ldots, K-1$ **do**
  for $j = 1, \ldots, J$ **do** ▷ Parallelizable
    Simulate: $\boldsymbol{y}^{sim,(j)} = \mathcal{M}(\boldsymbol{\theta}_k^{(j)})$
    Compute: $\mu^{sim,(j)}$ from sliding-window FC
    Evaluate: $L_j = \mathcal{W}_{2,\varepsilon}^2(\mu^{emp}, \mu^{sim,(j)})$
    Gradient: $\nabla_\theta L_j$ via autodiff through Sinkhorn
  end for
  Update particles via Stein Variational Gradient Descent (Liu & Wang, 2016):
    $\boldsymbol{\theta}_{k+1}^{(j)} = \boldsymbol{\theta}_k^{(j)} + \tau \phi^*(\boldsymbol{\theta}_k^{(j)})$
    where $\phi^* = \arg\max_{\|\phi\| \leq 1} \{-\mathbb{E}[\langle \nabla_\theta L + \lambda \nabla \log \pi_0, \phi \rangle + \nabla \cdot \phi]\}$
end for
return $\{\boldsymbol{\theta}_K^{(j)}\}_{j=1}^{J}$

4.8 Identifiability-Aware Inference

Given the degeneracy problem (Limitation 4.1), we extract *identifiable combinations* rather than individual parameters:

Definition 4.8 (Effective Parameters)**.** We define effective parameters as functions of biophysical parameters that are (approximately) identifiable from FC:

$$\theta_{E/I}^{eff} = \frac{w_{EE} - w_{IE}}{w_{EI} + w_{II}} \quad \text{(effective E-I ratio)} \tag{40}$$

$$\theta_{coup}^{eff} = \sum_{X,Y} |w_{XY}| \quad \text{(total coupling strength)} \tag{41}$$

$$\theta_{\tau}^{eff} = \frac{\sum_p T_p \nu_p^{ss}}{\sum_p \nu_p^{ss}} \quad \text{(activity-weighted time constant)} \tag{42}$$

Remark 4.3 (Validation Strategy). The identifiability of effective parameters should be validated via:

1. **Simulation recovery**: Generate synthetic data with known parameters; assess recovery of effective combinations
2. **Test-retest reliability**: Estimate parameters from same subject on two occasions; assess correlation
3. **Pharmacological challenge**: Administer drug with known mechanism; assess expected parameter shift

4.9 Comparison with DCM

5 Clinical Applications

This section details the clinical relevance of the proposed framework, specifying target circuits, associated disorders, and concrete study designs.

Table 1: Comparison of Inversion Approaches

Aspect	**DCM**	**TVB + Grid Search**	**Proposed (OT-MF-HH)**
Neural model	Phenomenological	Generic oscillators	Region-specific MF-HH
Parameters	Effective connectivity	Global coupling, delay	Biophysical (conductances, etc.)
Loss function	Free energy (Gaussian)	MSE, KS on FCD	Wasserstein on FC
Optimization	Variational Bayes	Grid search, BO	Wasserstein gradient flow
Uncertainty	Laplace approximation	None / bootstrap	Particle posterior
Scalability	Small networks (2-10)	Whole-brain	Intermediate (circuits)

5.1 Overview: Circuit-Disorder Mapping

Table 2: Target Circuits and Associated Disorders

Circuit		**Computational Role**	**Disorders**	**Key Predictions**
vmPFC/OFC → ventral striatum	→	Value encoding, reward prediction	MDD, Addiction, Suicidality	$\downarrow \theta^{eff}_{E/I}$, $\downarrow \theta^{eff}_{coup}$
Dorsal striatum motor ctx	↔	Action selection, habit	Addiction, OCD, Parkinson's	Altered D1/D2 balance
rTPJ ↔ mPFC		Belief updating, mentalizing	Schizophrenia, Autism	$\uparrow \theta^{eff}_{\tau}$ (slow updating)
Amygdala vmPFC	↔	Emotion regulation	Anxiety, PTSD	$\downarrow$ inhibitory coupling

5.2 Application 1: vmPFC/OFC-Striatal Circuit in Depression

5.2.1 Rationale

Major Depressive Disorder (MDD) is characterized by anhedonia—reduced capacity to experience pleasure from rewards. Computational studies consistently find:

- Reduced reward prediction error signals in ventral striatum (Kumar et al., 2008)
- Blunted value signals in vmPFC (Pizzagalli, 2014)
- Disrupted model-based/model-free arbitration (Park et al., 2021)

5.2.2 Circuit Model

5.2.3 Hypotheses

H1.a: MDD patients will show reduced effective E-I ratio ($\theta^{eff}_{E/I}$) in vmPFC, reflecting diminished excitatory drive or enhanced inhibition

H1.b: MDD patients will show reduced vmPFC→striatum coupling (θ^{eff}_{coup})

H1.c: These effective parameters will correlate with anhedonia severity (SHAPS scores) and predict treatment response

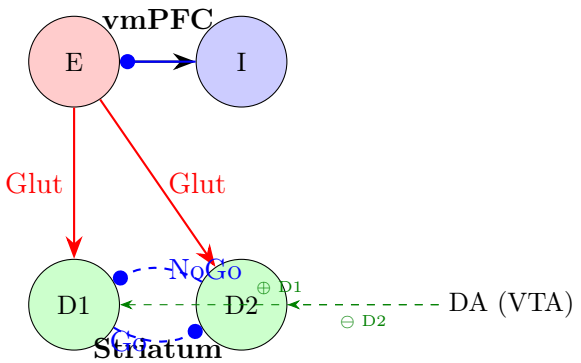

Figure 1: Schematic of vmPFC-striatal circuit model. E: excitatory pyramidal; I: fast-spiking interneuron; D1/D2: medium spiny neurons expressing D1/D2 dopamine receptors. Dopamine (DA) from VTA differentially modulates D1 (excitation) and D2 (inhibition) pathways.

5.2.4 Study Design

Study 1: Depression Protocol

Participants:

- $N = 50$ MDD patients (unmedicated or 2-week washout)
- $N = 50$ age/sex-matched healthy controls
- Exclusion: comorbid substance use, psychosis, neurological conditions

Assessments:

- Clinical: HDRS-17, BDI-II, SHAPS (anhedonia), DARS (anticipatory/consummatory)
- Cognitive: Probabilistic reversal learning task (behavioral RL parameters)

Neuroimaging:

- Structural: T1-MPRAGE, DWI (64 directions)
- Functional: Resting-state fMRI (15 min, eyes open, TR=1s), Task fMRI during reversal learning

Analysis:

1. Fit behavioral RL model to task data → individual learning rates α^+, α^-
2. Construct subject-specific vmPFC-striatum MF-HH model
3. Invert from resting-state FC using OT framework
4. Compare effective parameters between groups
5. Correlate with symptom dimensions and behavioral parameters

Power: With $N = 50$/group and $\alpha = 0.05$, 80% power to detect medium effect size ($d = 0.57$).

5.3 Application 2: TPJ Network in Schizophrenia

5.3.1 Rationale

Schizophrenia involves profound social cognitive deficits, including impaired theory of mind (ToM). Neuroimaging findings:

- Reduced right TPJ activation during false belief tasks (Lee et al., 2011)
- Sparser effective connectivity between TPJ and dmPFC (Bitsch et al., 2021)
- Computational accounts frame delusions as aberrant belief updating with disrupted precision weighting (Adams et al., 2013)

5.3.2 Circuit Model

The TPJ implements recursive Bayesian inference over others' mental states:

$$B_{t+1}(\theta) \propto p(o_t|\theta) \cdot B_t(\theta) \tag{43}$$

where B_t is the belief distribution over hidden states θ (e.g., another's intention), and o_t is observed behavior.

We model TPJ as an E-I circuit where:

- **Firing rate variability** encodes belief uncertainty (precision)
- **Time constant** θ_τ^{eff} reflects belief update speed
- **TPJ-mPFC coupling** integrates self-referential processing

5.3.3 Hypotheses

H2.a: Schizophrenia patients will show increased effective time constant (θ_τ^{eff}) in TPJ, reflecting sluggish belief updating

H2.b: Patients will show reduced TPJ→mPFC coupling

H2.c: Positive symptoms (PANSS-P) will correlate with altered precision encoding; negative symptoms (PANSS-N) with reduced coupling

5.3.4 Study Design

Study 2: Schizophrenia Protocol

Participants:

- $N = 40$ schizophrenia patients (stable on atypical antipsychotics ≥ 4 weeks)
- $N = 40$ age/sex/education-matched healthy controls
- Exclusion: substance use disorder, intellectual disability, ECT in past 6 months

Assessments:

- Clinical: PANSS (positive, negative, general), BNSS, CGI-S
- Social cognition: False belief task, Hinting Task, Reading the Mind in the Eyes

Neuroimaging:

- Resting-state fMRI (15 min, TR=1s)
- Task fMRI: False belief/false photograph paradigm (Saxe & Kanwisher, 2003)

Analysis:

1. Construct TPJ-mPFC-precuneus MF-HH model
2. Invert from resting-state + task-based FC
3. Compare effective parameters between groups
4. Correlate with PANSS subscales and ToM performance

5.4 Application 3: Transdiagnostic Dimensional Analysis

5.4.1 Rationale

The Research Domain Criteria (RDoC) framework emphasizes dimensional rather than categorical approaches to psychopathology. Circuit-level parameters may cut across diagnostic boundaries:

- Reward circuit dysfunction spans MDD, addiction, and schizophrenia negative symptoms
- Social cognition deficits span schizophrenia and autism

5.4.2 Study Design

Study 3: Transdiagnostic Protocol

Participants:

- $N = 30$ MDD
- $N = 30$ Schizophrenia
- $N = 30$ Autism Spectrum Disorder (without intellectual disability)
- $N = 30$ Healthy controls

Shared assessments across groups:

- Reward sensitivity: SHAPS, Effort Expenditure for Rewards Task
- Social cognition: RMET, Hinting Task
- General: PHQ-9, GAD-7

Analysis:

1. Estimate effective parameters for both circuits (vmPFC-striatum, TPJ-mPFC) in all participants
2. Cluster participants by parameter profiles (unsupervised)
3. Compare clustering to diagnostic categories
4. Test whether parameter-based clusters better predict functional outcomes

5.5 Validation Strategy

Table 3: Validation Experiments

Validation Type	**Method**	**Expected Outcome**
Simulation recovery	Generate synthetic data with known $\boldsymbol{\theta}^*$; invert; compare	Recovery of effective parameters within posterior credible intervals
Test-retest reliability	Same subjects, two sessions (1-2 weeks apart)	ICC > 0.7 for effective parameters
Pharmacological probe	Lorazepam (GABA agonist) in healthy volunteers	Increased $\theta^{eff}_{E/I}$ (more inhibition)
Cross-modal validation	Correlate with MRS glutamate/GABA in same subjects	Significant correlation between $\theta^{eff}_{E/I}$ and Glx/GABA ratio

6 Novelty, Limitations, and Feasibility

6.1 Novel Contributions

N1: Region-specific MF-HH for psychiatric circuits: First application of biophysically grounded mean-field models (beyond cerebellar/sensory regions) to circuits implicated in psychiatric disorders.

N2: Optimal transport model inversion: First use of Wasserstein distance and JKO scheme for neural model parameter estimation from fMRI. Provides principled geometry-aware loss function and uncertainty quantification.

N3: Identifiability-aware inference: Explicit treatment of the degeneracy problem; focus on effective parameters rather than claiming biophysical ground truth.

N4: Circuit-level biomarkers: If validated, effective parameters could serve as quantitative biomarkers with mechanistic interpretation, unlike purely statistical connectivity measures.

6.2 Limitations and Risks

Limitation 6.1 (Fundamental Limits of BOLD). BOLD provides indirect, temporally blurred access to neural activity. Even with optimal methods, some parameters will remain unidentifiable. We mitigate this by focusing on effective parameters and validating with complementary modalities.

Limitation 6.2 (Model Misspecification). The brain does not literally implement MF-HH equations. If the model class is wrong, fitted parameters may not correspond to any biological quantity. We mitigate this by: (a) using models grounded in neurophysiology; (b) testing multiple model variants; (c) validating predictions with independent methods.

Limitation 6.3 (Clinical Heterogeneity). Psychiatric disorders are heterogeneous. Group-level effects may be diluted by subgroups with different pathophysiology. We mitigate this by: (a) dimensional analyses alongside categorical; (b) clustering by parameter profiles; (c) adequate sample sizes.

6.3 Feasibility Assessment

Table 4: Feasibility Analysis

Component	**Risk Level**	**Justification / Mitigation**
MF-HH implementation	Low	Established theory; successful precedents (Lorenzi et al., Carlu et al.)
OT inversion	Medium	POT/GeomLoss libraries mature; Sinkhorn efficient; requires careful tuning of ε
Clinical recruitment (MDD)	Low	High prevalence; existing clinic collaborations
Clinical recruitment (SZ)	Medium	Requires specialized clinic; medication confounds
Clinical recruitment (ASD)	Medium	Requires collaboration; heterogeneous population
Computational resources	Low	GPU cluster available; parallelization straightforward

6.4 Timeline

Table 5: Proposed 3-Year Timeline

Period	Milestones
Year 1, H1	Literature review; MF-HH implementation for paradigmatic circuit; simulation studies
Year 1, H2	OT inversion algorithm development; validation on synthetic data; test-retest study
Year 2, H1	Study 1 (Depression) data collection and analysis
Year 2, H2	Study 2 (Schizophrenia) data collection; TPJ model development
Year 3, H1	Study 2 analysis; Study 3 (Transdiagnostic) data collection
Year 3, H2	Transdiagnostic analysis; dissertation writing; publications

7 Conclusion

This proposal develops a framework for extracting mechanistically interpretable information about neural circuit function from fMRI in psychiatric populations. By combining region-specific mean-field Hodgkin-Huxley models with optimal transport-based inversion, we aim to estimate *effective circuit parameters* that may systematically differ between patients and controls.

We adopt an epistemically honest position: we do not claim to recover true synaptic conductances from BOLD. Instead, we seek model-derived signatures of circuit function that can generate testable hypotheses about pathophysiology. The approach offers several advantages over existing methods:

- More biophysically grounded than DCM
- More region-specific than TVB with generic oscillators
- Geometry-aware loss function (Wasserstein) with principled uncertainty quantification
- Explicit treatment of identifiability limitations

If validated, this framework could contribute to precision psychiatry by providing quantitative, mechanistically interpretable biomarkers linking neural circuit function to clinical phenotypes.

References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4, 47.

Bitsch, F., Berger, P., Nagels, A., Falkenberg, I., & Straube, B. (2021). Characterizing the theory of mind network in schizophrenia reveals a sparser network structure. *Schizophrenia Research*, 228, 581–589.

Buxton, R. B., Wong, E. C., & Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magnetic Resonance in Medicine*, 39(6), 855–864.

Carlu, M., et al. (2020). A mean-field approach to the dynamics of networks of complex neurons, from nonlinear integrate-and-fire to Hodgkin-Huxley models. *Journal of Neurophysiology*, 123(3), 1042–1051.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2292–2300.

Daunizeau, J., David, O., & Stephan, K. E. (2011). Dynamic causal modelling: A critical review of the biophysical and statistical foundations. *NeuroImage*, 58(2), 312–322.

El Boustani, S., & Destexhe, A. (2009). A master equation formalism for macroscopic modeling of asynchronous irregular activity states. *Neural Computation*, 21(1), 46–100.

Ersche, K. D., et al. (2020). Brain networks underlying vulnerability and resilience to drug addiction. *Proceedings of the National Academy of Sciences*, 117(26), 15253–15261.

Everitt, B. J., & Robbins, T. W. (2016). Drug addiction: Updating actions to habits to compulsions ten years on. *Annual Review of Psychology*, 67, 23–50.

Friston, K. J., Mechelli, A., Turner, R., & Price, C. J. (2000). Nonlinear responses in fMRI: The balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, 12(4), 466–477.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4), 1273–1302.

Friston, K. J., Redish, A. D., & Gordon, J. A. (2016). Computational nosology and precision psychiatry. *Computational Psychiatry*, 1, 2–23.

Friston, K. J., et al. (2019). Dynamic causal modelling revisited. *NeuroImage*, 199, 730–744.

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 from neuroscience to clinical applications. *Nature Neuroscience*, 19(3), 404–413.

Jordan, R., Kinderlehrer, D., & Otto, F. (1998). The variational formulation of the Fokker-Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1), 1–17.

Kumar, P., et al. (2008). Abnormal temporal difference reward-learning signals in major depression. *Brain*, 131(8), 2084–2093.

Lee, K. H., et al. (2011). A functional magnetic resonance imaging study of social cognition in schizophrenia during an acute episode and after recovery. *American Journal of Psychiatry*, 163(11), 1926–1933.

Liu, Q., & Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2378–2386.

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.

Park, H., Lee, D., & Chey, J. (2021). Effects of subclinical depression on prefrontal-striatal model-based and model-free learning. *PLoS Computational Biology*, 17(5), e1008882.

Pizzagalli, D. A. (2014). Depression, stress, and anhedonia: Toward a synthesis and integrated model. *Annual Review of Clinical Psychology*, 10, 393–423.

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.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19(4), 1835–1842.

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 12(1), 1–24.

Zerlaut, Y., et al. (2018). Modeling mesoscopic cortical dynamics using a mean-field model of conductance-based networks of adaptive exponential integrate-and-fire neurons. *Journal of Computational Neuroscience*, 44(1), 45–61.