Neural subspaces, minimax entropy, and mean–field theory for networks of neurons

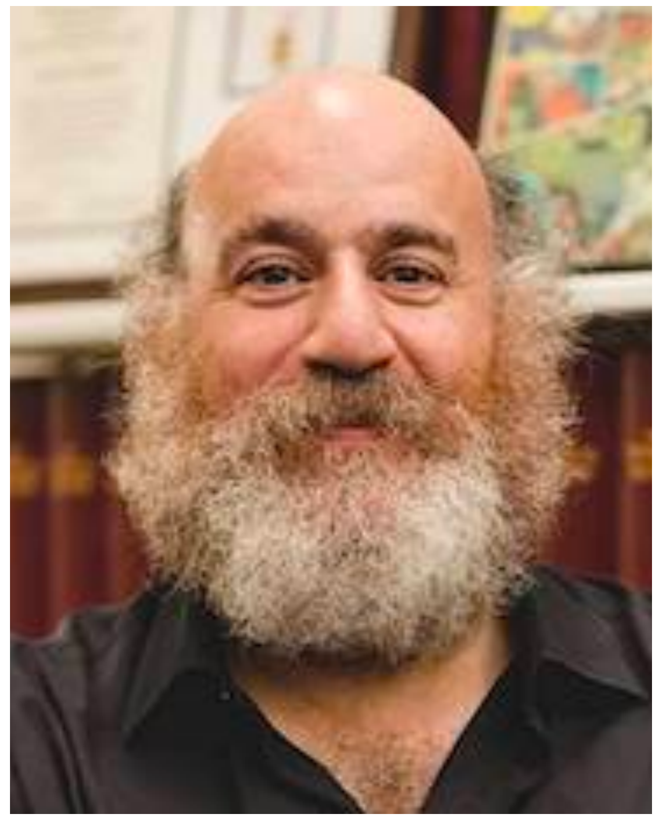
Francesca Mignacco
Princeton University & CUNY Graduate Center

A joint work with :

arXiv:2504.15197
\+ check arXiv today for a longer version: **arXiv:2508.02633**

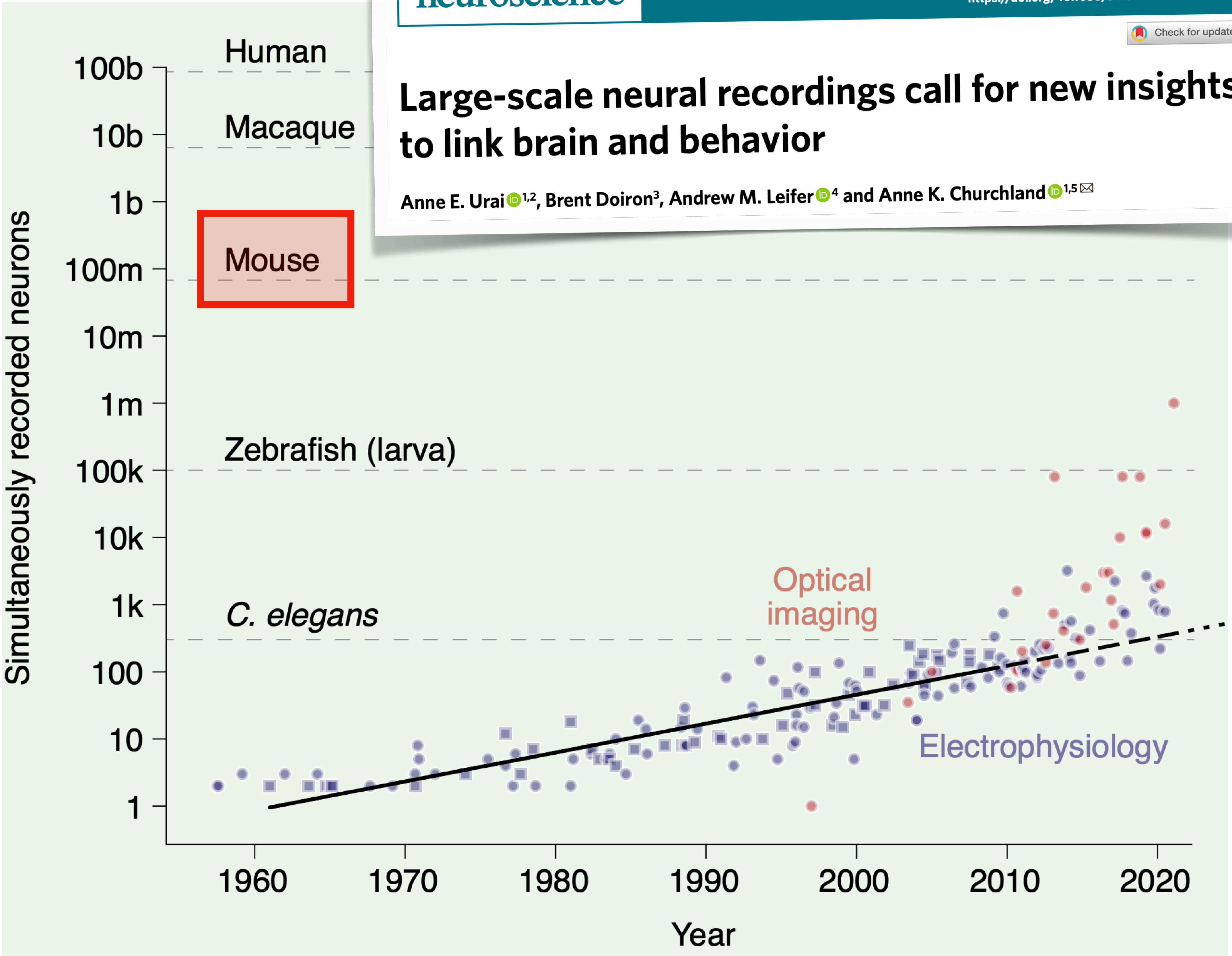
Luca Di Carlo
Princeton University

Christopher Lynn
Yale University



William Bialek
Princeton University

Large-scale neural recordings



nature neuroscience REVIEW ARTICLE
https://doi.org/10.1038/s41593-021-00980-9
Check for updates

Large-scale neural recordings call for new insights to link brain and behavior

Anne E. Urai[1,2], Brent Doiron[3], Andrew M. Leifer[4] and Anne K. Churchland[1,5]

Calcium imaging (10^6 neurons)

nature methods ARTICLES
https://doi.org/10.1038/s41592-021-01239-8
Check for updates

High-speed, cortex-wide volumetric recording of neuroactivity at cellular resolution using light beads microscopy

Jeffrey Demas[1,2], Jason Manley[1,2], Frank Tejera[1], Kevin Barber[1], Hyewon Kim[1], Francisca Martínez Traub[1], Brandon Chen[1] and Alipasha Vaziri[1,2]

Electrophysiology (10^4 neurons)

NEUROSCIENCE

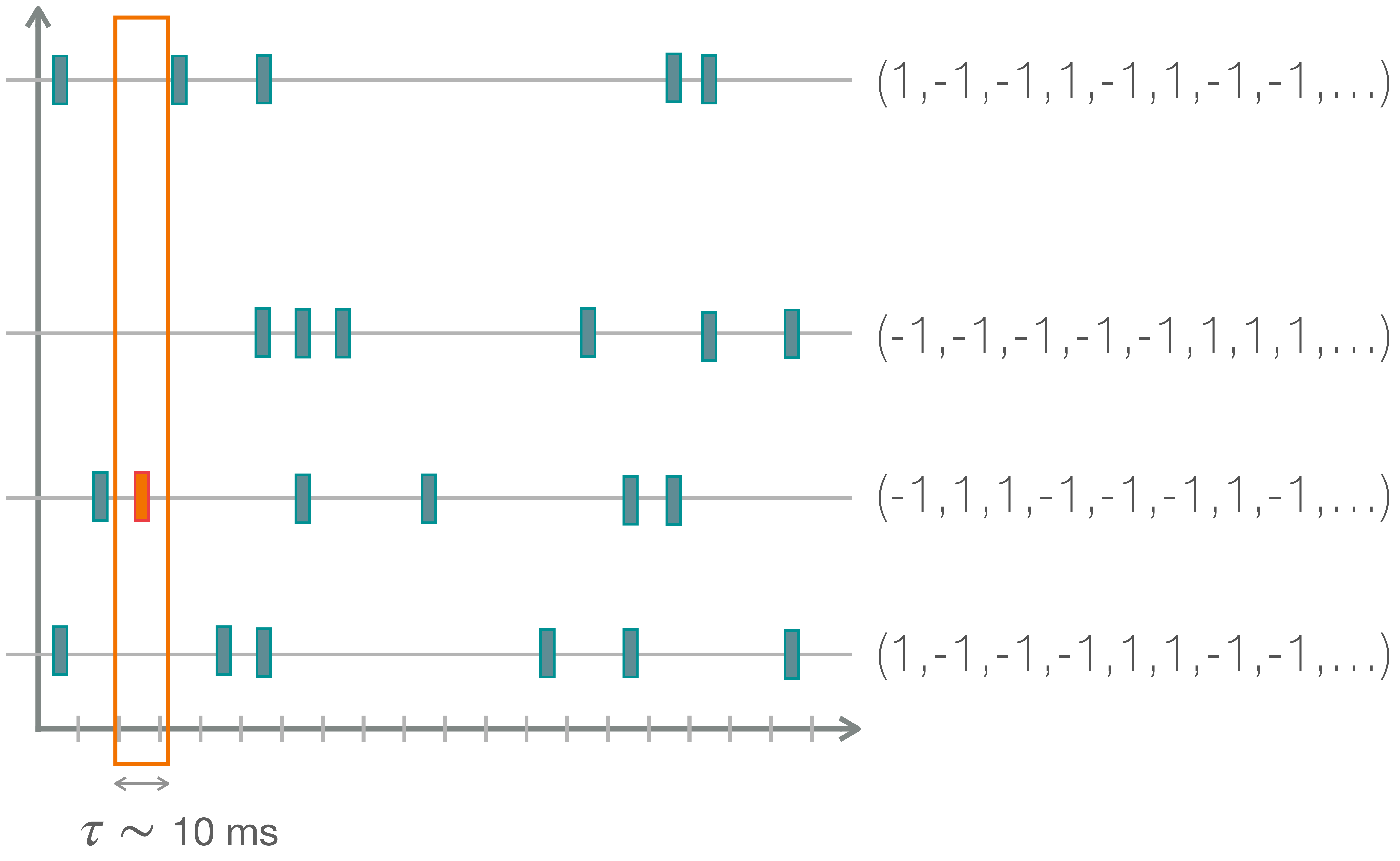
Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings

Nicholas A. Steinmetz*†, Cagatay Aydin†, Anna Lebedeva†, Michael Okun†, Marius Pachitariu†, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, Susu Chen, Jennifer Colonell, Richard J. Gardner, Bill Karsh, Fabian Kloosterman, Dimitar Kostadinov, Carolina Mora-Lopez, John O'Callaghan, Junchol Park, Jan Putzeys, Britton Sauerbrei, Rik J. J. van Daal, Abraham Z. Vollan, Shiwei Wang, Marleen Welkenhuysen, Zhiwen Ye, Joshua T. Dudman, Barundeb Dutta, Adam W. Hantman, Kenneth D. Harris, Albert K. Lee, Edvard I. Moser, John O'Keefe, Alfonso Renart, Karel Svoboda, Michael Häusser, Sebastian Haesler, Matteo Carandini*, Timothy D. Harris*

What physical principles underlie efficient neural computation?

How can macroscopic functions arise from microscopic neural interactions?

Large-scale neural recordings



At each time:

High-dimensional, sparse binary activity

$$\underline{\sigma} \in \{-1,1\}^N$$

$$\{\, \underline{\sigma}^{(1)}, \underline{\sigma}^{(2)}, \underline{\sigma}^{(3)}, \dots \}$$

Maximum entropy models

Consider some **microscopic measurements**

- Mean activities: $\frac{1}{T}\sum_{t=1}^{T}\bar{\sigma}_i^t = \mu_i^{\text{exp}}$ **(1)**
- Pairwise correlations: $\frac{1}{T-1}\sum_{t=1}^{T}(\bar{\sigma}_i^t - \mu_i^{\text{exp}})(\bar{\sigma}_j^t - \mu_j^{\text{exp}}) = C_{ij}^{\text{exp}}$ **(2)**

Can we predict **macroscopic features** ?

- State probability $P(\underline{\sigma})$
- Structure of interactions

Maximum entropy model: Search for the **least structured model** that matches these statistics

i.e., **maximize the entropy:** $S = -\left\langle \ln P(\underline{\sigma}) \right\rangle$ subject to the constraints **(1)** and **(2)**

Maximum entropy models

Consider some **microscopic measurements**

- Mean activities: $$\frac{1}{T}\sum_{t=1}^{T}\bar{\sigma}_i^t = \mu_i^{\text{exp}}$$
- Pairwise correlations: $$\frac{1}{T-1}\sum_{t=1}^{T}(\bar{\sigma}_i^t - \mu_i^{\text{exp}})(\bar{\sigma}_j^t - \mu_j^{\text{exp}}) = C_{ij}^{\text{exp}}$$

Can we predict **macroscopic features** ?

- State probability $P(\underline{\sigma})$
- Structure of interactions

Maximum entropy model:

[Martignon et al. (2000); Schneidman et al. (2006); Meshulam et al. (2017, 2024) ; …]

$$P(\underline{\sigma}) = \frac{1}{Z(J,h)}\exp\left(\frac{1}{2}\sum_{i\leq j} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{N} h_i\sigma_i\right)$$

(non-parametric model)

Maximum entropy models

Consider some **microscopic measurements**

- Mean activities: $$\frac{1}{T}\sum_{t=1}^{T}\bar{\sigma}_i^t = \mu_i^{\mathrm{exp}}$$
- Pairwise correlations: $$\frac{1}{T-1}\sum_{t=1}^{T}(\bar{\sigma}_i^t - \mu_i^{\mathrm{exp}})(\bar{\sigma}_j^t - \mu_j^{\mathrm{exp}}) = C_{ij}^{\mathrm{exp}}$$

Can we predict **macroscopic features** ?

- State probability $P(\underline{\sigma})$
- Structure of interactions

Maximum entropy model:

[Martignon et al. (2000); Schneidman et al. (2006); Meshulam et al. (2017, 2024) ; …]

$$P(\underline{\sigma}) = \frac{1}{Z(J,h)} \exp\left(\frac{1}{2}\sum_{i\le j} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{N} h_i\sigma_i\right)$$

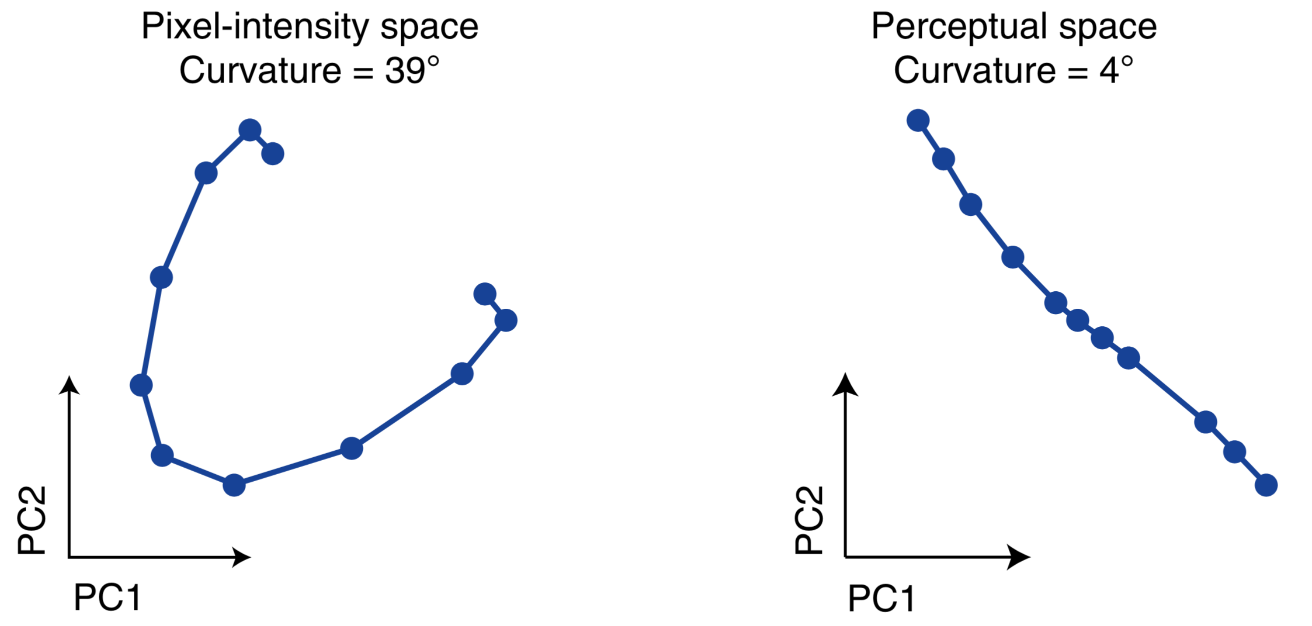
(non-parametric model)

Statistical bottleneck: $\sim N^2 \quad \gg \quad TN$ **(Limited to $N \sim 100$)**

Theory of neural population structures

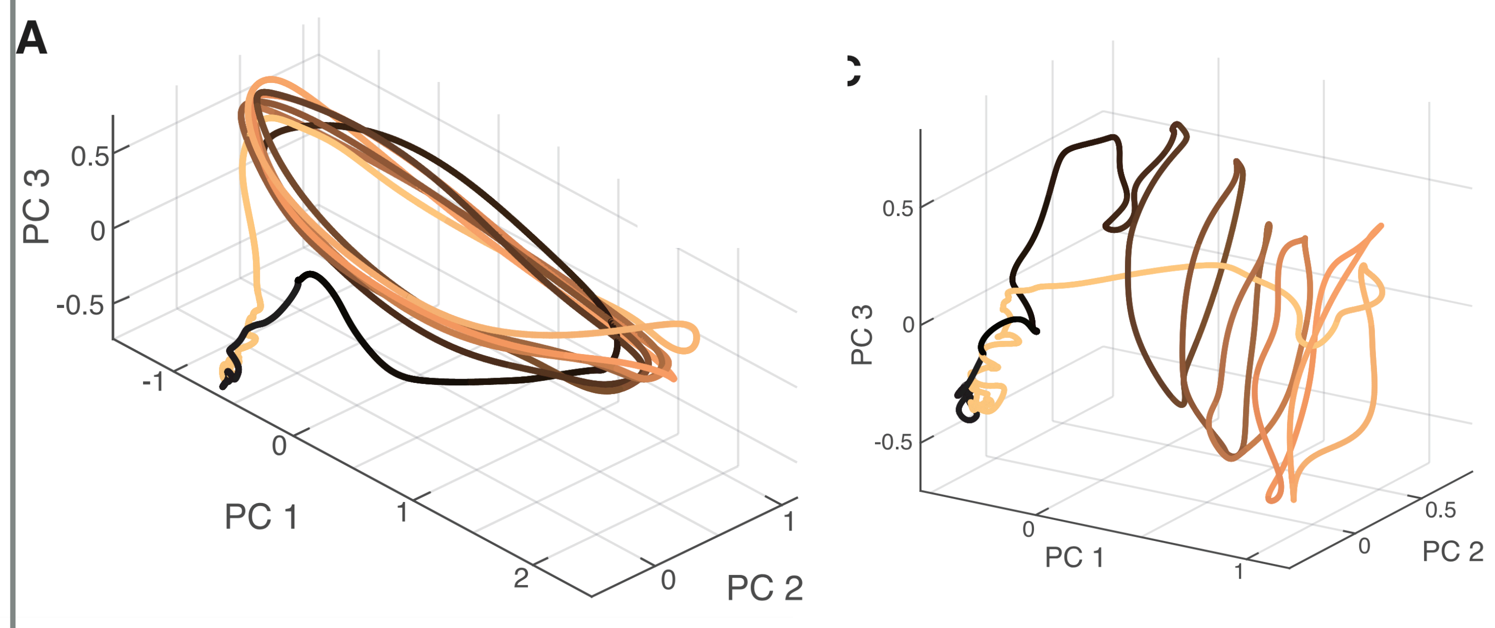
(Non-exhaustive list)

Perceptual straightening (Vision)



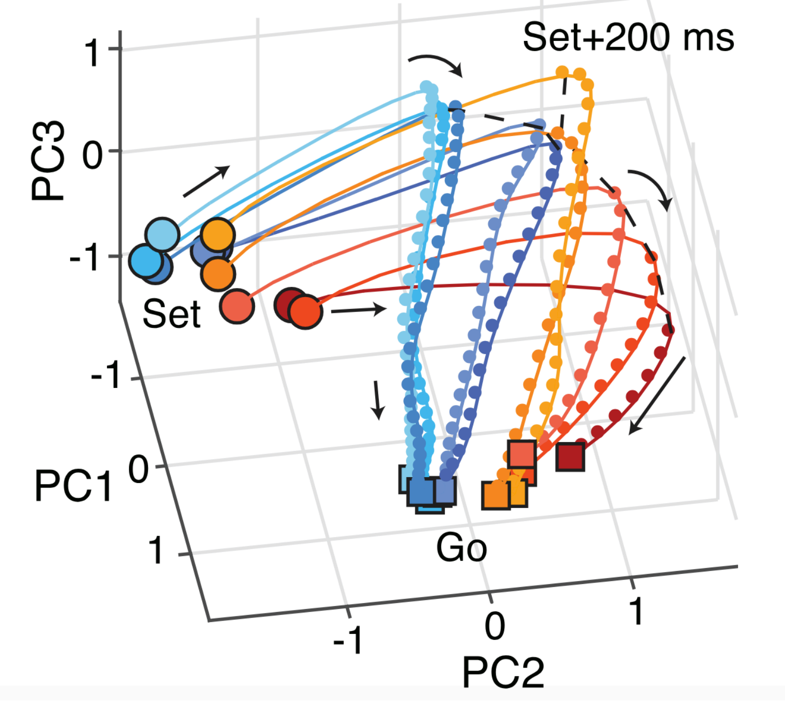
[Hénaff, Goris, Simoncelli (2019)]

Dynamical untangling (Motor)



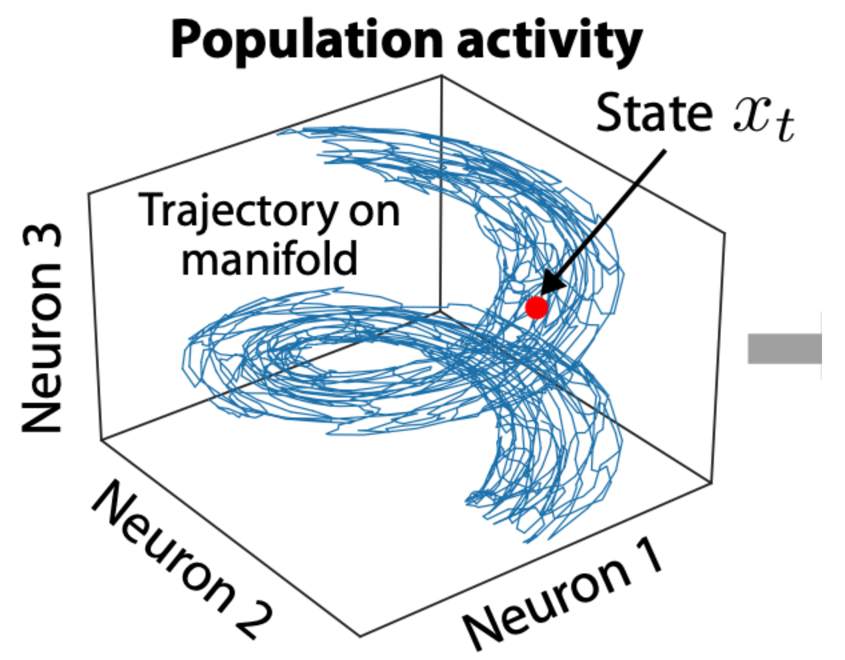
[Russo, Churchland, Abbott (2020)]

Bayesian computation through cortical latent dynamics



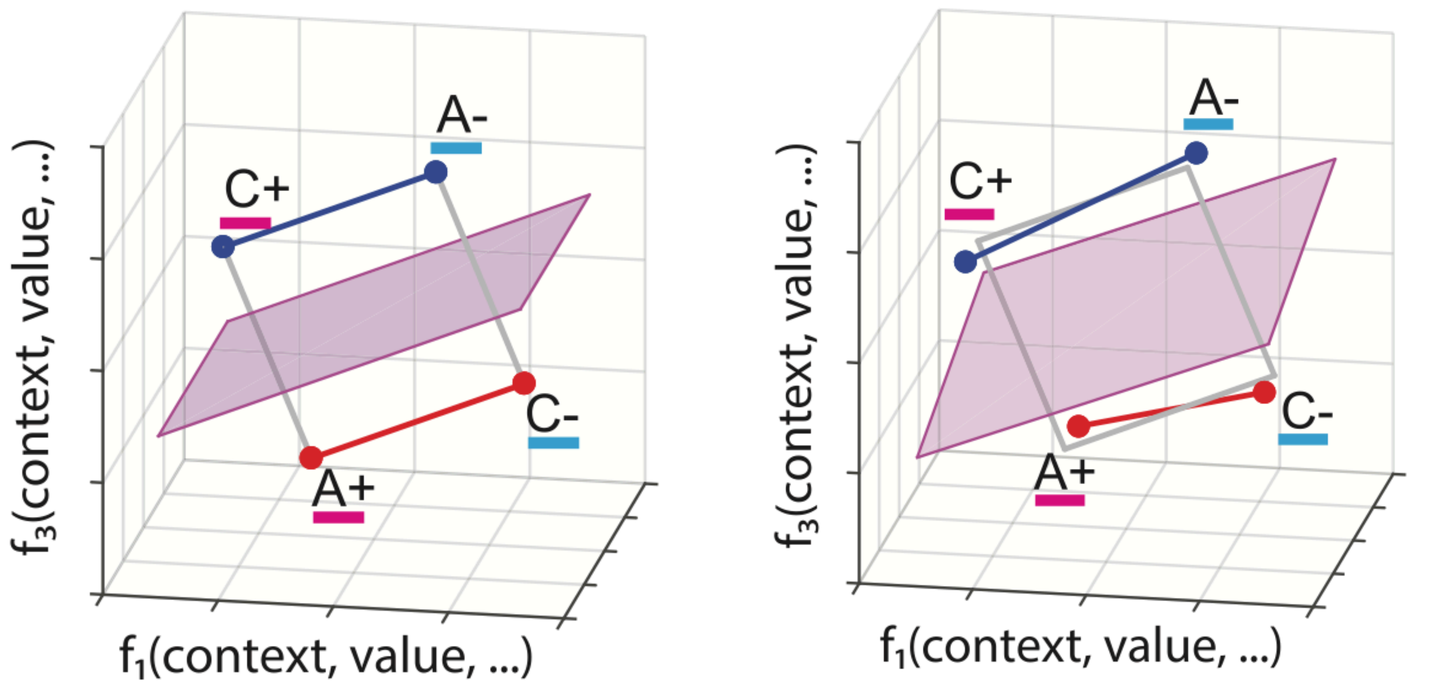
[Sohn, Jazayeri, et al. (2019)]

Topology underlying navigation, state-transition tasks



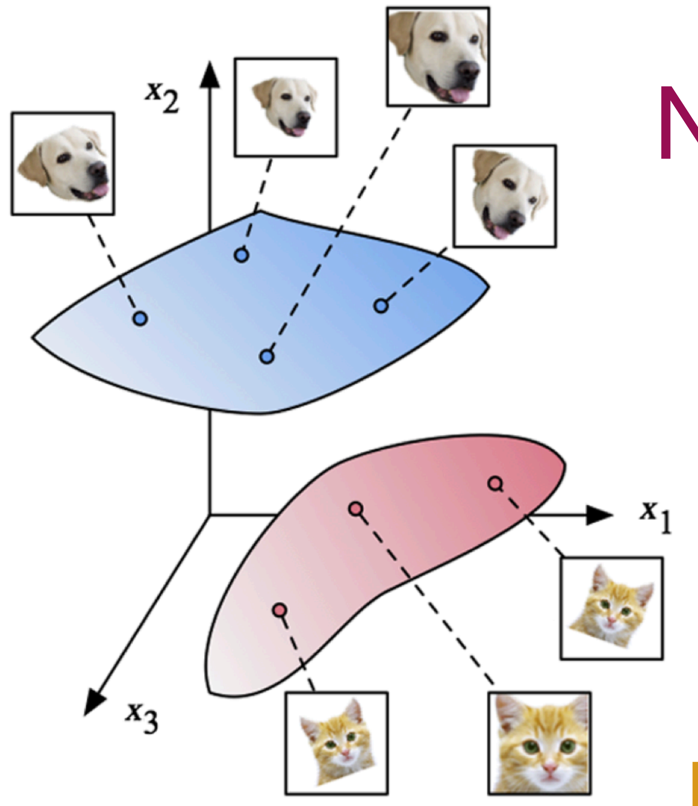
[Low, et al. (2018)]

Disentangling for abstraction tasks (Hippocampus, Pre-frontal cortex)



[Bernardi, et al. (2021)]

Neural manifold capacity



[Chung et al. (2018)]

[Review: S. Chung, L. F. Abbott, *Current opinion in neurobiology*, 70, 137-144]

Can we scale max-ent models to large neural systems ?

Can we describe these systems with a small number ($\ll N$) of *collective* variables*?

(the “neural manifold hypothesis”)

* “order parameters”, “summary statistics”, …

To what extent are these reduced descriptions captured by a mean-field theory?

Can we leverage these theories to solve efficiently the inverse problem of fitting maximum entropy parameters?

Max-ent on informative collective coordinates

Projections of the neural activity: $\varphi_\alpha = \sum_{n=1}^{N} W_n^\alpha \sigma_n$ given $W^\alpha \in \mathbb{R}^N$, $\alpha = 1,\ldots,M \ll N$

Measurements:

- Mean activities: $\frac{1}{T}\sum_{t=1}^{T} \bar{\sigma}_i^t = \mu_i^{\exp}$
- Pairwise correlations **among projections**: $\frac{1}{T-1}\sum_{t=1}^{T} (\bar{\varphi}_\alpha^t - \langle \bar{\varphi}_\alpha^t \rangle)(\bar{\varphi}_\beta^t - \langle \bar{\varphi}_\beta^t \rangle) = \chi_{\alpha\beta}^{\exp}$

Maximum entropy model:
($\sim NM$ parameters)

$$P(\underline{\sigma}) = \frac{1}{Z(\Lambda, h)} \exp\left(\frac{1}{2}\sum_{\alpha \le \beta} \Lambda_{\alpha\beta}\, \varphi_\alpha \varphi_\beta + \sum_{i=1}^{N} h_i \sigma_i \right)$$

[Cocco, Monasson, Sessak, PRE (2011)]

… then, select the **most informative directions** W (mini-max entropy)

[Lynn et al. (2023), Carcamo & Lynn (2024)]

The simplest case: population activity

- Match the mean and variance of the average firing rate: $m = \sum_{n=1}^{N} \sigma_n$

Maximum entropy model:

(Fully-connected ferromagnet)

$$P(\underline{\sigma}) = \frac{1}{Z(\lambda, h)} \exp\left(\frac{\lambda}{2N} \left(\sum_{i=1}^{N} \sigma_i \right)^2 + h \sum_{i=1}^{N} \sigma_i \right)$$

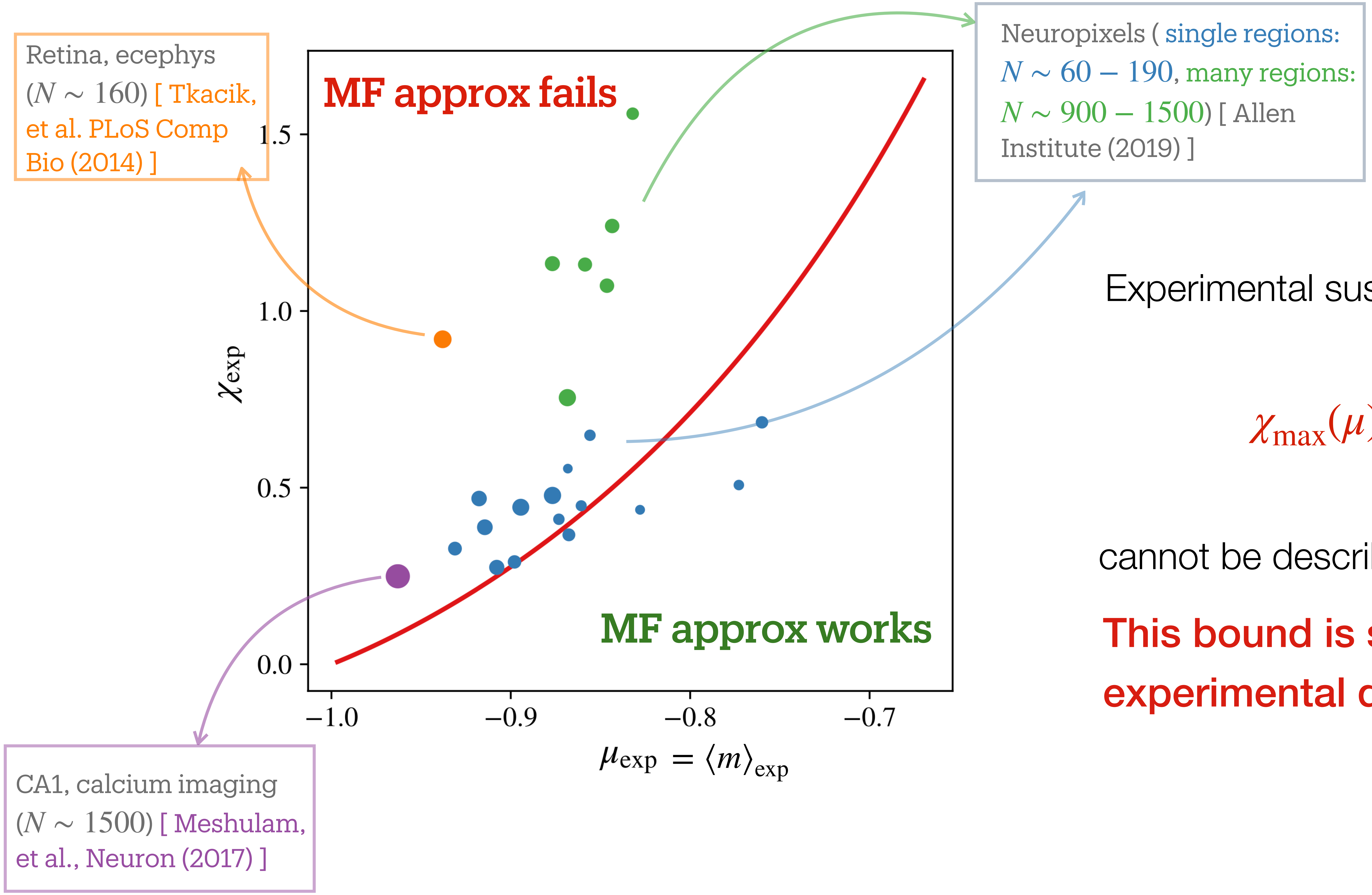
Mean-field approximation:

$$Z(\lambda, h) = \sqrt{\frac{N}{2\pi\lambda}}\, 2^N \int \mathrm{d}\psi\, \mathrm{e}^{-Nf(\psi)} \simeq \sqrt{\frac{1}{\lambda f''(\psi_{\mathrm{sp}})}}\, 2^N\, \mathrm{e}^{-Nf(\psi_{\mathrm{sp}})}$$

Saddle point approximation

Local free energy: $f(\psi) = \frac{1}{2\lambda}\psi^2 - \ln\cosh\left(h + \psi\right)$

The simplest case: population activity



Experimental susceptibilities above

$$\chi_{\max}(\mu) = \frac{\mu(1-\mu^2)}{\mu - \operatorname{atanh}(\mu)(1-\mu^2)}$$

cannot be described by naive MF approximation .

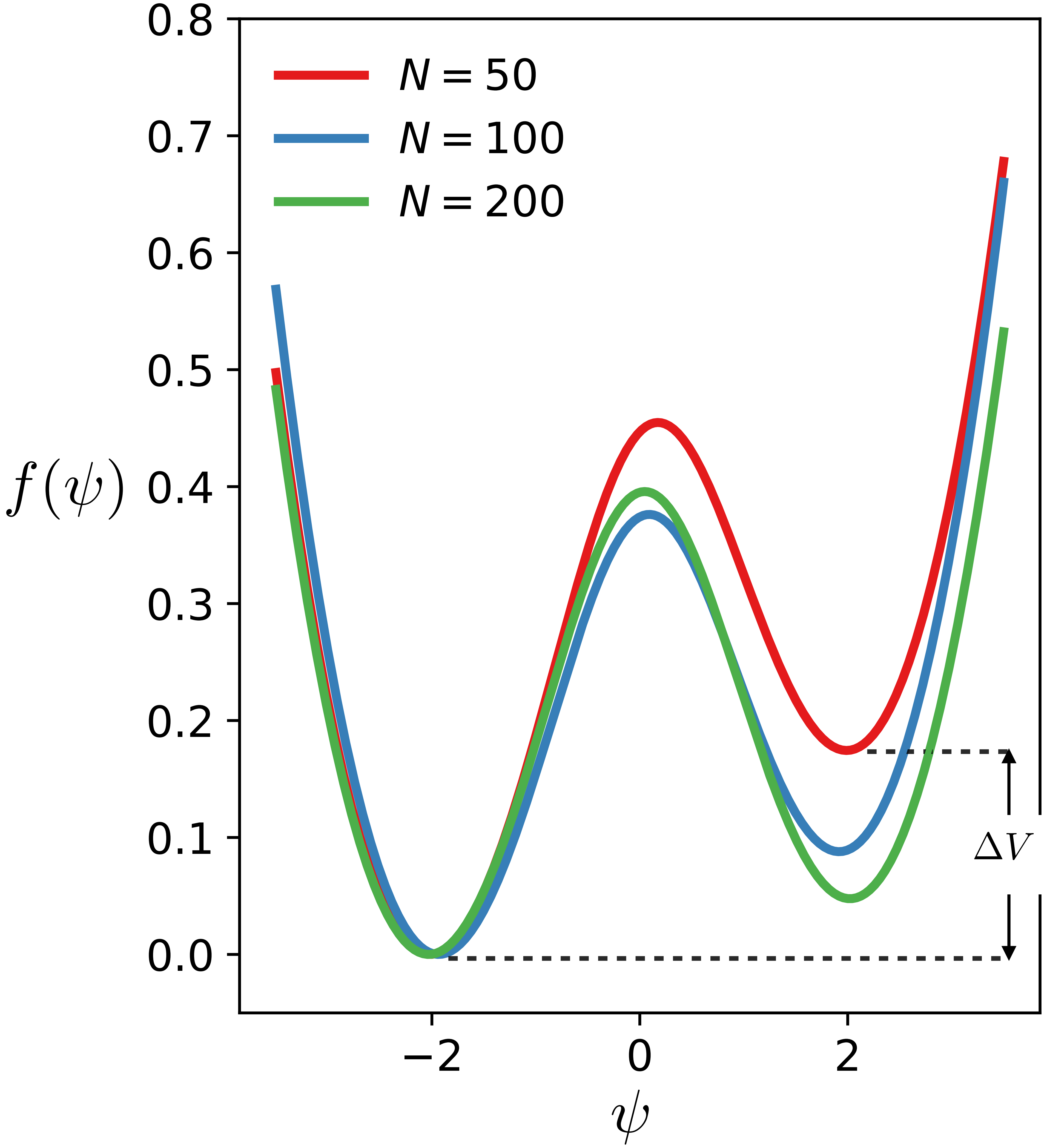
This bound is systematically violated by experimental data.

The simplest case: population activity

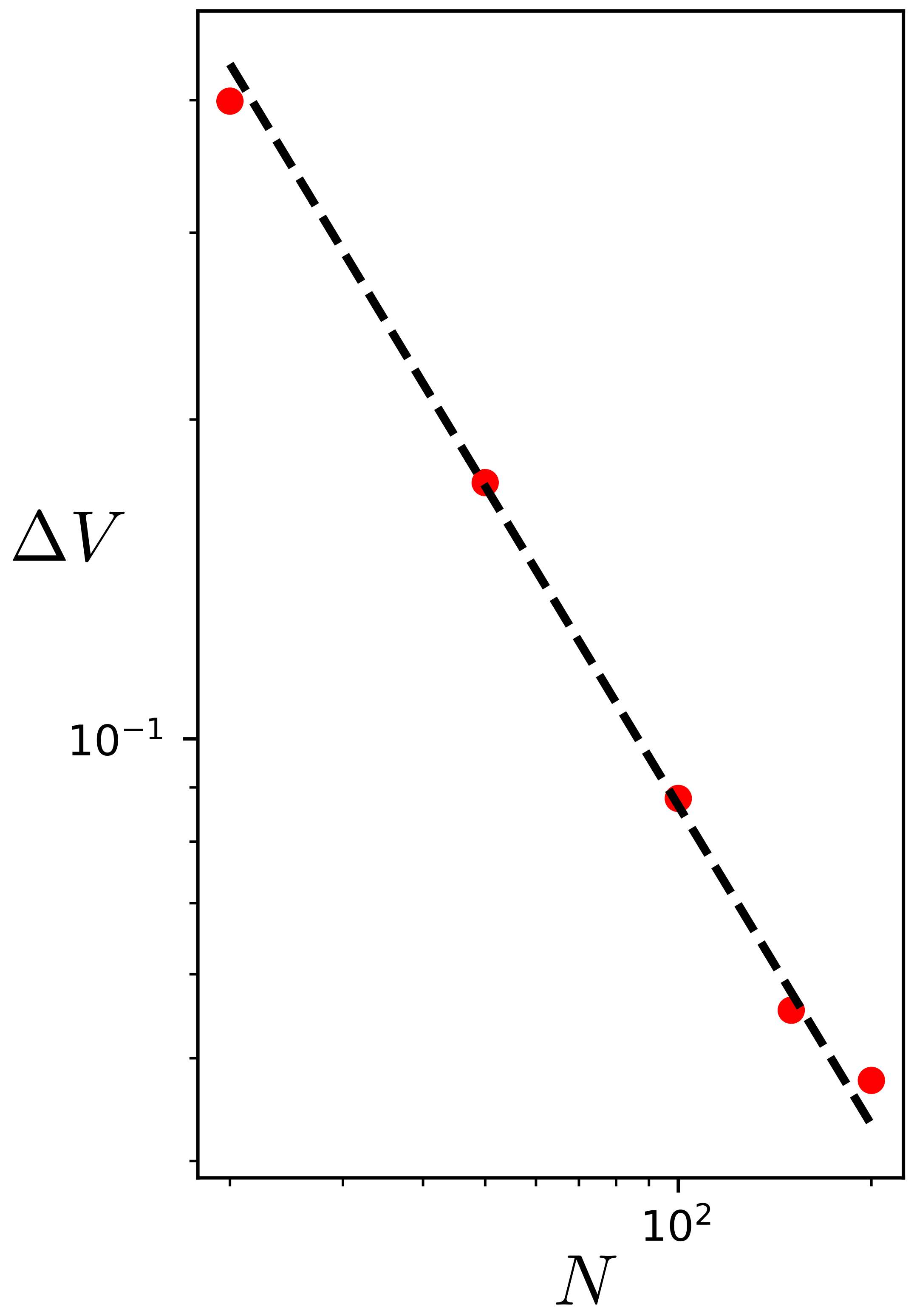
What went wrong?

The exact solution of the model shows that the **local free energy** has **two nearly-degenerate minima**.

This model leads to **fundamentally wrong predictions**, such as a bimodal distribution of population activity.



CA1, calcium imaging ($N \sim 1500$) [Meshulam, et al., Neuron (2017)]



The difference vanishes as the system size grows: $\Delta V \sim N^{-1}$.

Max-ent on multiple projections

Projections of the neural activity: $\varphi_\alpha = \sum_{n=1}^{N} W_n^\alpha \sigma_n$ given $W^\alpha \in \mathbb{R}^N$, $\alpha = 1,\ldots,M \ll N$

Measurements:

- Mean activities: $\frac{1}{T}\sum_{t=1}^{T} \bar{\sigma}_i^t = \mu_i^{\exp}$
- Pairwise correlations **among projections:** $\frac{1}{T-1}\sum_{t=1}^{T} (\bar{\varphi}_\alpha^t - \langle \bar{\varphi}_\alpha^t \rangle)(\bar{\varphi}_\beta^t - \langle \bar{\varphi}_\beta^t \rangle) = \chi_{\alpha\beta}^{\exp}$

Maximum entropy model:
(inverse Hopfield)

$$P(\underline{\sigma}) = \frac{1}{Z(\Lambda, h)} \exp\left(\frac{1}{2} \sum_{\alpha \le \beta} \Lambda_{\alpha\beta} \sum_{n=1}^{N} W_n^\alpha \sigma_n \sum_{m=1}^{N} W_m^\beta \sigma_m + \sum_{i=1}^{N} h_i \sigma_i \right)$$

MF approximation:

$$\Lambda_{\mathrm{MF}} = \left((\chi_0^{\exp})^{-1} \chi^{\exp} - \mathbf{I} \right) (\chi_0^{\exp})^{-1}, \qquad h_{\mathrm{MF}} = \mathrm{atanh}(\mu^{\exp}) - \frac{1}{N} W^\top \Lambda_{\mathrm{MF}} W \mu^{\exp}$$

Max-ent on multiple projections

How do we select the *most informative* directions ?

→ Measure the information gain with respect to the **independent model**:

$$P_0(\underline{\sigma}) = \frac{1}{Z(h_0)} \exp\left(\sum_{i=1}^{N} h_{0,i}\, \sigma_i \right)$$

Entropy reduction:

$$\Delta S = S_0 - S \;\; = \frac{1}{2} \mathrm{Tr} \left[\chi_0^{-1} \chi - \ln(\chi_0^{-1} \chi) - \mathbb{1} \right]$$

Fluctuations of the independent model (χ_0)

Fluctuations of the pairwise model (χ)

Max-ent on multiple projections

How do we select the *most informative* directions ?

→ Measure the information gain with respect to the **independent model**:

$$P_0(\underline{\sigma}) = \frac{1}{Z(h_0)} \exp\left(\sum_{i=1}^{N} h_{0,i}\, \sigma_i \right)$$

Entropy reduction:

$$\Delta S = S_0 - S = \frac{1}{2} \mathrm{Tr}\left[\chi_0^{-1}\chi - \ln(\chi_0^{-1}\chi) - \mathbb{1} \right]$$

By taking $W_n^\alpha = U_n^\alpha \Big/ \sqrt{1-\mu_n^2}$,

where U^α are the eigenvectors of the data correlation matrix.

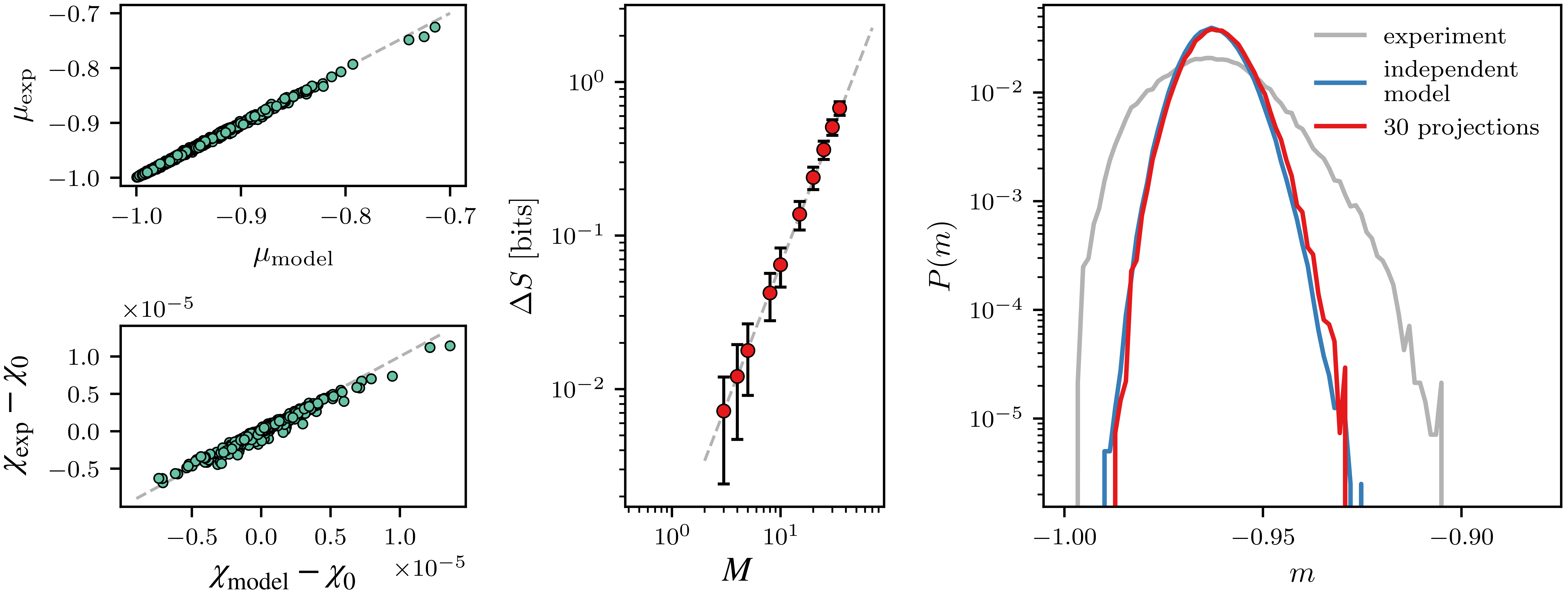
$$= \frac{1}{2} \sum_{\alpha=1}^{M} \left[\rho_\alpha - \ln \rho_\alpha - 1 \right]$$

Eigenvalues of the data correlation matrix

Max-ent on multiple projections

However:

1. Random projections are **consistent** with MF approximation, **but uninformative**.

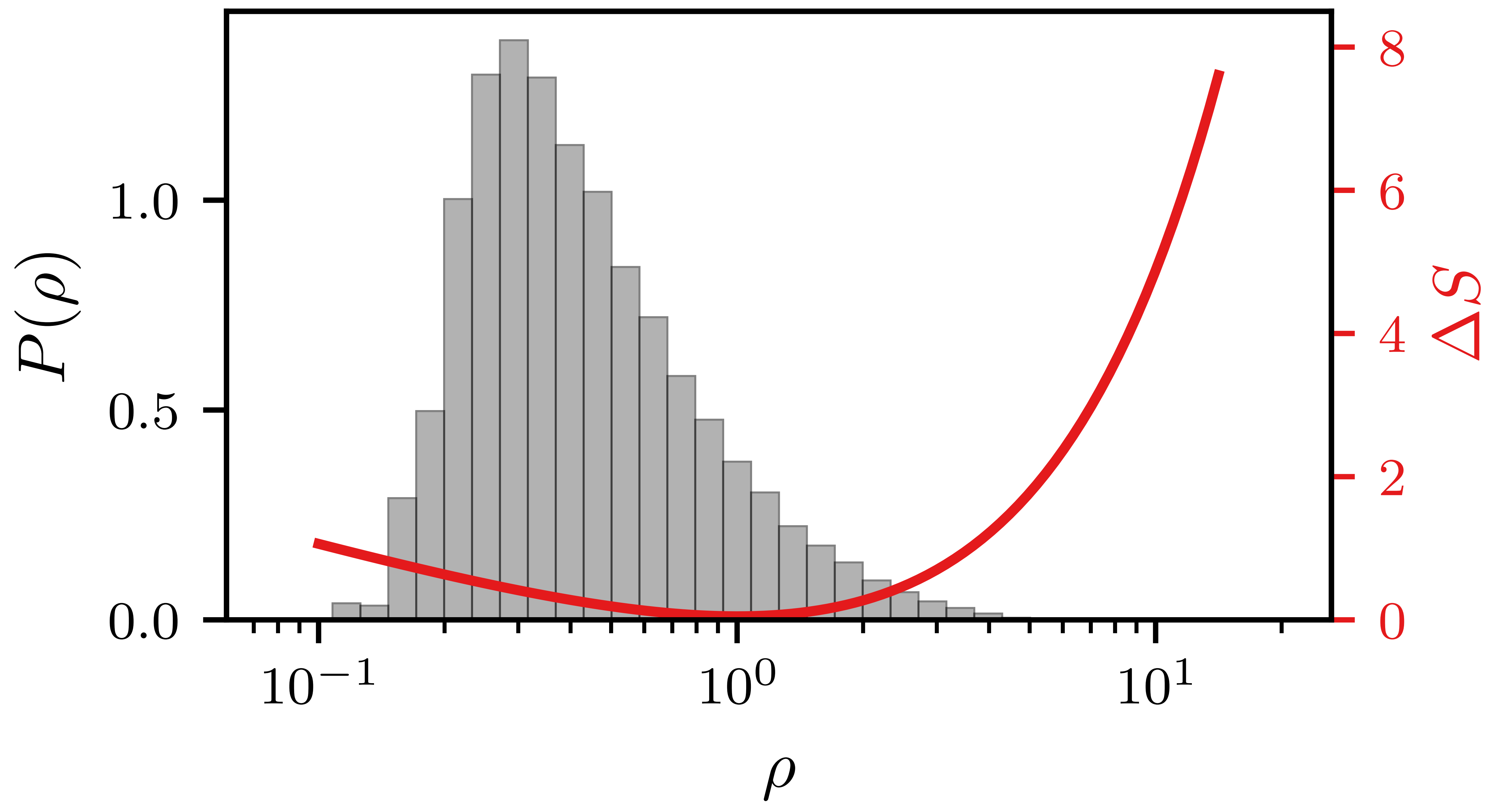


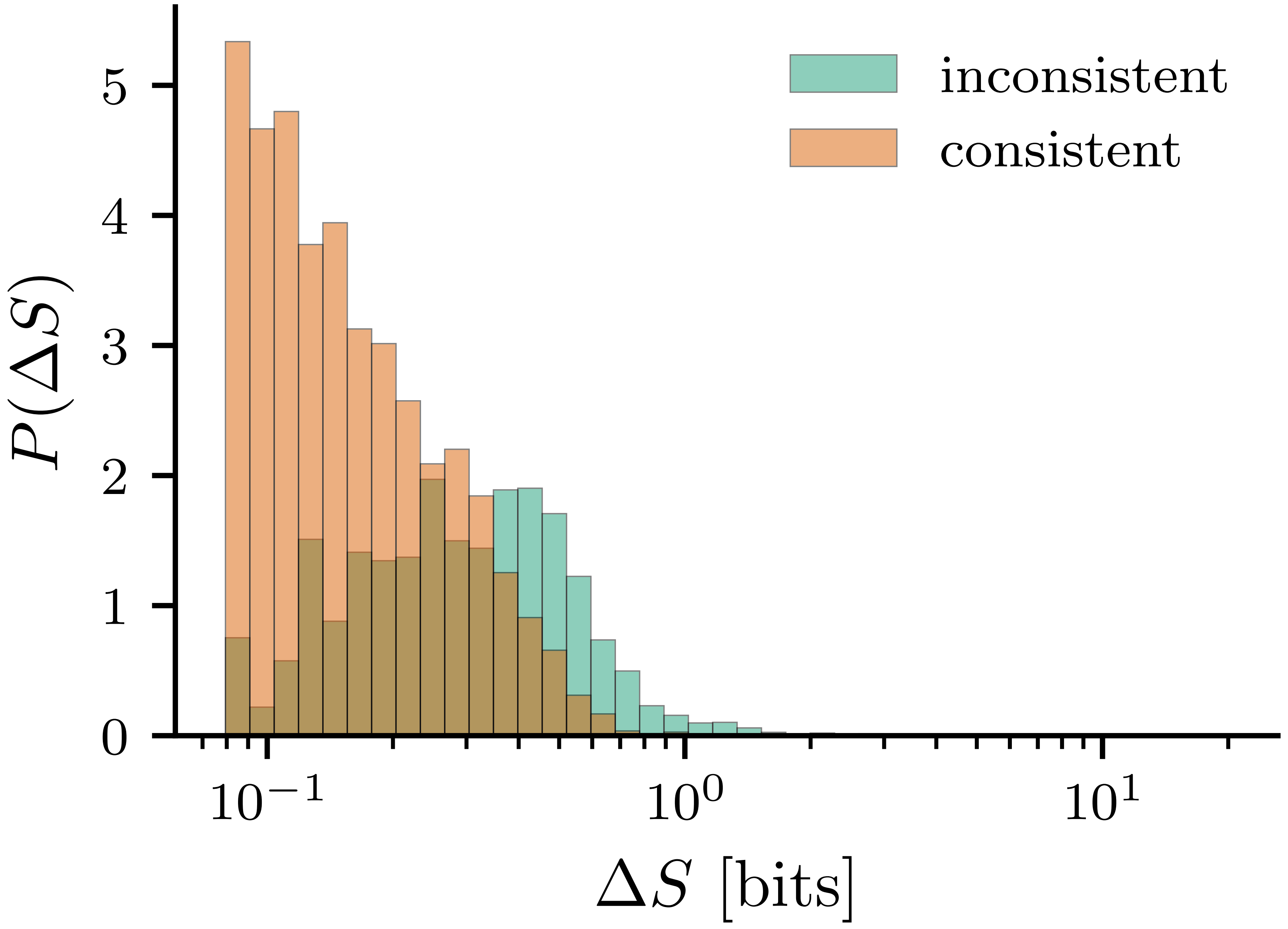
Max-ent on multiple projections

However:

1. **Random projections** are **consistent** with MF approximation, **but uninformative**.

2. **Informative projections** are **inconsistent** with MF approximation.





Max-ent on multiple projections

However:

1. **Random projections** are **consistent** with MF approximation, **but uninformative**.

2. **Informative projections** are **inconsistent** with MF approximation.

Can we do better ?

Max-ent on the full distribution of a projection

Projection of the neural activity: $\varphi = \sum_{n=1}^{N} W_n \sigma_n$ given $W \in \mathbb{R}^N$

Measurements:

- Mean activities: $\frac{1}{T}\sum_{t=1}^{T} \bar{\sigma}_i^t = \mu_i^{\text{exp}}$
- The full distribution of the projection: $P_{\text{theory}}(\varphi) = \sum_{\sigma} \delta\left(\varphi - \sum_{n=1}^{N} W_n \sigma_n\right) P(\sigma) = P_{\text{exp}}(\varphi)$

Maximum entropy model: $E(\sigma) = -\sum_{n=1}^{N} h_n \sigma_n + N U(\varphi)$

Fit the potential from experimental data.

Max-ent on the full distribution of a projection

MF approximation:

$$Z = 2^N \int \frac{\mathrm{d}z}{2\pi} \int \mathrm{d}\varphi\, \mathrm{e}^{-Nf(\varphi,z)} \underset{N\gg 1}{\simeq} 2^N \frac{\mathrm{e}^{-Nf(\varphi_{sp},z_{sp})}}{\sqrt{\det \mathscr{H}_f(\varphi_{sp}, z_{sp})}}$$

Local free energy:

$$f(\varphi, z) = U(\varphi) + \frac{1}{N}\left[\mathrm{i}z\varphi - \sum_{n=1}^{N} \ln\cosh\left(h_n + \mathrm{i}zW_n\right)\right]$$

Solve self-consistently:

$$\varphi = \sum_{n=1}^{N} W_n \tanh\left(h_n + \mathrm{i}z^*(\varphi)W_n\right)$$

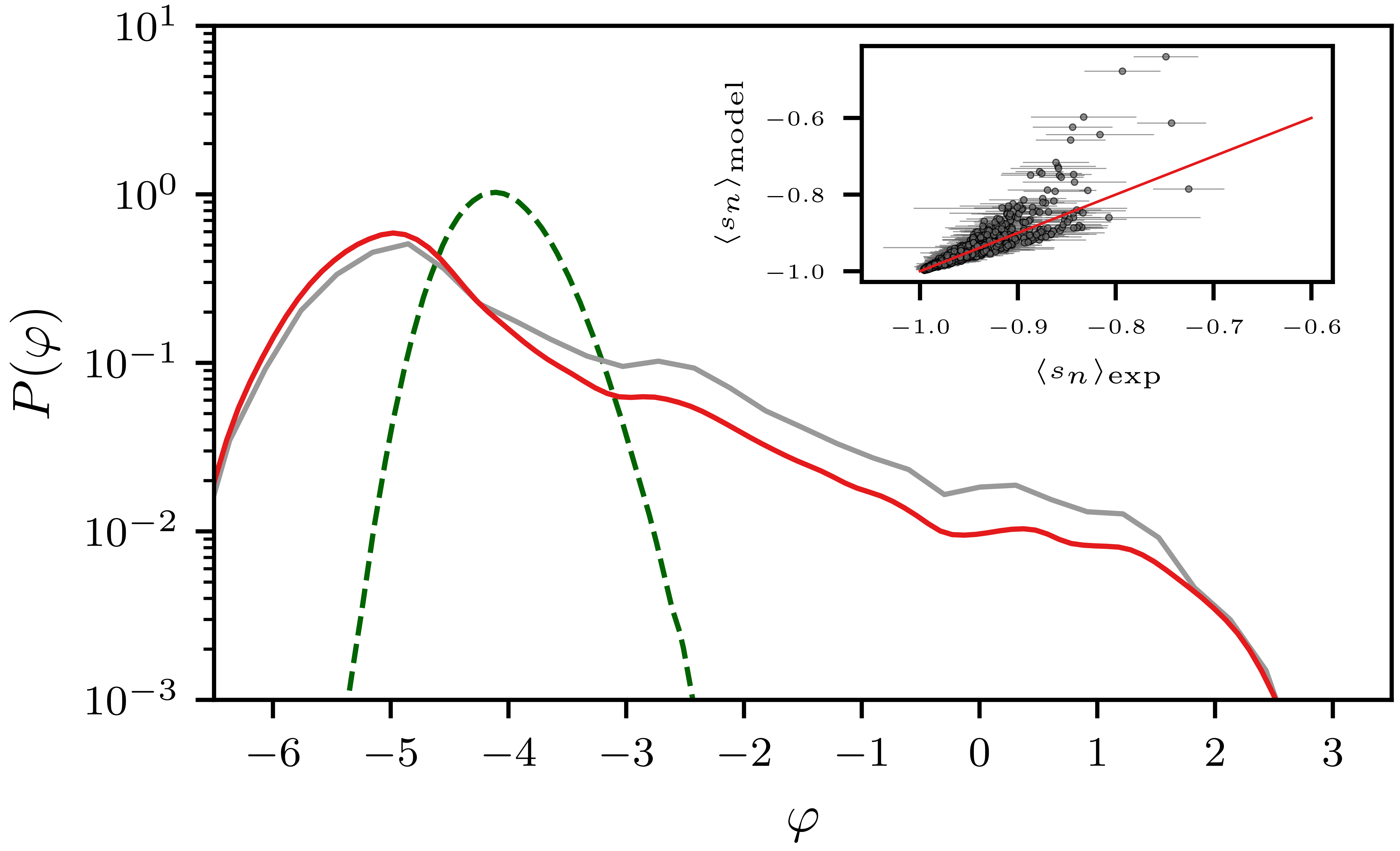
$$h_n = \mathrm{atanh}(\mu_n^{\exp})$$

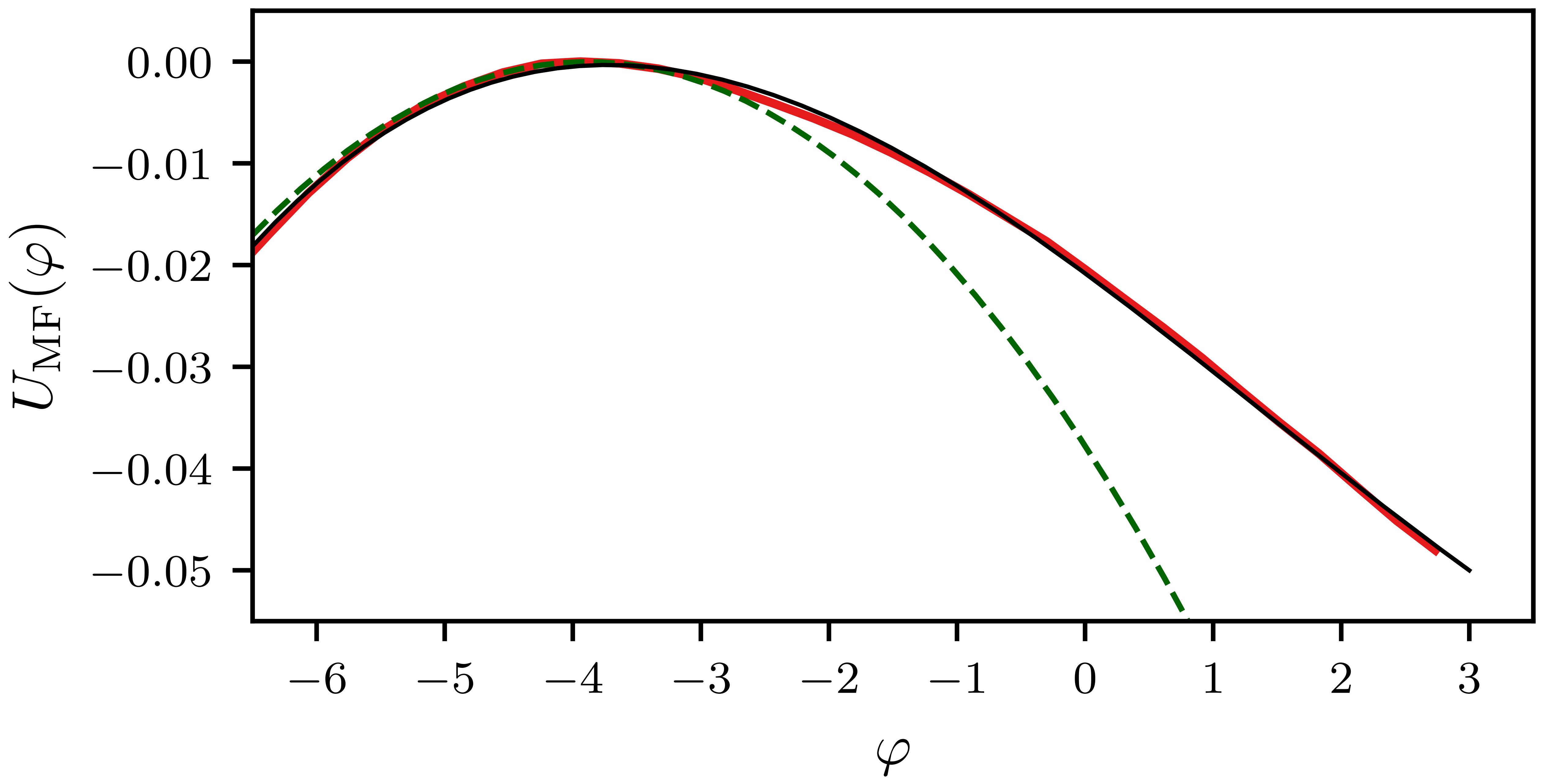
$$U_{MF}(\varphi) = \frac{1}{N}\left[-\ln P_{\exp}(\varphi) - \mathrm{i}z^*(\varphi)\varphi + \sum_n \ln\cosh\left(h_n + \mathrm{i}z^*(\varphi)W_n\right)\right]$$

The MF approximation is consistent

$W =$ principal component of the data correlation matrix.

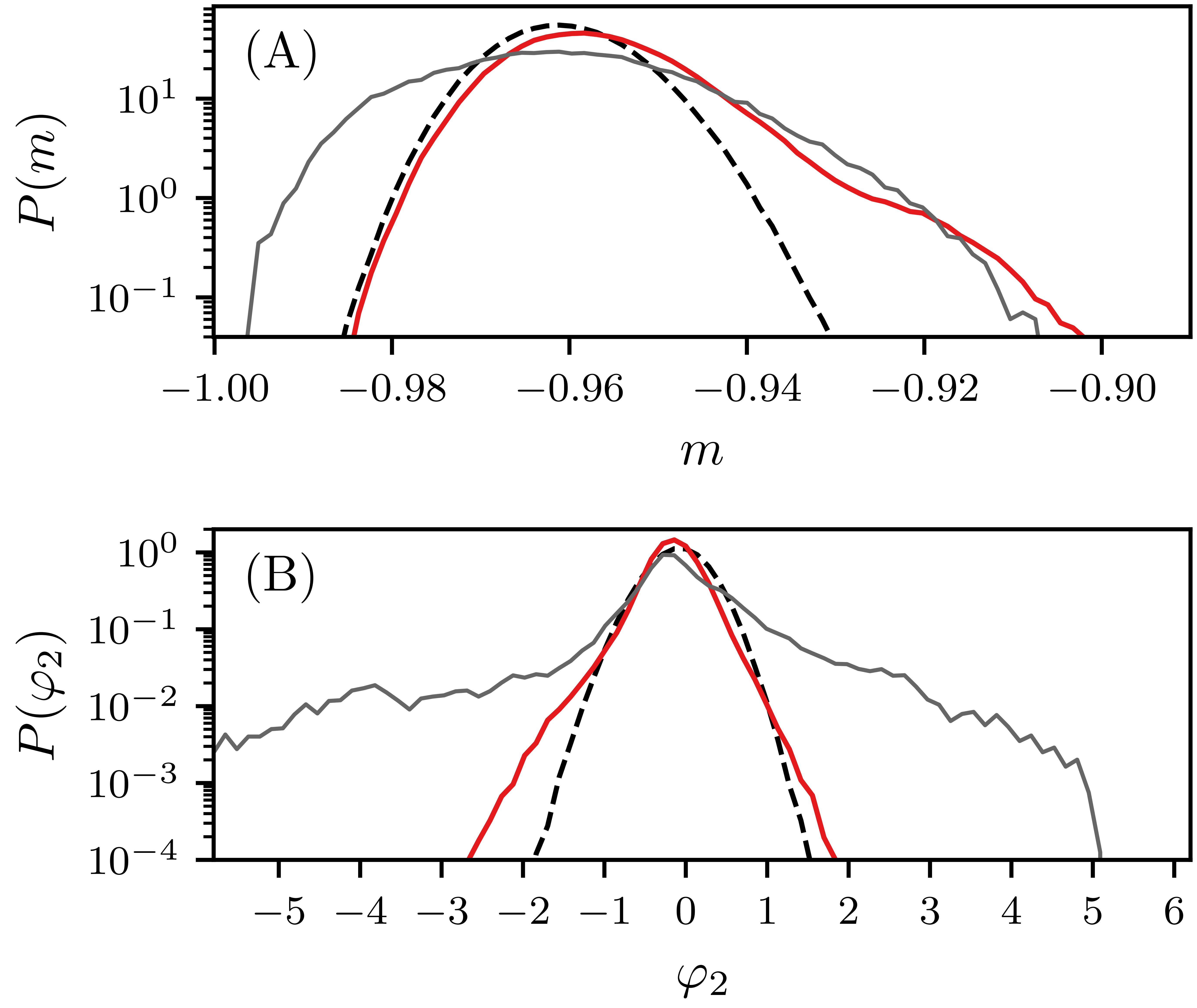
CA1, calcium imaging ($N \sim 1500$) [Meshulam, et al., Neuron (2017)]





✔ The extended mean-field theory on the projection distribution reproduces the maximum-entropy constraints on experimental data.

Is the model predictive?



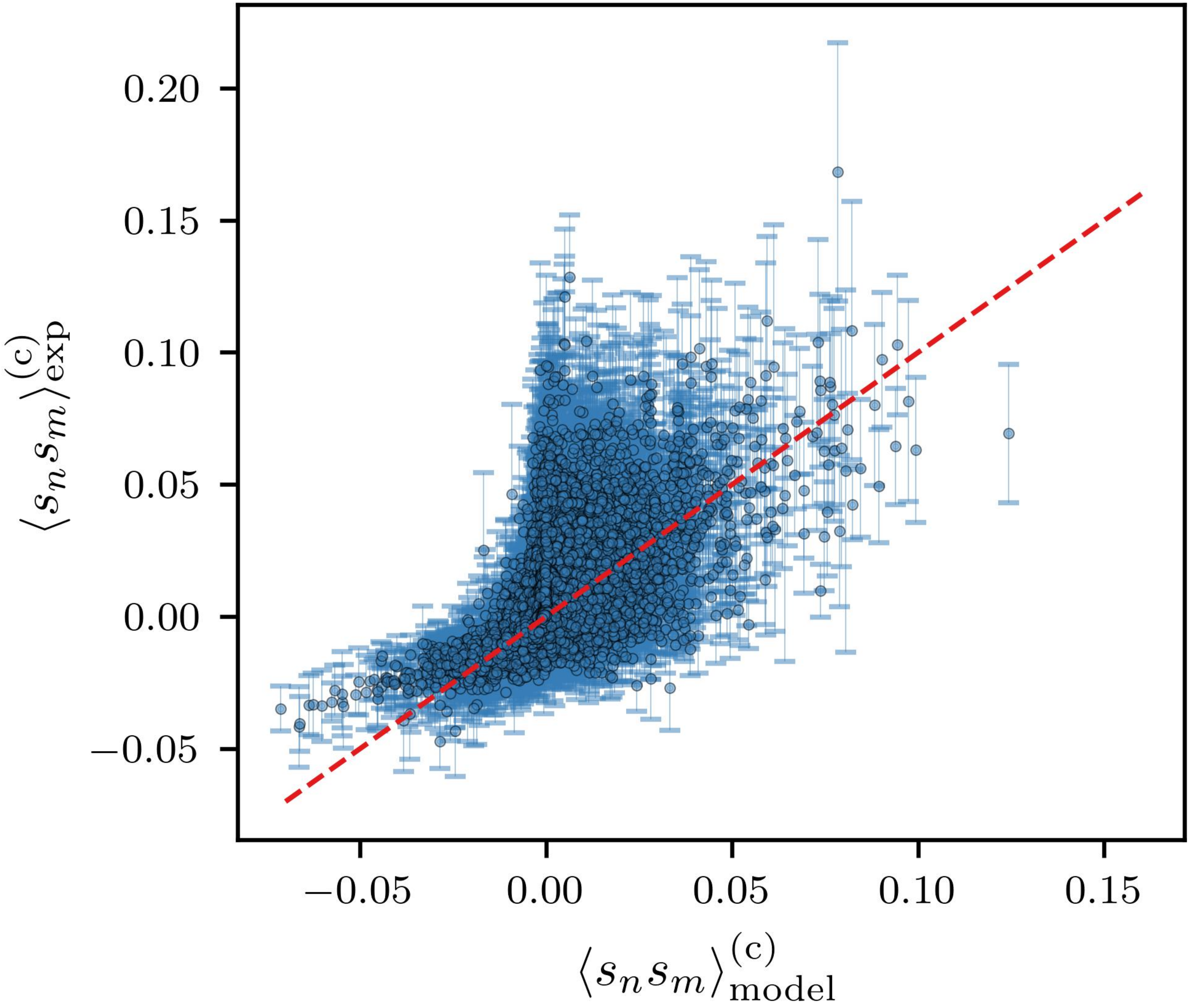
Total population activity

The model reproduces the highly non Gaussian right tail of the distribution.

2nd principal component

Only marginally better than the independent model.

Is the model predictive?



Although pairwise correlations are not explicitly constrained, the model captures the overall trend in the correlation matrix.

(W = 1st PC)

Which directions are most informative?

Entropy reduction from the independent model:

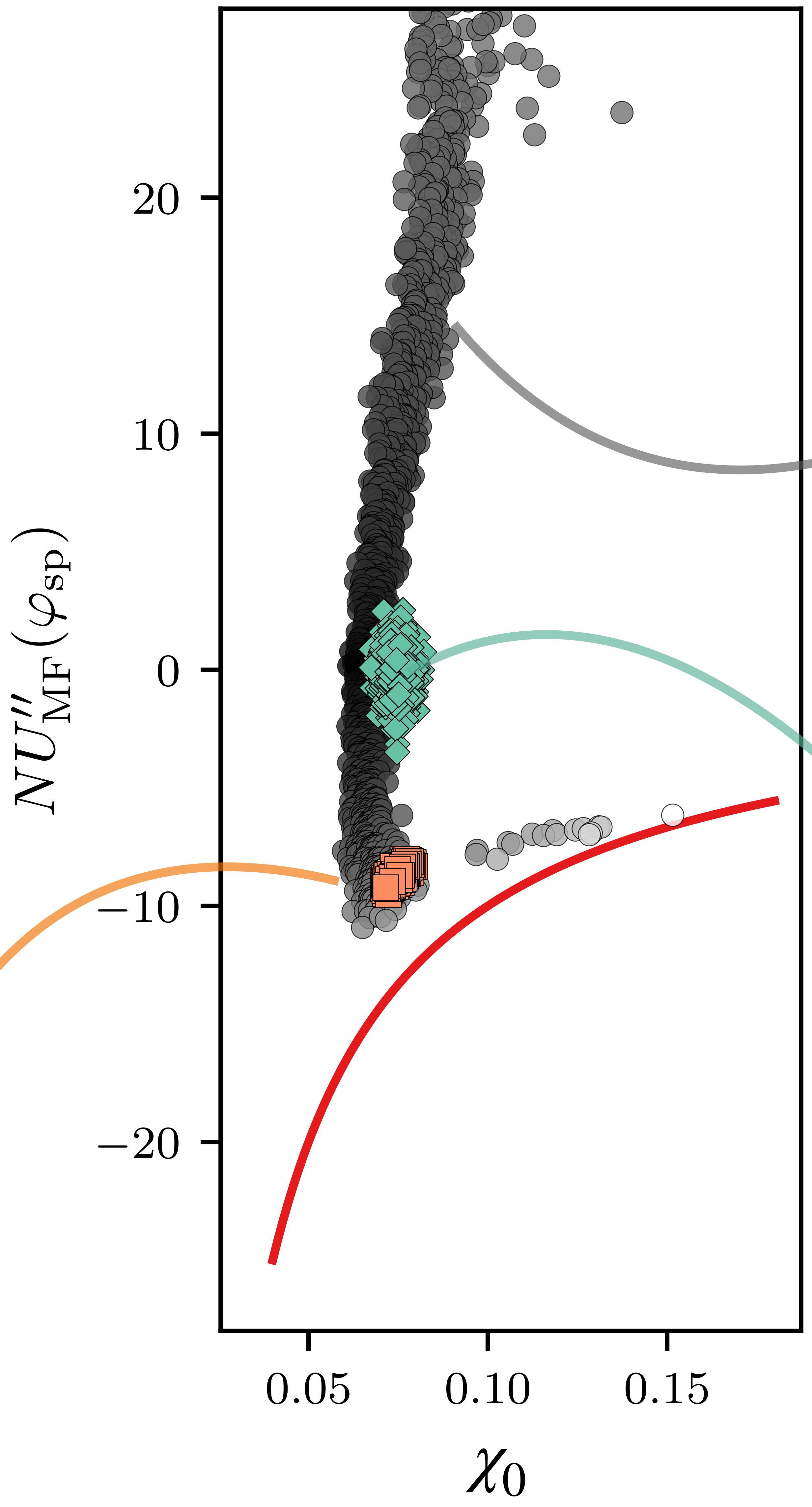
$$\Delta S = N\left(U(\varphi_{sp}) - \langle U(\varphi)\rangle\right) + \frac{1}{2}\ln\left(NU''(\varphi_{sp})\chi_0 + 1\right)$$

$W =$ eigenvectors of the data correlation matrix

The entropy reduction is highest at the boundaries of the spectrum of the neural correlation matrix.

The 1st PC of the correlation matrix reduces the entropy of approximately 5% of the independent entropy.

The system is poised near a critical point



Eigenvectors of data correlation matrix

(Lighter = higher ΔS)

Random projections

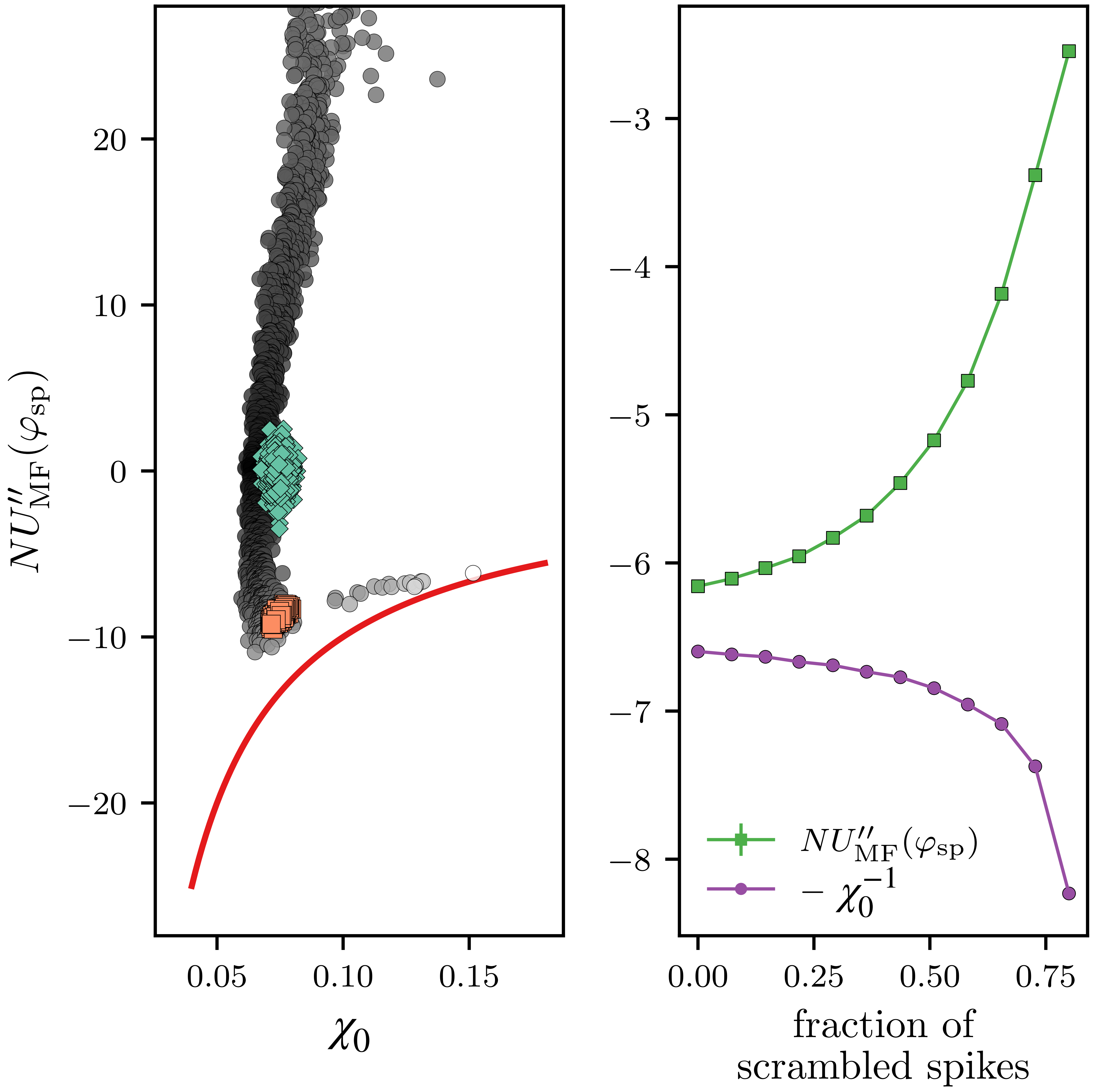
Random positive projections

Critical line: $NU''(\varphi_{sp}) = -\chi_0^{-1}$

Most **informative directions** describe neural systems **close to a second order phase transition.**

[See e.g.: Meshulam et al (2019); | Meshulam and Bialek (2024)]

The system is poised near a critical point



Critical line: $NU''(\varphi_{\mathrm{sp}}) = -\chi_0^{-1}$

Most **informative directions** describe neural systems **close to a second order phase transition**.

[See e.g.: Meshulam et al (2019); | Meshulam and Bialek (2024)]

Plausible neural populations (same mean activity, weaker correlations) are farther away from criticality than the real network.

Conclusions & perspectives

- Understanding large neural populations requires **scalable methods that avoid undersampling** relevant statistics.

- A promising path in this direction involves **maximum entropy models** constrained **on informative collective coordinates** in neural activity.

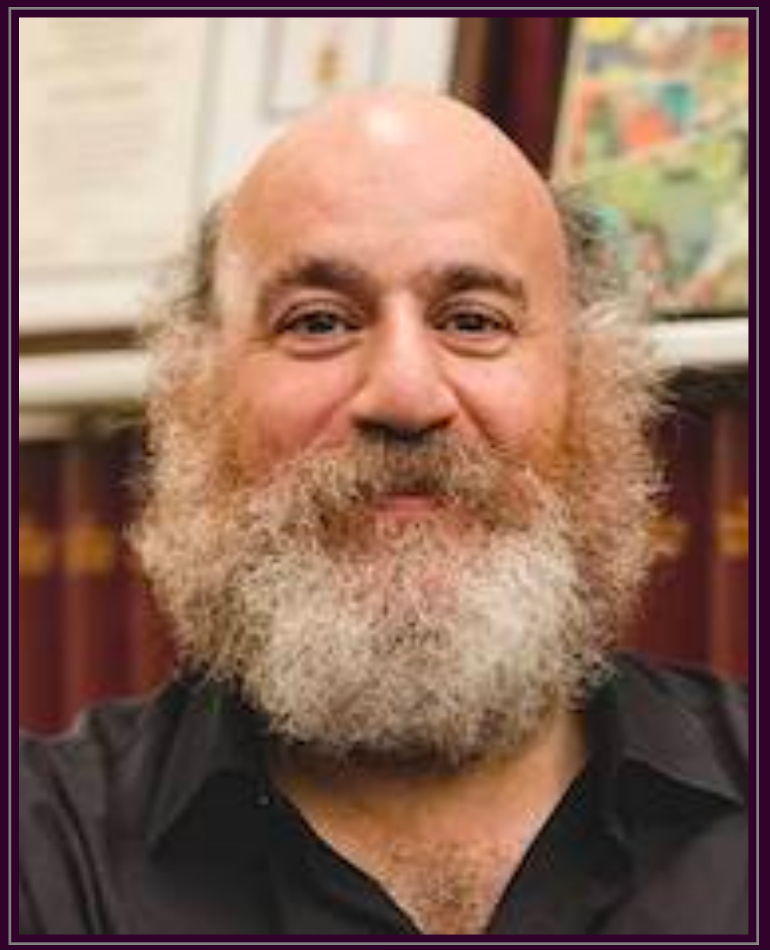
- Naive mean-field theories limited to pairwise correlations in these neural subspaces fail to capture the complexity of neural activity.

- Extending the theory to the **full distribution along one informative projection**, we obtain consistent and accurate models.

- **Expanding this approach to constrain the distribution of neural activity along multiple projections is a key next step to advance large-scale neural modeling.**

Luca Di Carlo
Princeton University

Christopher Lynn
Yale University



William Bialek
Princeton University

arXiv:2504.15197

\+ check arXiv today for a longer version: ***arXiv:2508.02633***

Thank you!

THE GRADUATE CENTER
CITY UNIVERSITY OF NEW YORK

Center for the Physics of Biological Function

SIMONS FOUNDATION