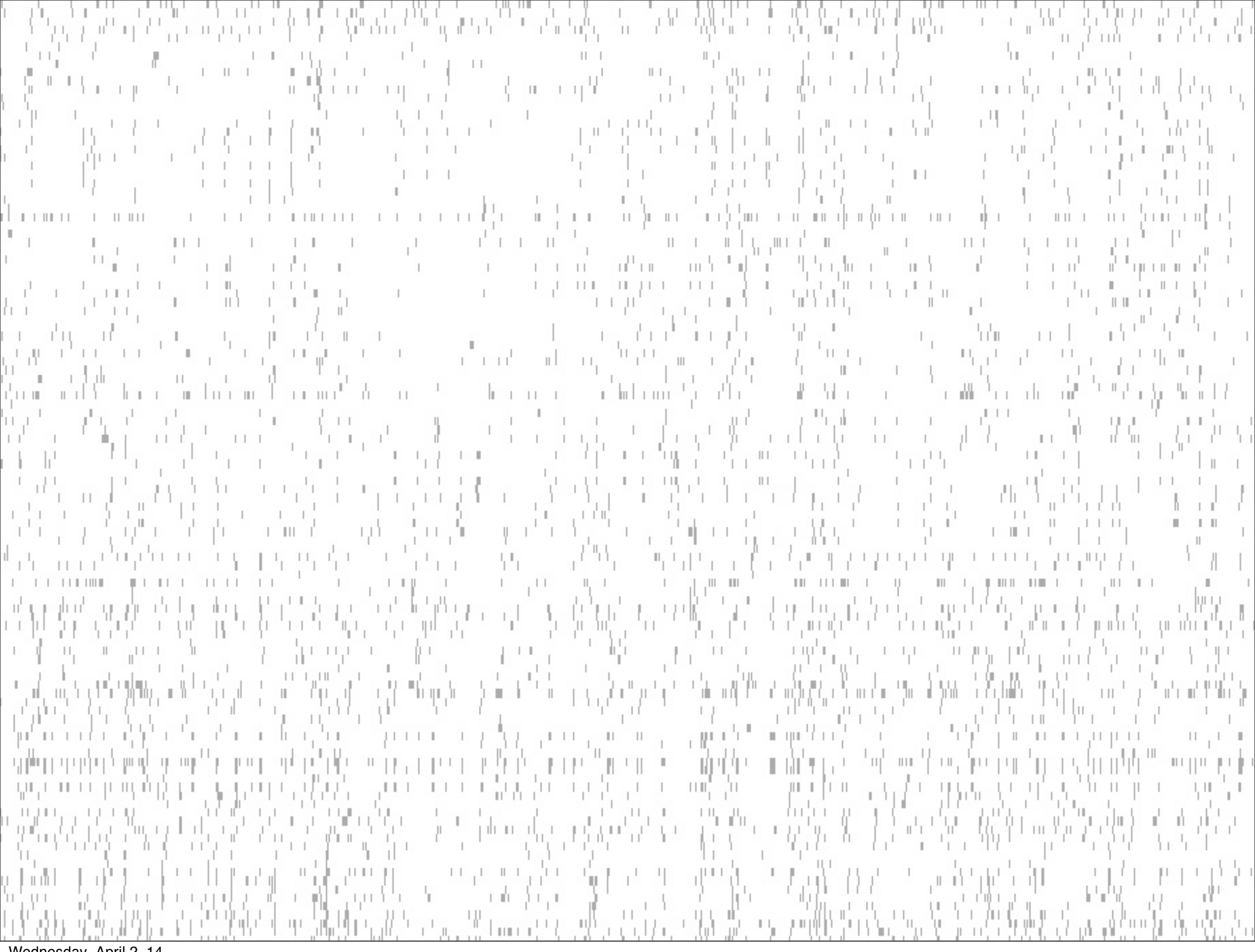
Critical behavior in networks of real neurons

Gašper Tkačik, IST Austria

Olivier Marre, Dario Amodei, Michael Berry
Thierry Mora, William Bialek



Questions

Questions

- How can we make sense of the joint activity of many neurons encoding rich (naturalistic) stimuli at the single-spike level?

Questions

- How can we make sense of the joint activity of many neurons encoding rich (naturalistic) stimuli at the single-spike level?
- Is the "code" independent? If not, what is the nature of the collective activity? Is it combinatorial?

Questions

- How can we make sense of the joint activity of many neurons encoding rich (naturalistic) stimuli at the single-spike level?
- Is the "code" independent? If not, what is the nature of the collective activity? Is it combinatorial?
- How can we connect theoretical work on neural coding to new, high-dimensional data?

Questions

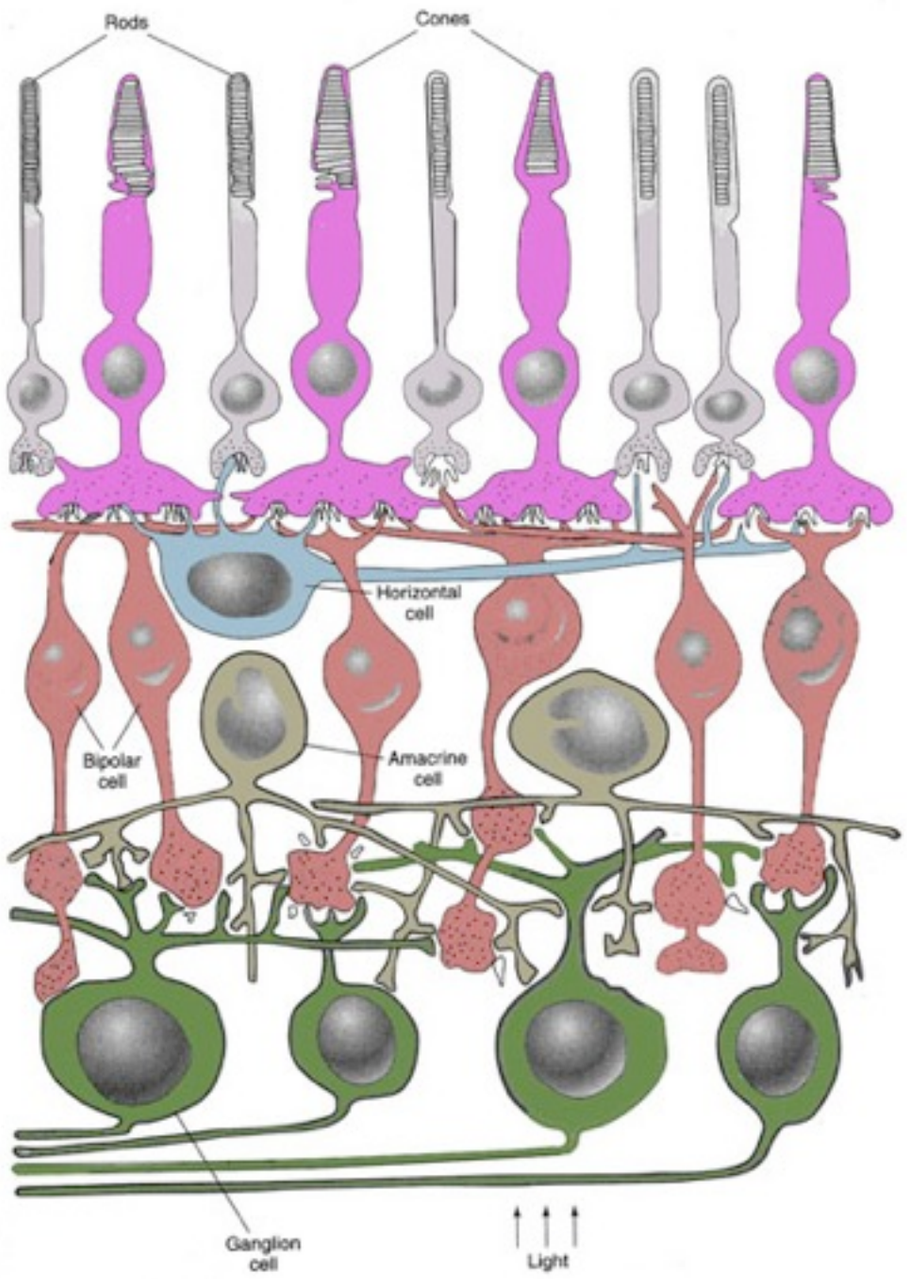
- How can we make sense of the joint activity of many neurons encoding rich (naturalistic) stimuli at the single-spike level?
- Is the "code" independent? If not, what is the nature of the collective activity? Is it combinatorial?
- How can we connect theoretical work on neural coding to new, high-dimensional data?

Overview

1. Start with data
2. Create inverse statistical physics models for the joint activity
3. Are the models good descriptions of the data?
4. Study the behavior of these models, and compare with predicted signatures in the data.

Retina as an encoding device

Retina as an encoding device



Retina:

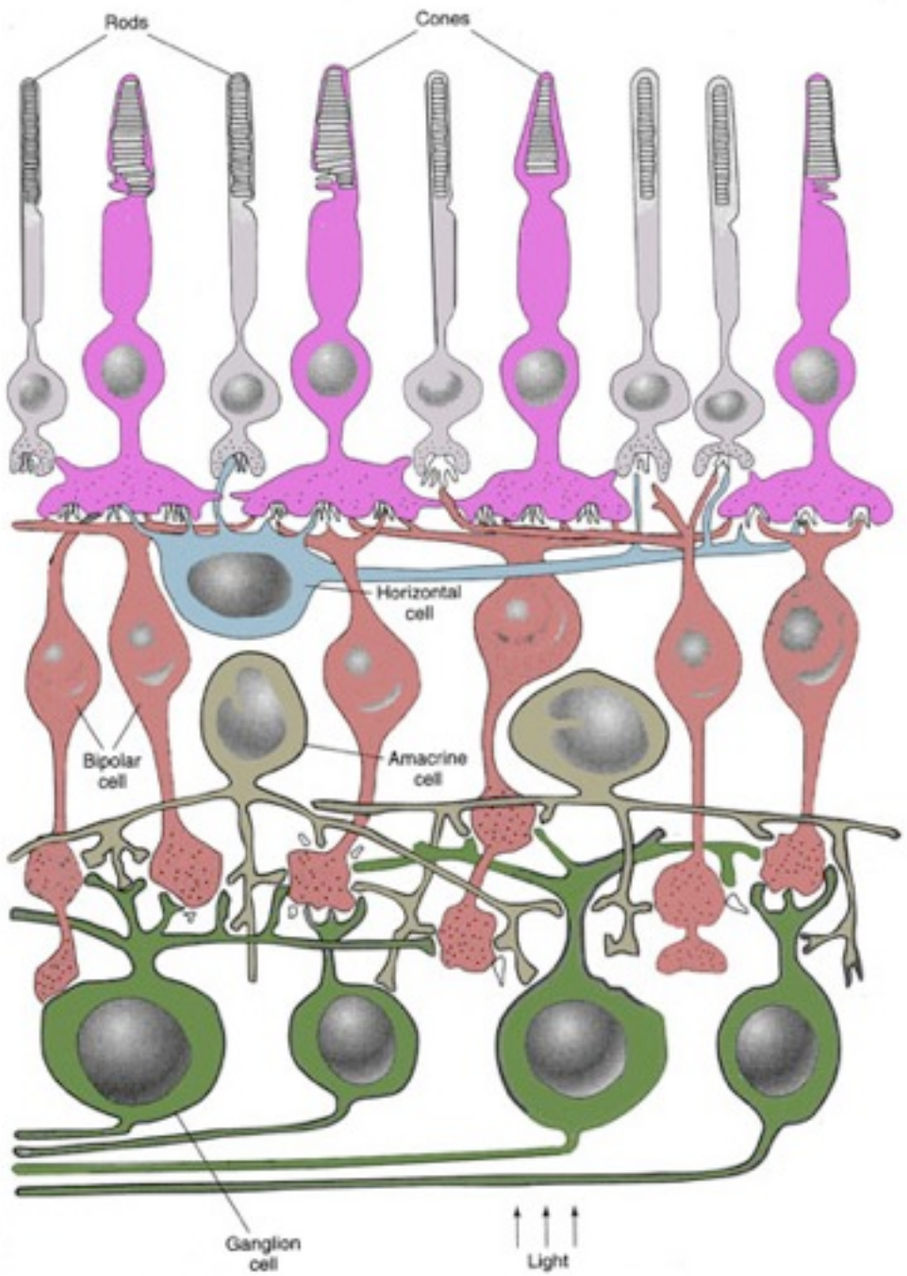
- planar tissue, easy experimental access, no feedback

- **input:** light, transduced by photoreceptors

- **output:** ganglion cell layer patterns of neural activity (spiking / silence)

- what is the mapping between stimuli and neural outputs?

Retina as an encoding device



Retina:

- planar tissue, easy experimental access, no feedback

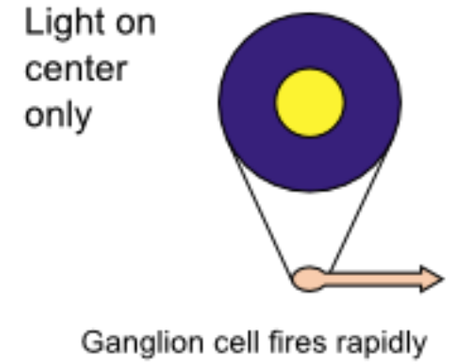
- **input:** light, transduced by photoreceptors

- **output:** ganglion cell layer patterns of neural activity (spiking / silence)

- **what is the mapping between stimuli and neural outputs?**

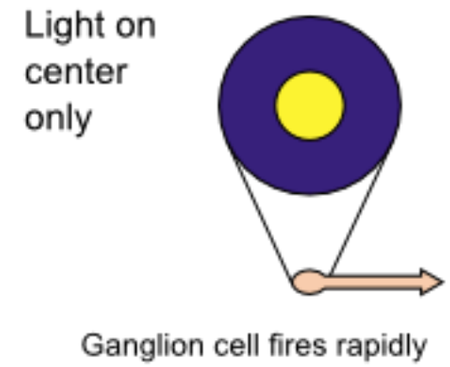
Population coding = one of the basic questions of sensory neuroscience, **especially interesting with complex / naturalistic stimuli.**
Retina is uniquely suitable for probing this question.

Wait, isn't the retina just a camera?



Wait, isn't the retina just a camera?

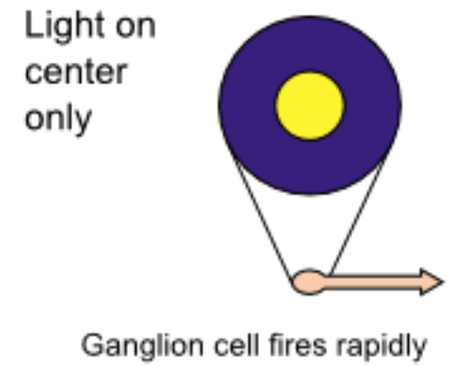
1. Each retinal ganglion cell (RGC) effectively performs center-surround filtering of the light stimulus



Wait, isn't the retina just a camera?

1. Each retinal ganglion cell (RGC) effectively performs center-surround filtering of the light stimulus

2. We know how this is implemented anatomically

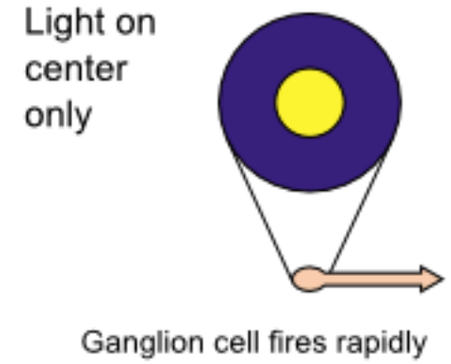


Wait, isn't the retina just a camera?

1. Each retinal ganglion cell (RGC) effectively performs center-surround filtering of the light stimulus

2. We know how this is implemented anatomically

3. Theory tells us why center-surround filtering is efficient (= to decorrelate natural inputs)



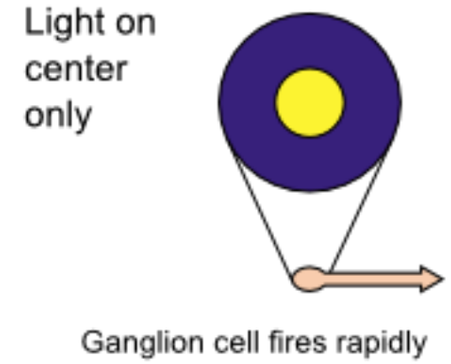
Wait, isn't the retina just a camera?

1. Each retinal ganglion cell (RGC) effectively performs center-surround filtering of the light stimulus

2. We know how this is implemented anatomically

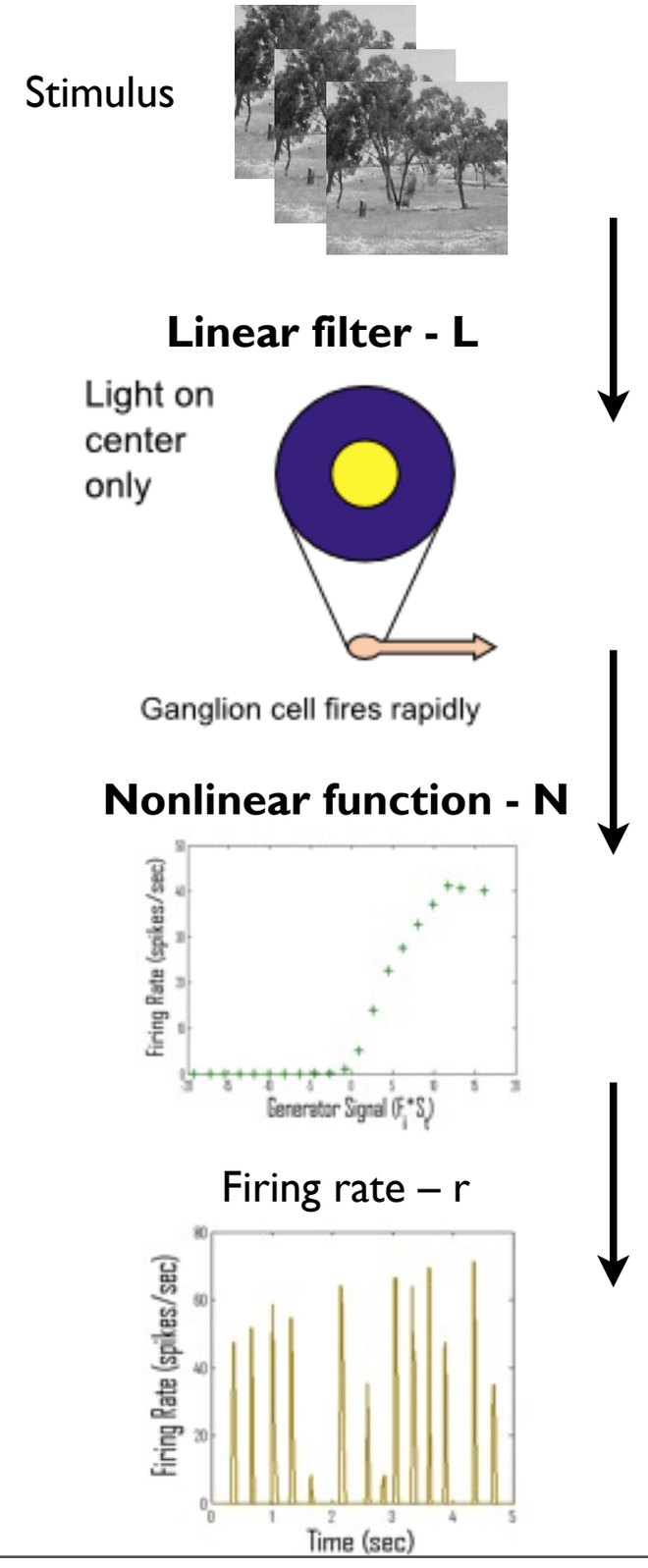
3. Theory tells us why center-surround filtering is efficient (= to decorrelate natural inputs)

4. We can build mathematical models that predict spikes given the stimulus for each cell



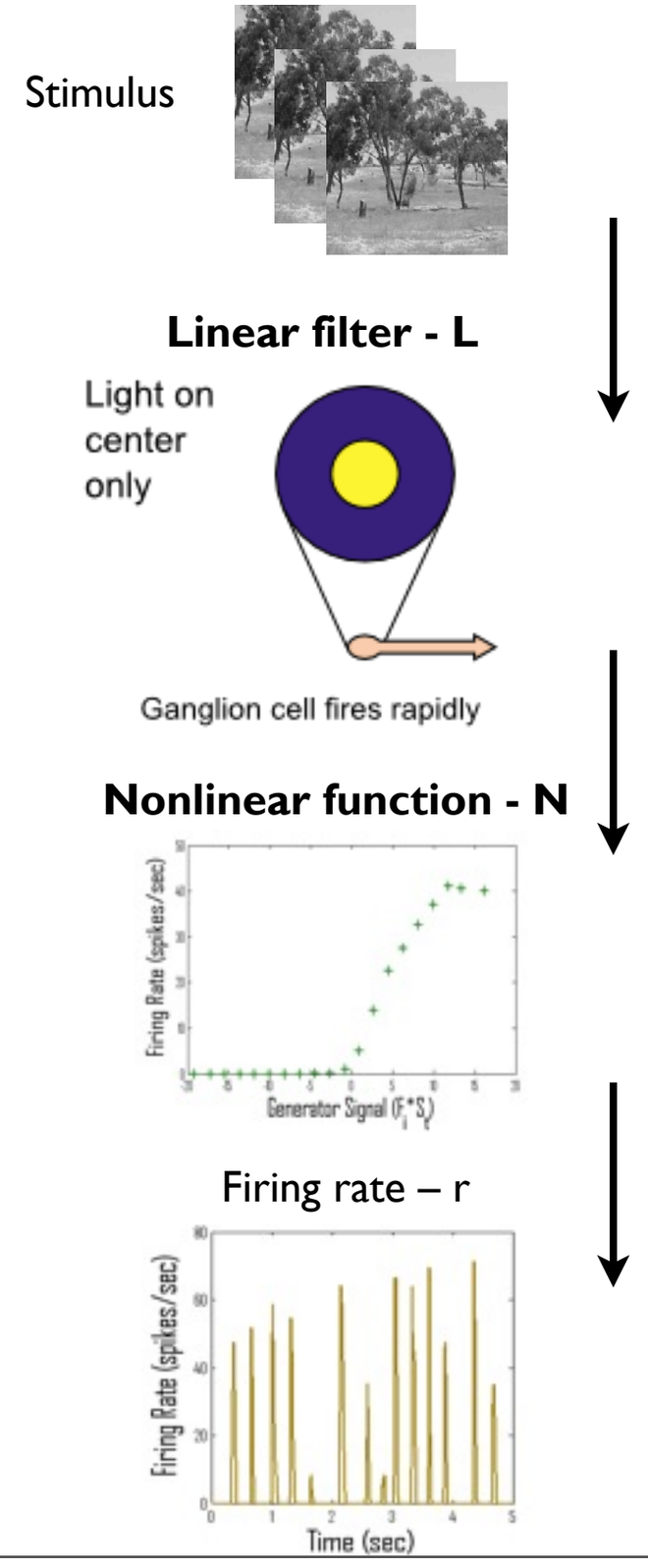
Wait, isn't the retina just a camera?

1. Each retinal ganglion cell (RGC) effectively performs center-surround filtering of the light stimulus
2. We know how this is implemented anatomically
3. Theory tells us why center-surround filtering is efficient (= to decorrelate natural inputs)
4. We can build mathematical models that predict spikes given the stimulus for each cell



Wait, isn't the retina just a camera?

1. Each retinal ganglion cell (RGC) effectively performs center-surround filtering of the light stimulus
2. We know how this is implemented anatomically
3. Theory tells us why center-surround filtering is efficient (= to decorrelate natural inputs)
4. We can build mathematical models that predict spikes given the stimulus for each cell
5. RGCs (or RGC subtypes) tile the visual space

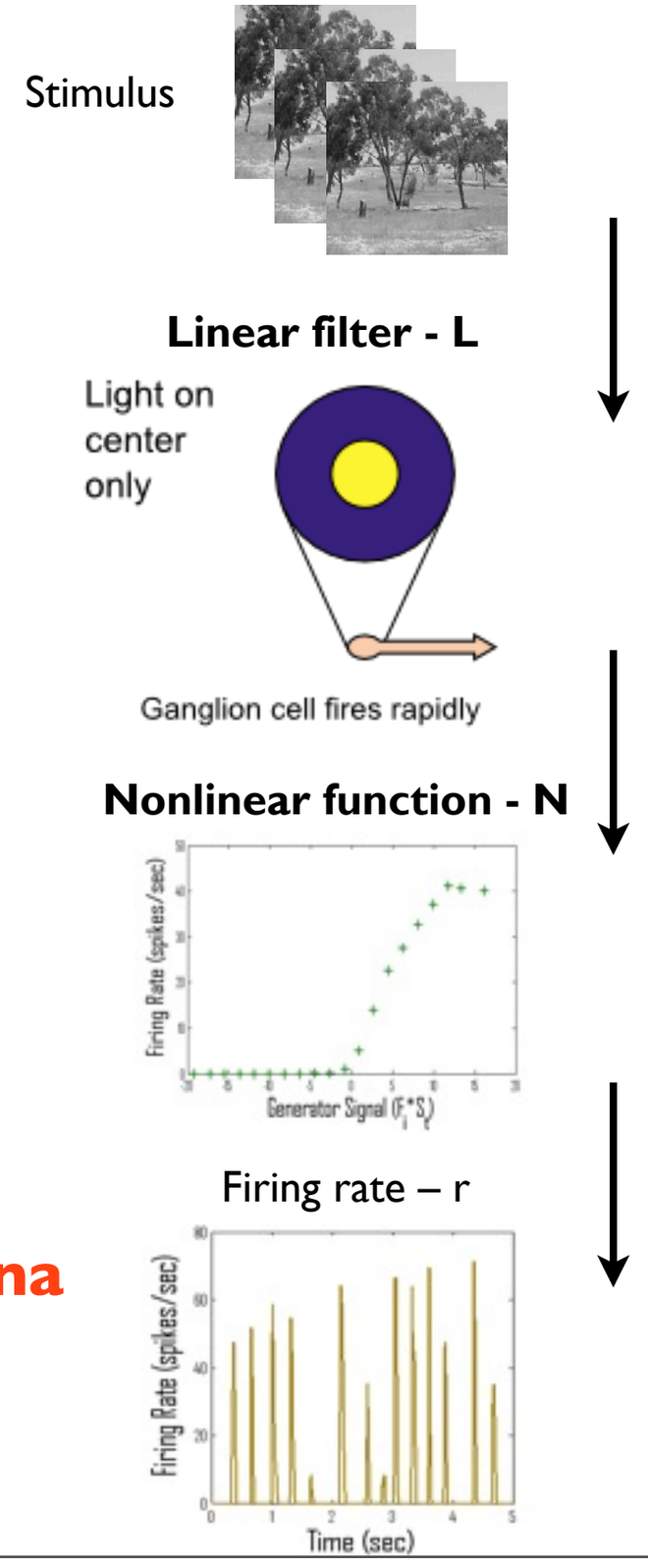


Wait, isn't the retina just a camera?

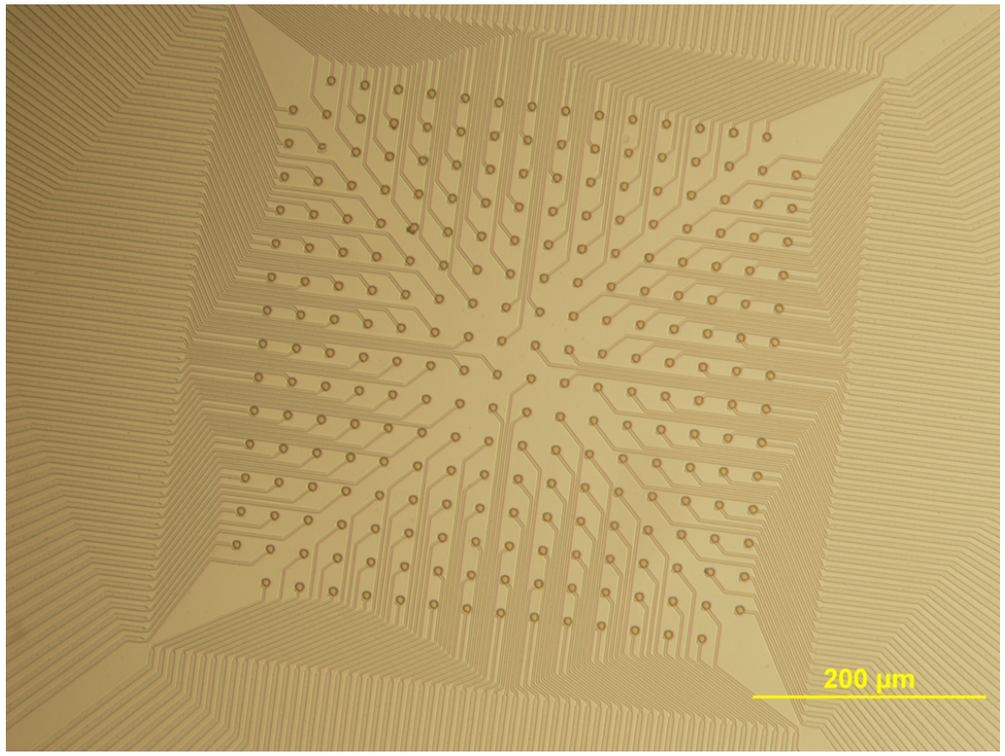
1. Each retinal ganglion cell (RGC) effectively performs center-surround filtering of the light stimulus
2. We know how this is implemented anatomically
3. Theory tells us why center-surround filtering is efficient (= to decorrelate natural inputs)
4. We can build mathematical models that predict spikes given the stimulus for each cell
5. RGCs (or RGC subtypes) tile the visual space

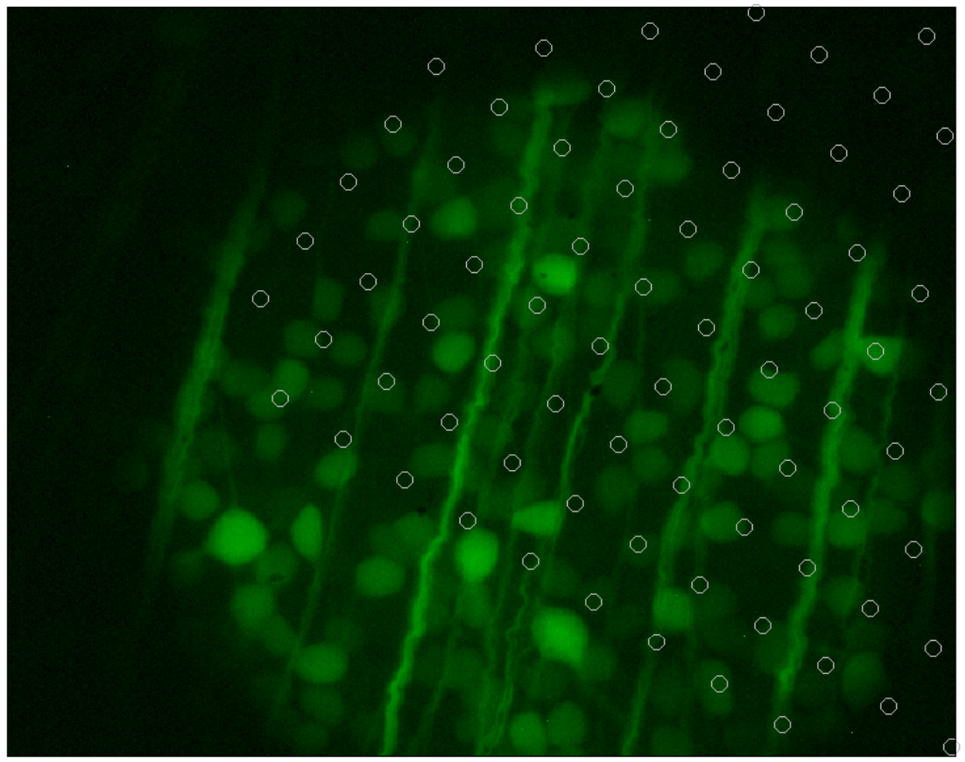
⇒ We *should* know everything about how the retina represents stimulus information

⇒ Prediction: decorrelated spike trains



Studying complete neural populations





O Marre et al, J Neurosci (2012)

- 252 electrode array, dense spacing, salamander retina
- able to record ~200-300 RGCs in a dense patch, overlapping RFs
- >90 % coverage
- This is (almost) a complete population encoding the stimuli in a small visual angle... A rare case in neuroscience!

Watching the retinal output

One ellipse = the receptive field of one recorded cell.

Watching the retinal output

Watching the retinal output

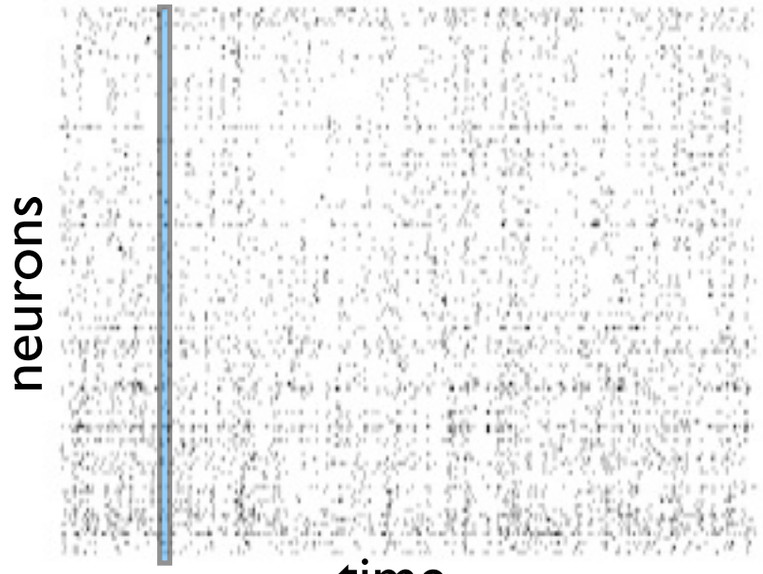
The ellipse appears when the cell fires.

Watching the retinal output

Watching the retinal output

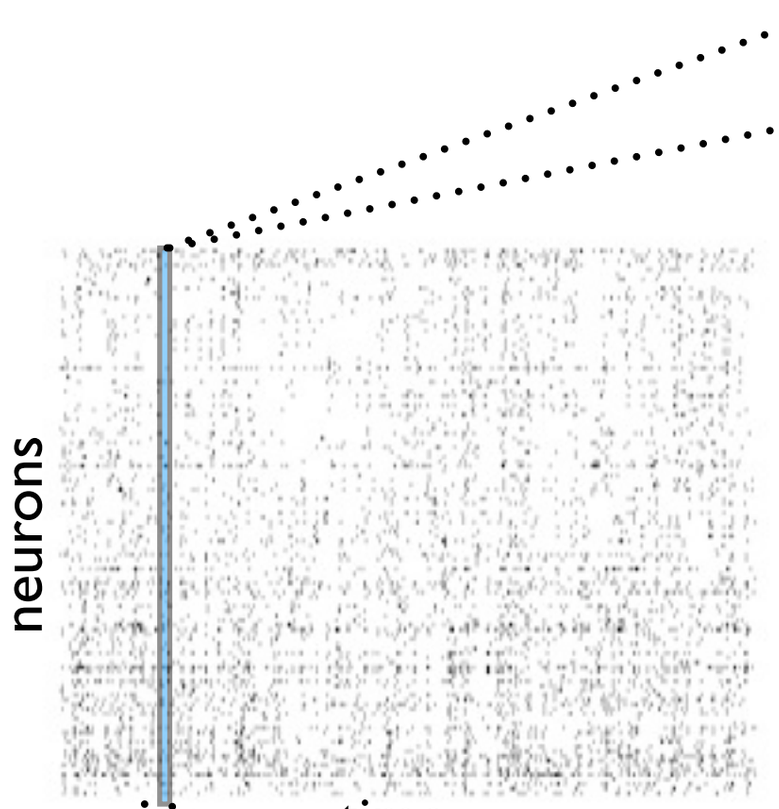
Neural codebook, dictionary, and vocabulary

natural stimulus
P(s)



Neural codebook, dictionary, and vocabulary

natural stimulus
P(s)



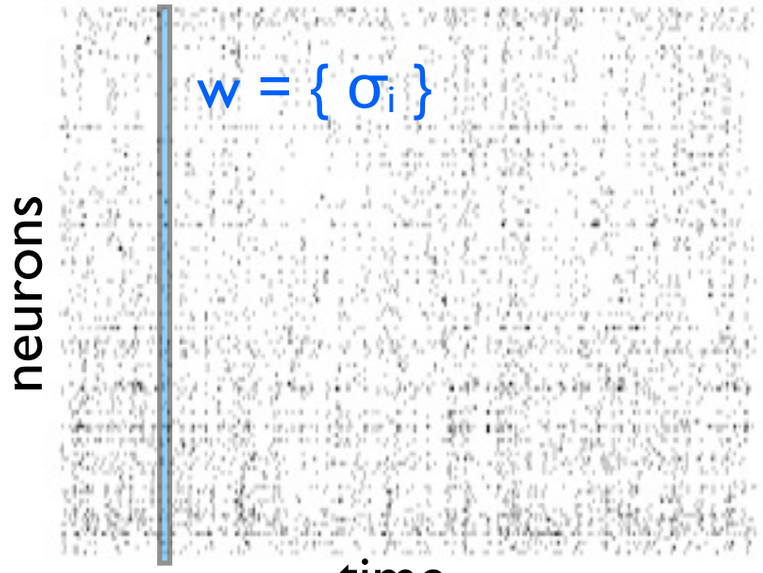
state of the retina at t
$w = \{ \sigma_i \}$

neuron 1	0
	1
	0
	1
	0
	0
	0
	0
	1
...	1
	0
	0
	1
	0
	1
	0
	1
	0
neuron 100	1

codeword in 20 ms bin

Neural codebook, dictionary, and vocabulary

natural stimulus
P(s)



vocabulary

which codewords occur how often?

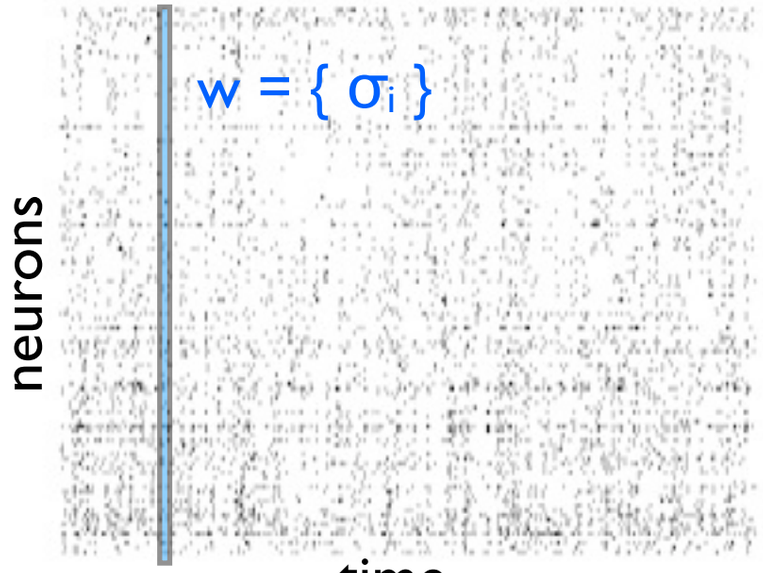
are some codewords "forbidden"?

are the codewords organized into clusters?

P(w)

Neural codebook, dictionary, and vocabulary

natural stimulus
P(s)



vocabulary
which codewords occur how often?
are some codewords "forbidden"?
are the codewords organized into clusters?
P(w)

Major problem: curse of dimensionality. For N neurons we have 2^N possible firing patterns!

Maxent models for the neural vocabulary, $P(\{\sigma_i\})$

What is the overall structure of retinal codewords $\{\sigma_i\}$? Which population states are frequent, which ones are "forbidden"...? This is encoded in $P(\{\sigma_i\})$ -- but how can we find this distribution from (finite) data?

Maxent models for the neural vocabulary, P({σi})

What is the overall structure of retinal codewords {σi}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σi}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σi }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

Maxent models for the neural vocabulary, P({σ_i})

What is the overall structure of retinal codewords {σ_i}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σ_i}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σ_i }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

Maxent models for the neural vocabulary, P({σ_i})

What is the overall structure of retinal codewords {σ_i}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σ_i}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σ_i }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

We look for a distribution that is as random as possible (max entropy), but matches the ensemble averages of f to their expectations in data

$$P(\{\sigma_i\}) = \frac{1}{Z(\{g_\mu\})} \exp\left(-\sum_{\mu=1}^{L} g_\mu f_\mu(\{\sigma_i\})\right)$$

Maxent models for the neural vocabulary, P({σi})

What is the overall structure of retinal codewords {σi}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σi}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σi }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

We look for a distribution that is as random as possible (max entropy), but matches the ensemble averages of f to their expectations in data

$$P(\{\sigma_i\}) = \frac{1}{Z(\{g_\mu\})} \exp\left(-\sum_{\mu=1}^{L} g_\mu f_\mu(\{\sigma_i\}) \right)$$

Maxent models for the neural vocabulary, P({σi})

What is the overall structure of retinal codewords {σi}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σi}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σi }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

We look for a distribution that is as random as possible (max entropy), but matches the ensemble averages of f to their expectations in data

$$P(\{\sigma_i\}) = \frac{1}{Z(\{g_\mu\})} \exp\left(-\sum_{\mu=1}^{L} g_\mu f_\mu(\{\sigma_i\}) \right)$$

Independent neurons $\{f_\mu\} = \{\sigma_i\}$

Maxent models for the neural vocabulary, P({σi})

What is the overall structure of retinal codewords {σi}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σi}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σi }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

We look for a distribution that is as random as possible (max entropy), but matches the ensemble averages of f to their expectations in data

$$P(\{\sigma_i\}) = \frac{1}{Z(\{g_\mu\})} \exp\left(-\sum_{\mu=1}^{L} g_\mu f_\mu(\{\sigma_i\}) \right)$$

Independent neurons $\{f_\mu\} = \{\sigma_i\}$

Pairwise (Ising-like) models $\{f_\mu\} = \{\sigma_i, \sigma_i\sigma_j\}$

Maxent models for the neural vocabulary, P({σi})

What is the overall structure of retinal codewords {σi}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σi}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σi }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

We look for a distribution that is as random as possible (max entropy), but matches the ensemble averages of f to their expectations in data

$$P(\{\sigma_i\}) = \frac{1}{Z(\{g_\mu\})} \exp\left(-\sum_{\mu=1}^{L} g_\mu f_\mu(\{\sigma_i\}) \right)$$

Independent neurons $\{f_\mu\} = \{\sigma_i\}$

Pairwise (Ising-like) models $\{f_\mu\} = \{\sigma_i, \sigma_i\sigma_j\}$

Schneidman et al, Nature 440 (2006)
GT et al, arxiv.org:q-bio/0611072 (2006)

Maxent models for the neural vocabulary, P({σi})

What is the overall structure of retinal codewords {σi}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σi}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σi }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

We look for a distribution that is as random as possible (max entropy), but matches the ensemble averages of f to their expectations in data

$$P(\{\sigma_i\}) = \frac{1}{Z(\{g_\mu\})} \exp\left(-\sum_{\mu=1}^{L} g_\mu f_\mu(\{\sigma_i\}) \right)$$

Independent neurons $\{f_\mu\} = \{\sigma_i\}$

Pairwise (Ising-like) models $\{f_\mu\} = \{\sigma_i, \sigma_i\sigma_j\}$

K-spike models $\{f_\mu\} = \{\delta(k - \sum_i \sigma_i)\}$

Schneidman et al, Nature 440 (2006)
GT et al, arxiv.org:q-bio/0611072 (2006)

Maxent models for the neural vocabulary, P({σi})

What is the overall structure of retinal codewords {σi}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σi}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σi }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

We look for a distribution that is as random as possible (max entropy), but matches the ensemble averages of f to their expectations in data

$$P(\{\sigma_i\}) = \frac{1}{Z(\{g_\mu\})} \exp\left(-\sum_{\mu=1}^{L} g_\mu f_\mu(\{\sigma_i\}) \right)$$

Independent neurons	$\{f_\mu\} = \{\sigma_i\}$	Schneidman et al, Nature 440 (2006) GT et al, arxiv.org:q-bio/0611072 (2006)
Pairwise (Ising-like) models	$\{f_\mu\} = \{\sigma_i, \sigma_i\sigma_j\}$	
K-spike models	$\{f_\mu\} = \{\delta(k - \sum_i \sigma_i)\}$	GT. et al, J Stat Mech P03011 (2013)

Maxent models for the neural vocabulary, P({σi})

What is the overall structure of retinal codewords {σi}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σi}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σi }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

We look for a distribution that is as random as possible (max entropy), but matches the ensemble averages of f to their expectations in data

$$P(\{\sigma_i\}) = \frac{1}{Z(\{g_\mu\})} \exp\left(-\sum_{\mu=1}^{L} g_\mu f_\mu(\{\sigma_i\}) \right)$$

Independent neurons	$\{f_\mu\} = \{\sigma_i\}$	Schneidman et al, Nature 440 (2006) GT et al, arxiv.org:q-bio/0611072 (2006)
Pairwise (Ising-like) models	$\{f_\mu\} = \{\sigma_i, \sigma_i\sigma_j\}$	
K-spike models	$\{f_\mu\} = \{\delta(k - \sum \sigma_i)\}$	GT. et al, J Stat Mech P03011 (2013)
K-pairwise models	$\{f_\mu\} = \{\sigma_i, \sigma_i\sigma_j^i, \delta(k - \sum_i \sigma_i)\}$	

Maxent models for the neural vocabulary, P({σi})

What is the overall structure of retinal codewords {σi}? Which population states are frequent, which ones are "forbidden"...? This is encoded in P({σi}) -- but how can we find this distribution from (finite) data?

We choose L functions of the state of the system of N neurons { σi }:

$$\{f_\mu(\sigma_1, \ldots, \sigma_N)\}$$

such that their average values can be estimated from recorded data

$$\langle f_1 \rangle, \langle f_2 \rangle, \ldots, \langle f_\mu \rangle, \ldots, \langle f_L \rangle$$

We look for a distribution that is as random as possible (max entropy), but matches the ensemble averages of f to their expectations in data

$$P(\{\sigma_i\}) = \frac{1}{Z(\{g_\mu\})} \exp\left(-\sum_{\mu=1}^{L} g_\mu f_\mu(\{\sigma_i\}) \right)$$

Independent neurons	$\{f_\mu\} = \{\sigma_i\}$	Schneidman et al, Nature 440 (2006) GT et al, arxiv.org:q-bio/0611072 (2006)
Pairwise (Ising-like) models	$\{f_\mu\} = \{\sigma_i, \sigma_i\sigma_j\}$	
K-spike models	$\{f_\mu\} = \{\delta(k - \sum \sigma_i)\}$	GT. et al, J Stat Mech P03011 (2013)
K-pairwise models	$\{f_\mu\} = \{\sigma_i, \sigma_i\sigma_j^i, \delta(k - \sum_i \sigma_i)\}$	GT. et al, PLOS CB 10 (2014)

What are maxent models (and what they are not)?

What are maxent models (and what they are not)?

They ARE a description of a stationary probability distribution (of responses).

What are maxent models (and what they are not)?

They ARE a description of a stationary probability distribution (of responses).

Interpretation-wise, they ARE NOT identical to equilibrium distributions (note the absence of T). The retina is "driven" by external stimulus.

What are maxent models (and what they are not)?

They ARE a description of a stationary probability distribution (of responses).

Interpretation-wise, they ARE NOT identical to equilibrium distributions (note the absence of T). The retina is "driven" by external stimulus.

The couplings ARE functional (which neurons tend to spike together), but NOT physiological (which neurons are wired together). The output distribution depends on the retina and on the input.

What are maxent models (and what they are not)?

They ARE a description of a stationary probability distribution (of responses).

Interpretation-wise, they ARE NOT identical to equilibrium distributions (note the absence of T). The retina is "driven" by external stimulus.

The couplings ARE functional (which neurons tend to spike together), but NOT physiological (which neurons are wired together). The output distribution depends on the retina and on the input.

They DO NOT assume any dynamics (like Glauber etc); many dynamics could result in the same stationary distribution. Maxent models can be extended to dynamics.

What are maxent models (and what they are not)?

They ARE a description of a stationary probability distribution (of responses).

Interpretation-wise, they ARE NOT identical to equilibrium distributions (note the absence of T). The retina is "driven" by external stimulus.

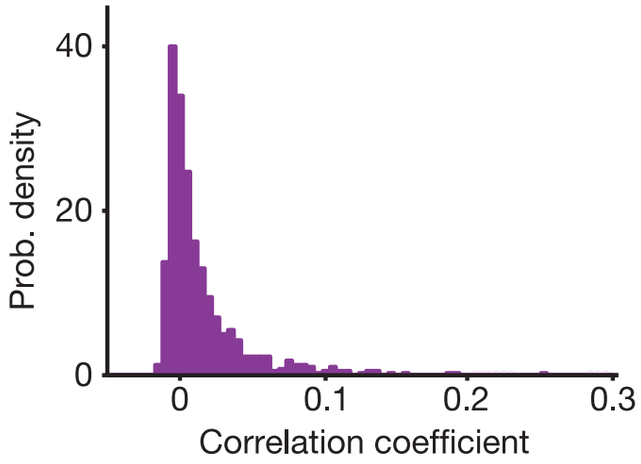
The couplings ARE functional (which neurons tend to spike together), but NOT physiological (which neurons are wired together). The output distribution depends on the retina and on the input.

They DO NOT assume any dynamics (like Glauber etc); many dynamics could result in the same stationary distribution. Maxent models can be extended to dynamics.

Despite these caveats, one can learn a lot about the neural code by studying $P(\{\sigma_i\})$.

Pairwise models for small networks

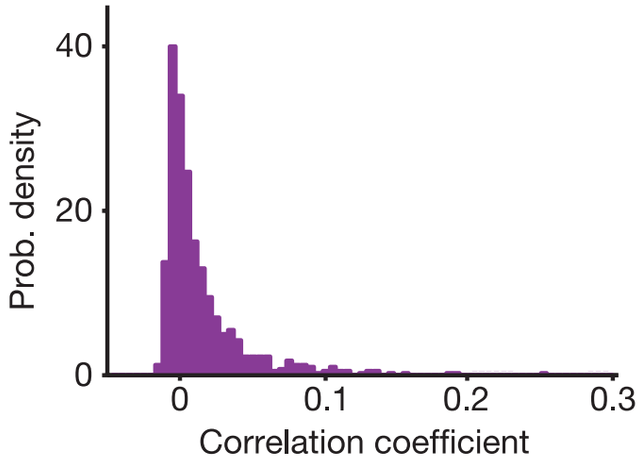
Pairwise correlations between neurons σ_i are weak...



Schneidman et al (2006), 10 neurons

Pairwise models for small networks

Pairwise correlations between neurons σ_i are weak...

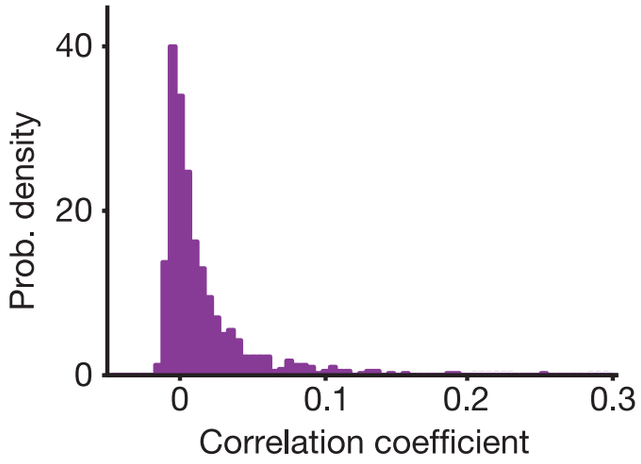


...so perhaps they can be neglected?

Schneidman et al (2006), 10 neurons

Pairwise models for small networks

Pairwise correlations between neurons σ_i are weak...



...so perhaps they can be neglected?

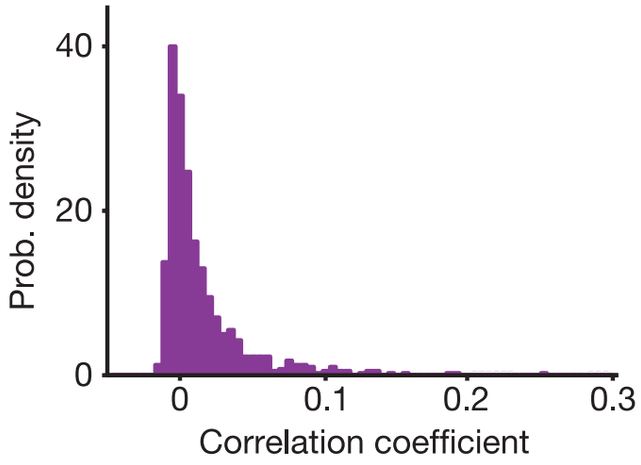
Ising-like models for small networks

A distribution over codewords that reproduces measured mean firing rates and all pairwise correlations

Schneidman et al (2006), 10 neurons

Pairwise models for small networks

Pairwise correlations between neurons σ_i are weak...



Schneidman et al (2006), 10 neurons

...so perhaps they can be neglected?

Ising-like models for small networks

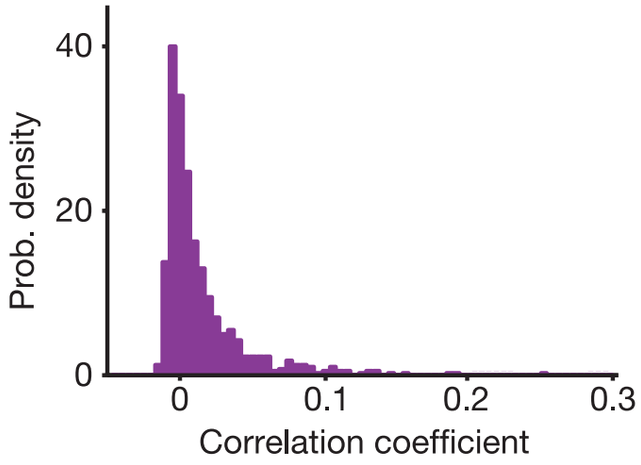
A distribution over codewords that reproduces measured mean firing rates and all pairwise correlations

$$P(\{\sigma_i\}) = Z^{-1} \exp(-E)$$

$$E = -\sum_i h_i \sigma_i - \frac{1}{2} \sum_{i,j} J_{ij} \sigma_i \sigma_j$$

Pairwise models for small networks

Pairwise correlations between neurons σ_i are weak...



...so perhaps they can be neglected?

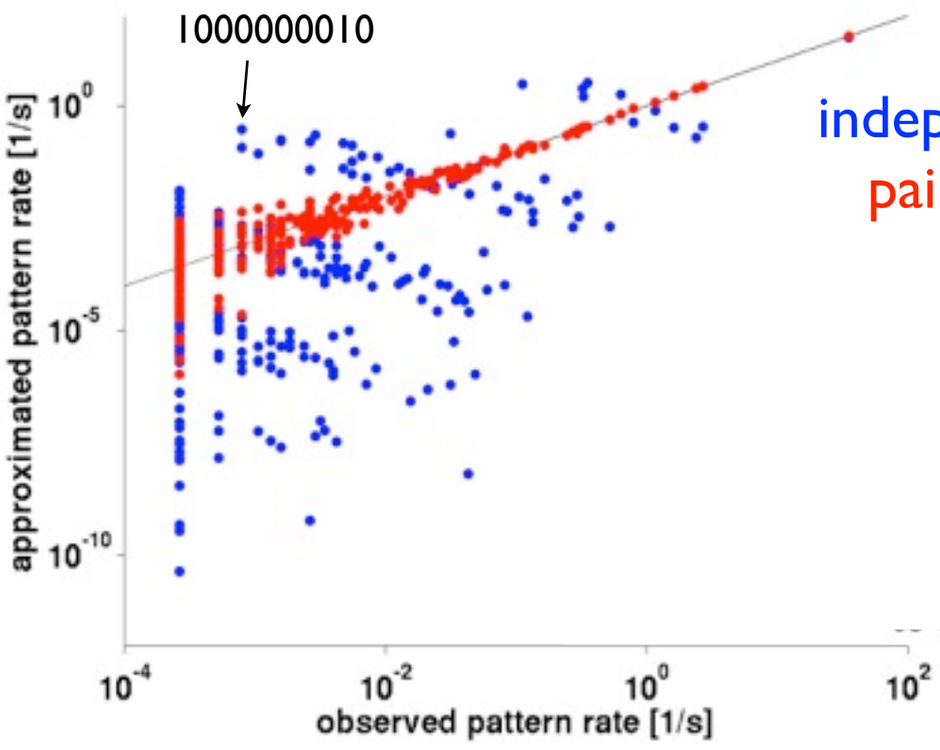
Ising-like models for small networks

A distribution over codewords that reproduces measured mean firing rates and all pairwise correlations

$$P(\{\sigma_i\}) = Z^{-1} \exp(-E)$$

$$E = -\sum_i h_i \sigma_i - \frac{1}{2} \sum_{i,j} J_{ij} \sigma_i \sigma_j$$

Schneidman et al (2006), 10 neurons

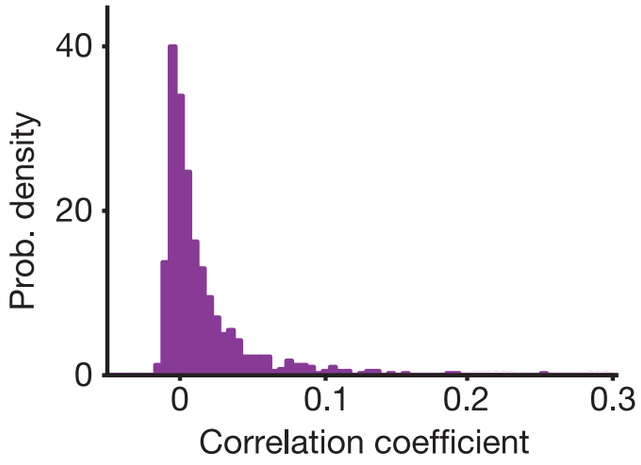


independent

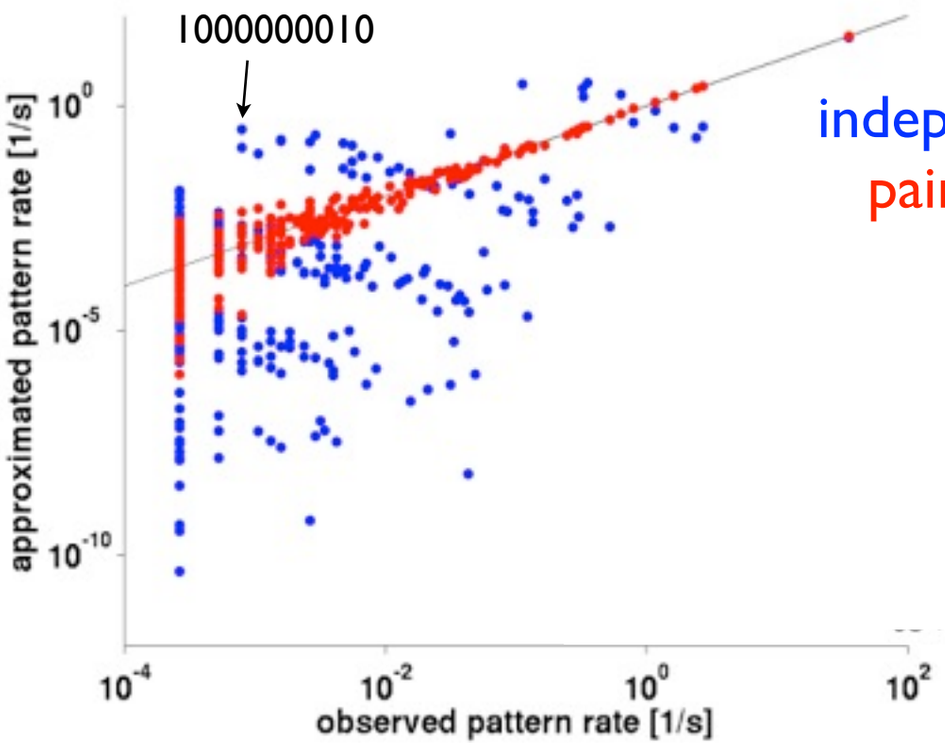
pairwise

Pairwise models for small networks

Pairwise correlations between neurons σ_i are weak...



Schneidman et al (2006), 10 neurons



independent

pairwise

...so perhaps they can be neglected?

Ising-like models for small networks

A distribution over codewords that reproduces measured mean firing rates and all pairwise correlations

$$P(\{\sigma_i\}) = Z^{-1} \exp(-E)$$

$$E = -\sum_i h_i \sigma_i - \frac{1}{2} \sum_{i,j} J_{ij} \sigma_i \sigma_j$$

- Exact solution for the model of $P(\{\sigma_i\})$; check by explicit sampling
- Independent model is a failure
- Pairwise correlations are weak, but their collective effects are strong already for small groups of neurons

Models for > 100 neurons: **pairwise models**

- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N)
  pairwise
  models

Models for > 100 neurons: **pairwise models**

- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N)
  pairwise
  models

constrain N mean rates + N(N-1)/2 covariances

Models for > 100 neurons: **pairwise models**

- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N)

pairwise models

constrain N mean rates + N(N-1)/2 covariances

measured correlations → computed interactions

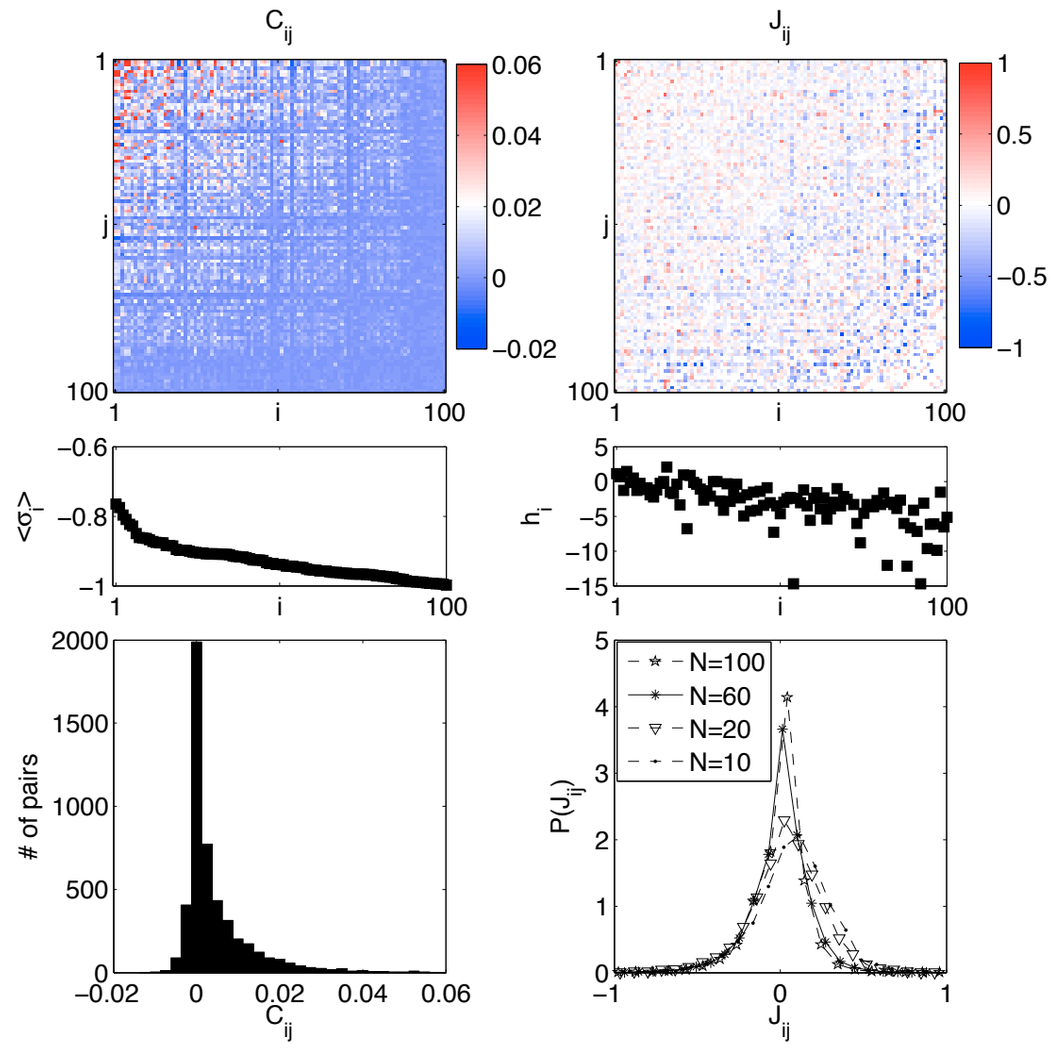
Models for > 100 neurons: **pairwise models**

- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N)

pairwise models

constrain N mean rates + N(N-1)/2 covariances

measured correlations → computed interactions

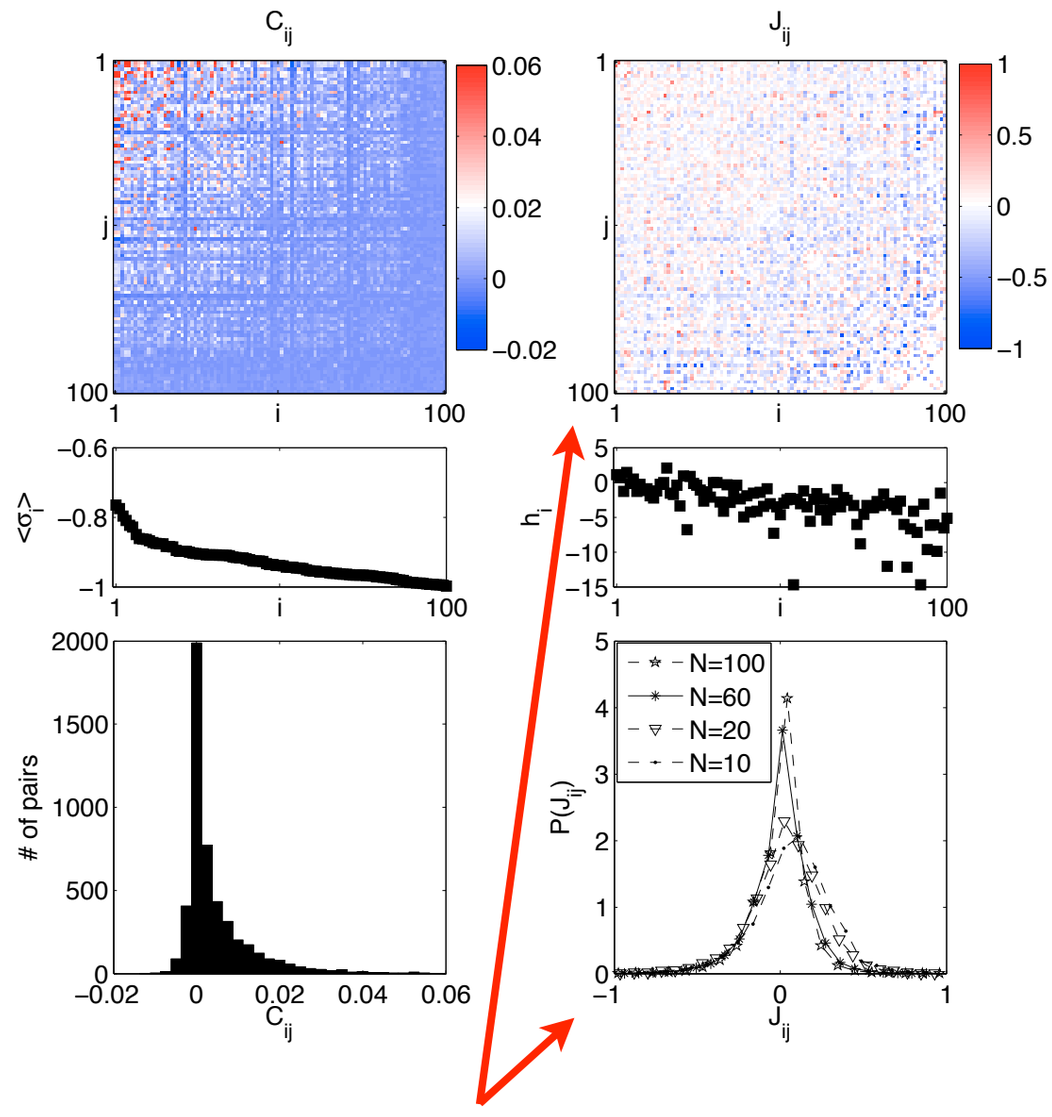


Models for > 100 neurons: **pairwise models**

- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N)
  pairwise models

constrain N mean rates + N(N-1)/2 covariances

measured correlations → computed interactions



note the frustrated interactions + all-to-all connectivity

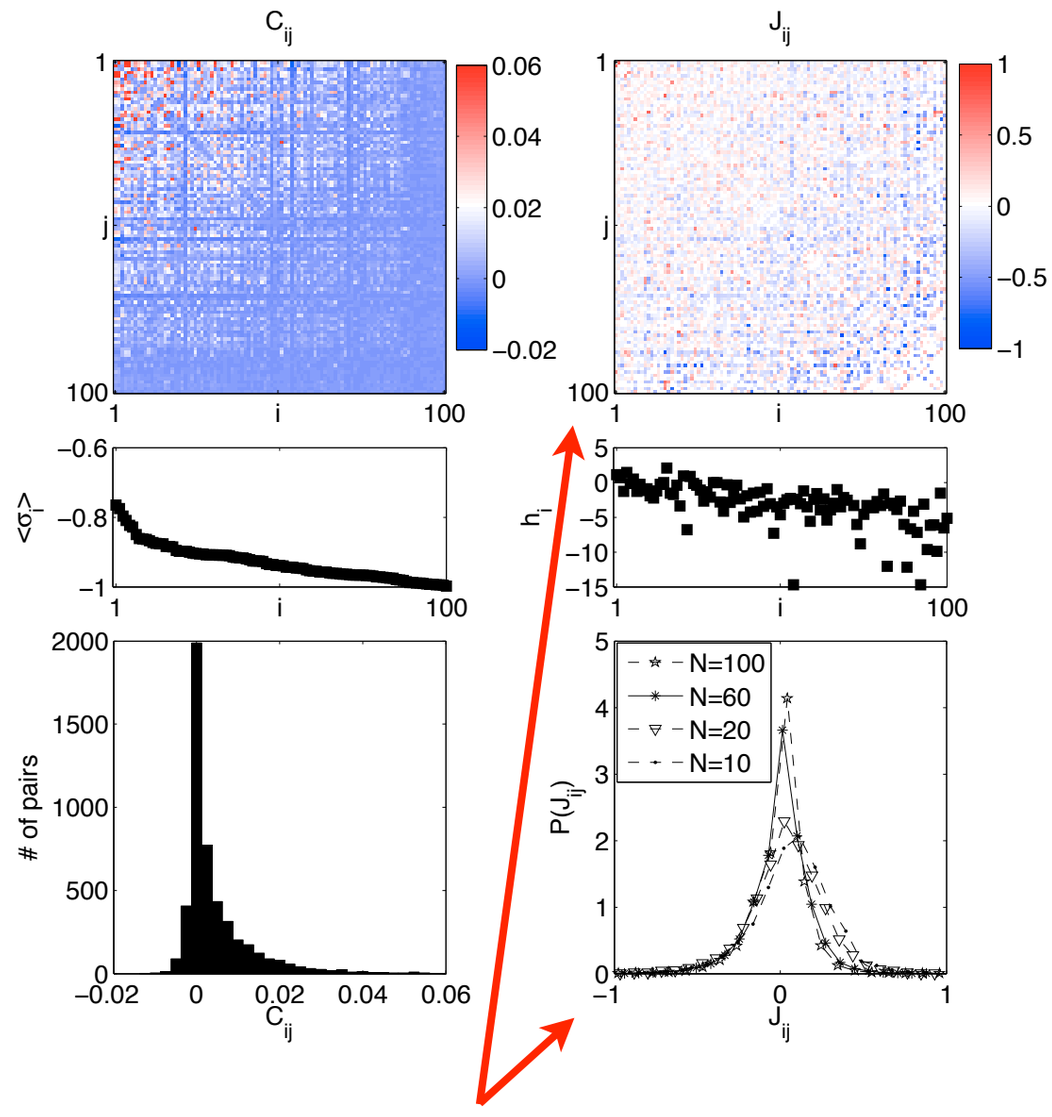
Models for > 100 neurons: **pairwise models**

- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N)
  pairwise
  models

constrain N mean rates + N(N-1)/2 covariances

- the models reconstruct the observables to within exp't error bars; no overfitting

measured correlations → computed interactions



note the frustrated interactions + all-to-all connectivity

Models for > 100 neurons: **pairwise** models are no longer sufficient

- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N)
  pairwise models

constrain N mean rates + N(N-1)/2 covariances

Pairwise models fail to reproduce the distribution of synchronous activity, P(K)

Models for > 100 neurons: **pairwise** models are no longer sufficient

- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N)
  pairwise models

constrain N mean rates + N(N-1)/2 covariances

Pairwise models fail to reproduce the distribution of synchronous activity, P(K)

network size N →

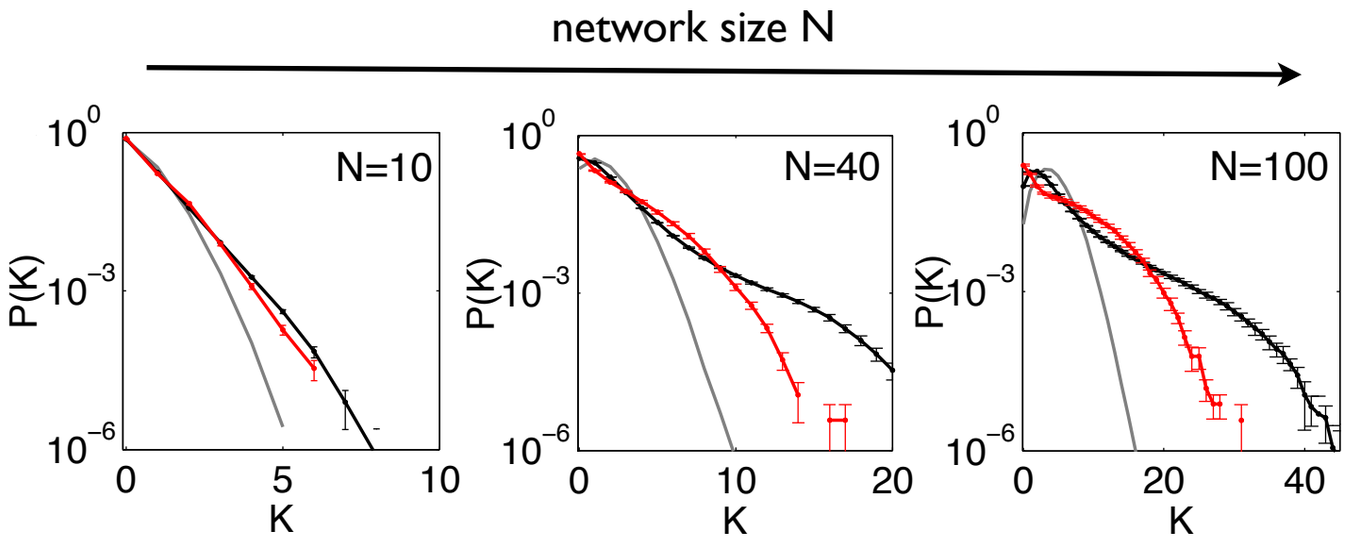
Models for > 100 neurons: **pairwise** models are no longer sufficient

- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N)

pairwise models

constrain N mean rates + N(N-1)/2 covariances

Pairwise models fail to reproduce the distribution of synchronous activity, P(K)



Models for > 100 neurons: **pairwise** models are no longer sufficient

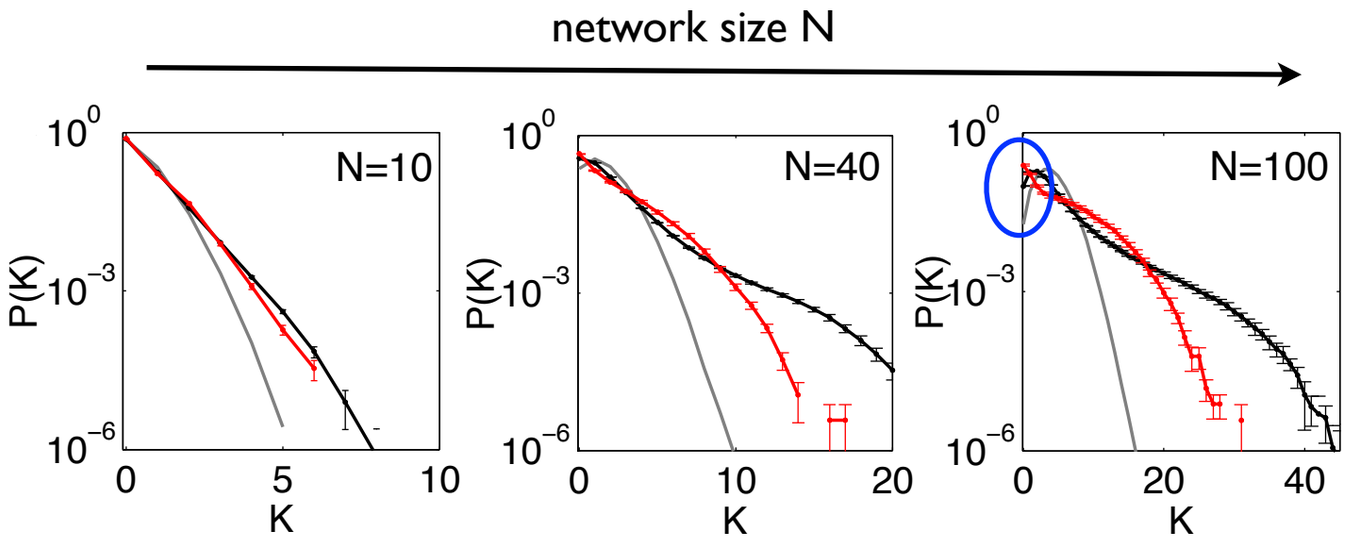
- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N) **pairwise** and **K-pairwise** models

constrain N mean rates + N(N-1)/2 covariances

... + probability of K simultaneous spikes

$$E = -\sum_i h_i \sigma_i - \frac{1}{2}\sum_{i,j} J_{ij}\sigma_i\sigma_j - V(k)$$

Pairwise models fail to reproduce the distribution of synchronous activity, P(K)



Models for > 100 neurons: **pairwise** models are no longer sufficient

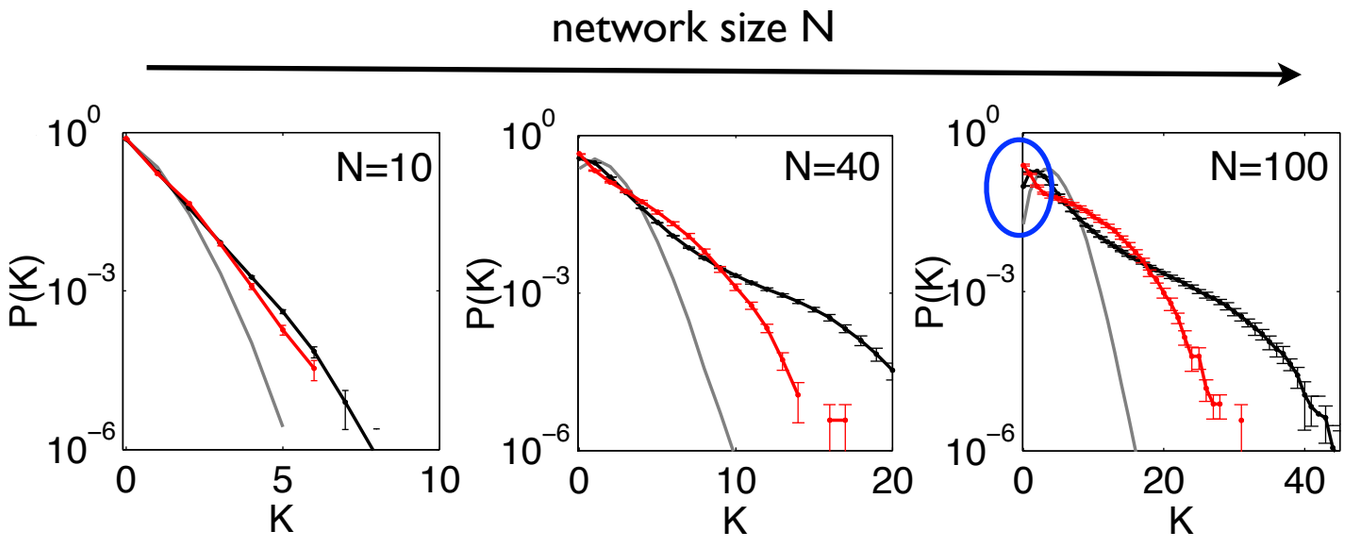
- **stimulus:** ~300 repeats of ~20s fishmovie clip
- **data:** 160 neurons, 20 ms bins (300K binary words)
- **models:** MC reconstruction for N=10, 20, ... 120 neuron subgroups (30 random groups per N) **pairwise** and **K-pairwise** models

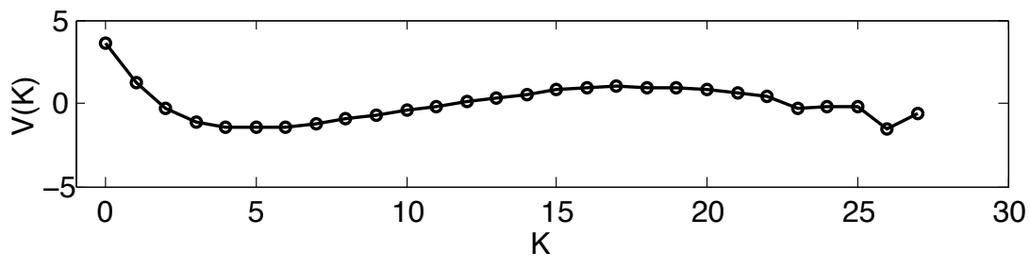
constrain N mean rates + N(N-1)/2 covariances

... + probability of K simultaneous spikes

$$E = -\sum_i h_i \sigma_i - \frac{1}{2}\sum_{i,j} J_{ij}\sigma_i\sigma_j - V(k)$$

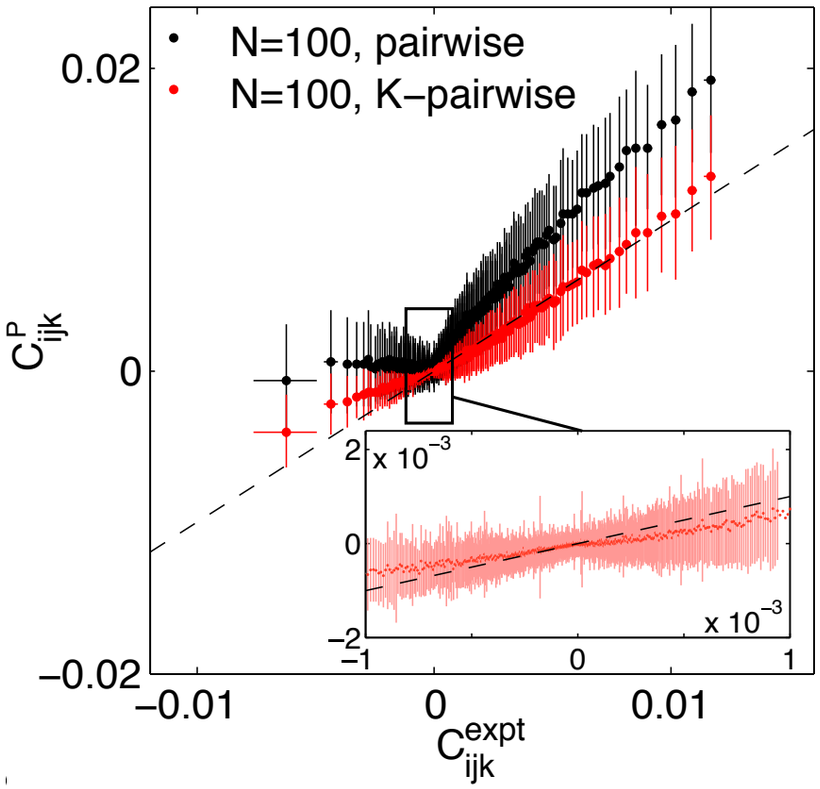
Pairwise models fail to reproduce the distribution of synchronous activity, P(K)



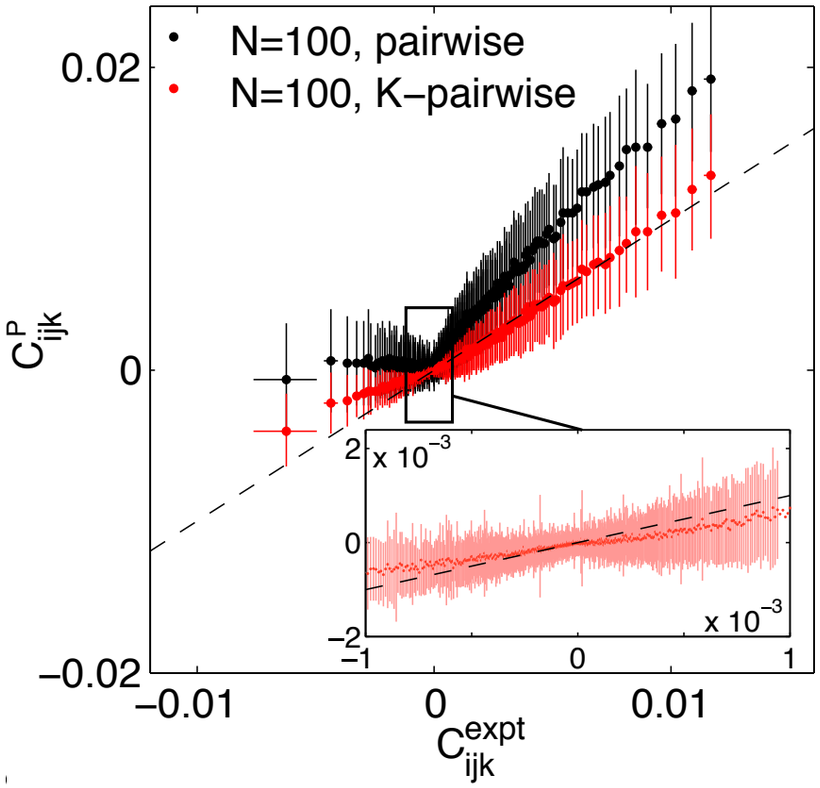


potential V(K) enforces the new P(K) constraint

I: Predicting higher-order statistics

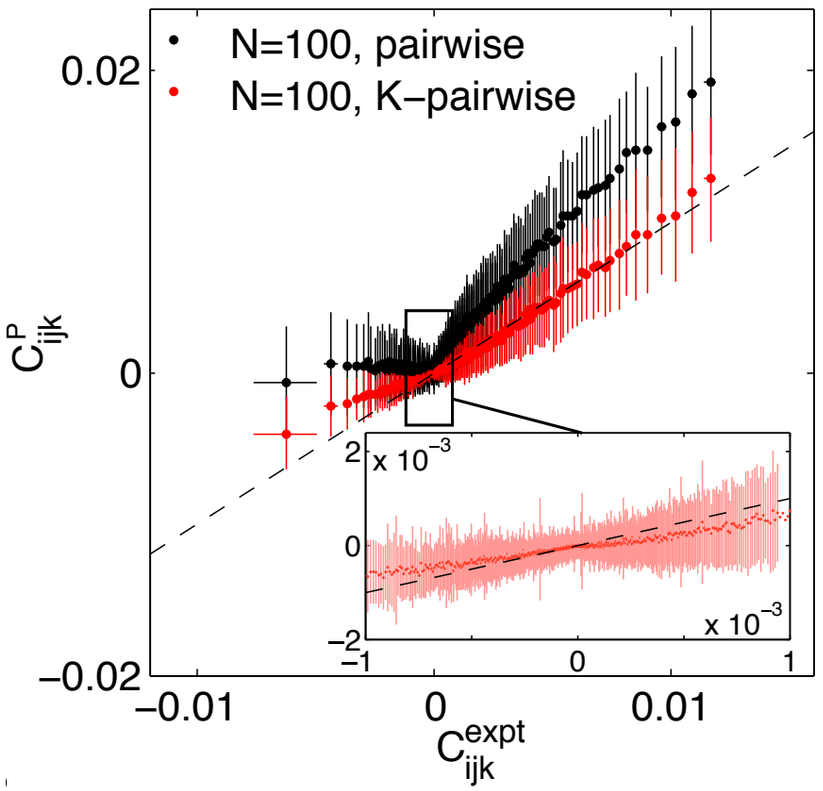


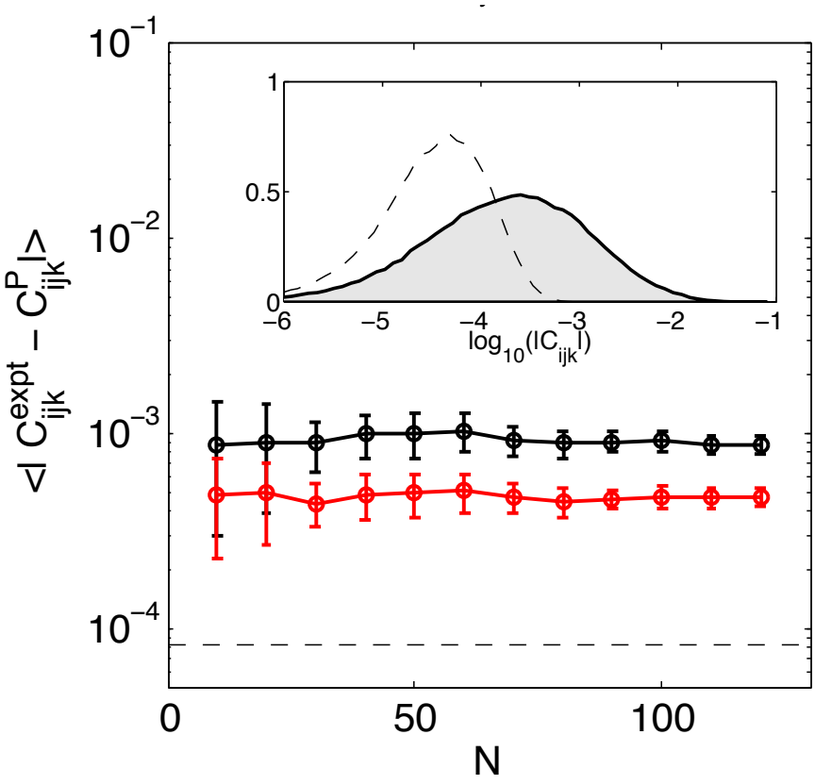
I: Predicting higher-order statistics



- K-pairwise models predict 3-point correlations without significant bias

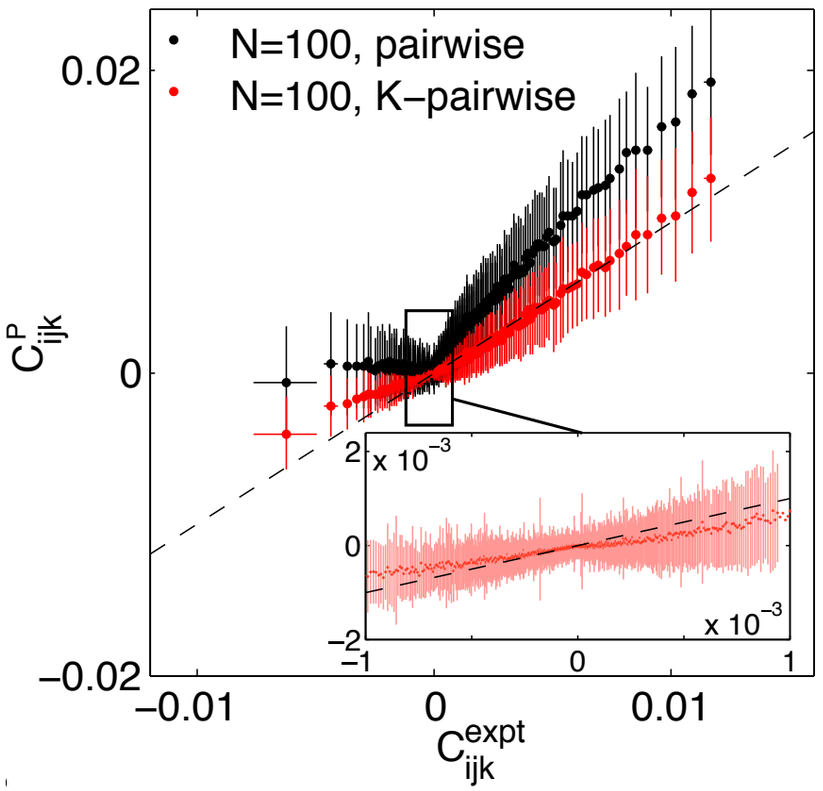
I: Predicting higher-order statistics

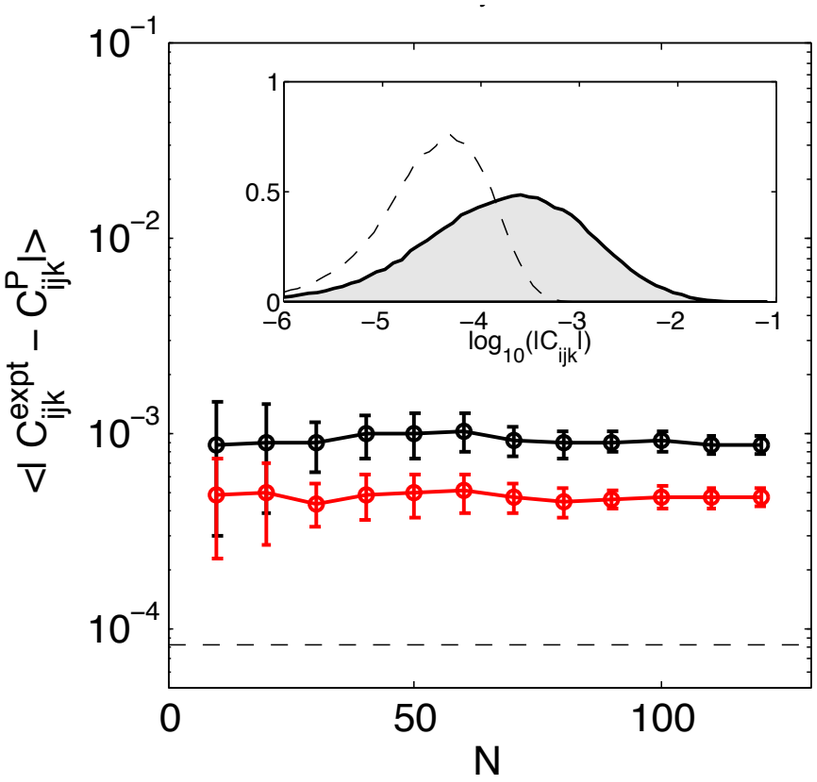




- K-pairwise models predict 3-point correlations without significant bias

I: Predicting higher-order statistics

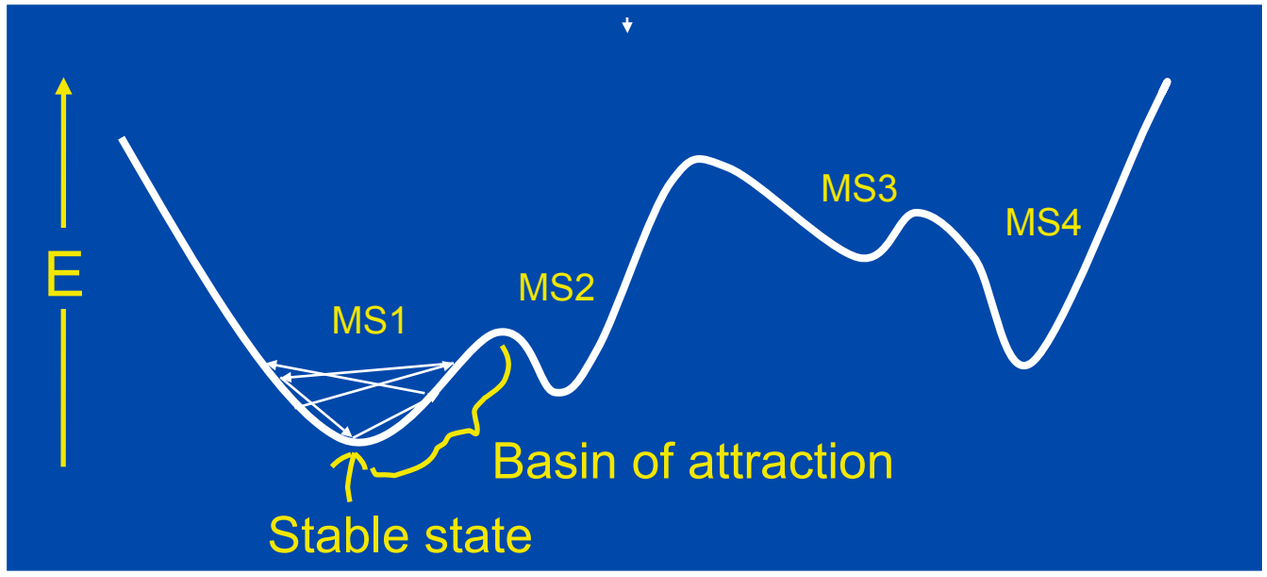




- K-pairwise models predict 3-point correlations without significant bias
- The error in predicted 3-point correlations does not increase with system size (perhaps we are accounting for “hidden” nodes?)

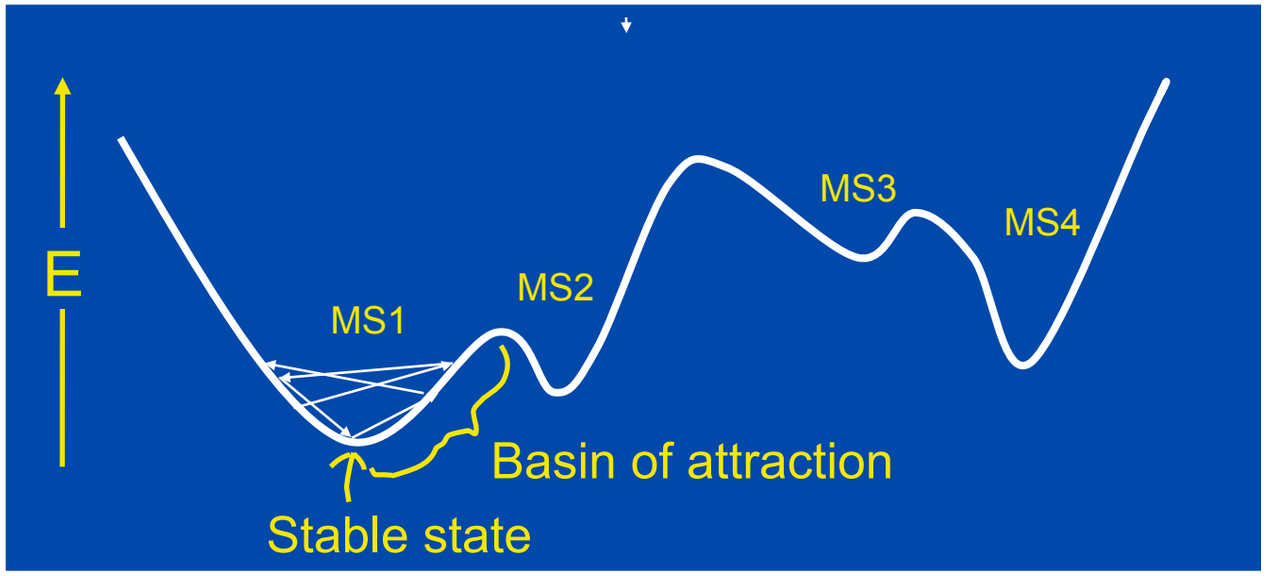
II: Energy landscape

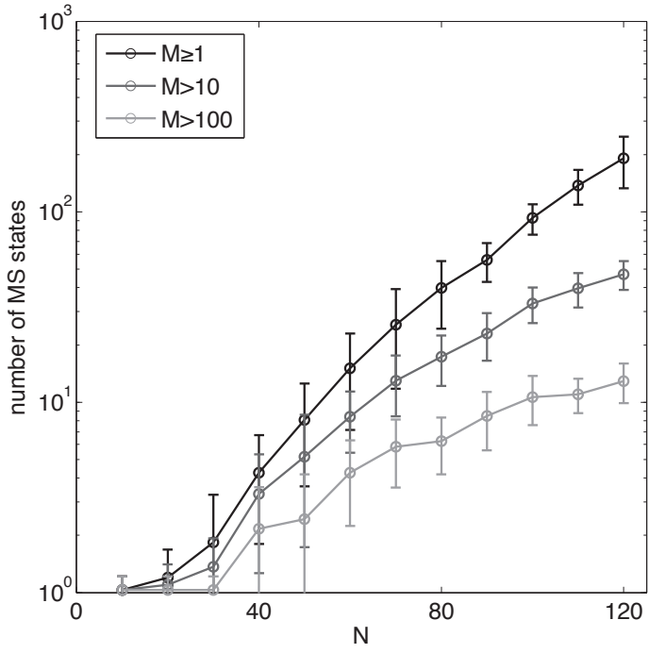
- The energy landscape of the resulting models has many local minima (MS states)
- Stimuli could be encoded by the identity of the basin of attraction and not by the detailed micro-state



II: Energy landscape

- The energy landscape of the resulting models has many local minima (MS states)
- Stimuli could be encoded by the identity of the basin of attraction and not by the detailed micro-state

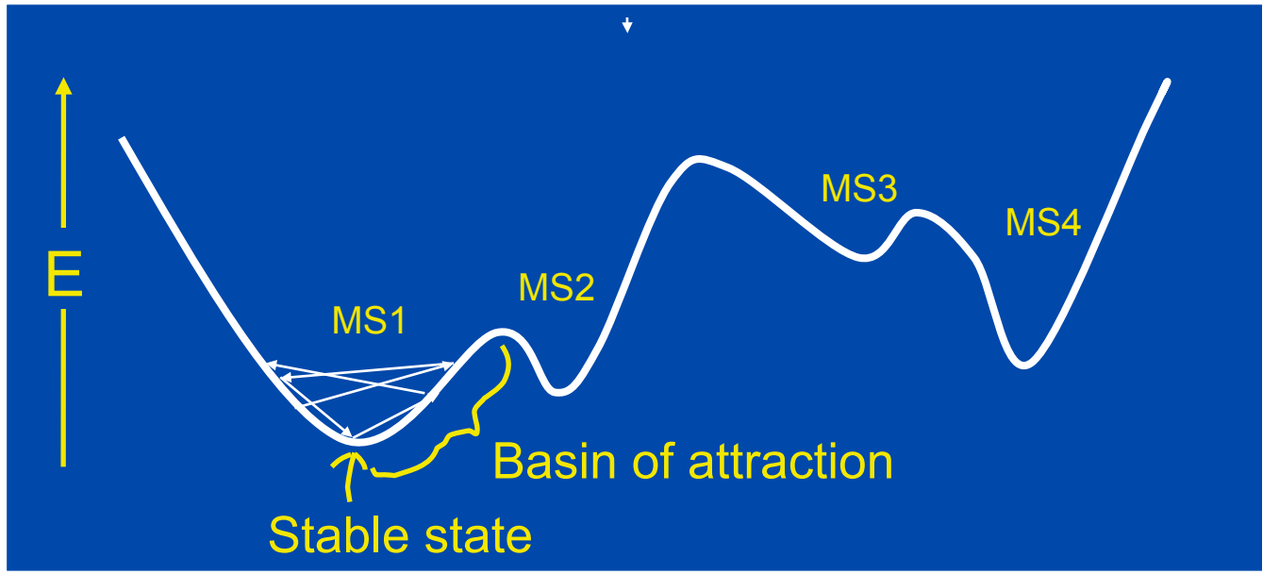


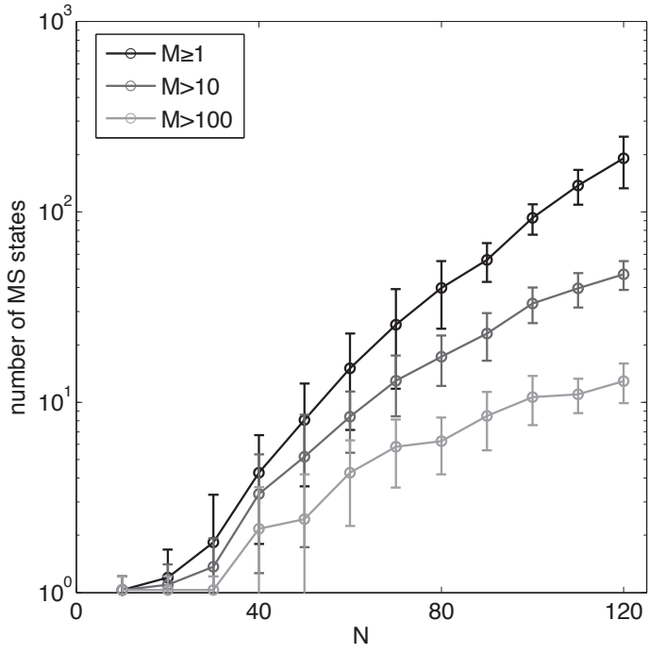


the number of MS states grows with the size of neural network N

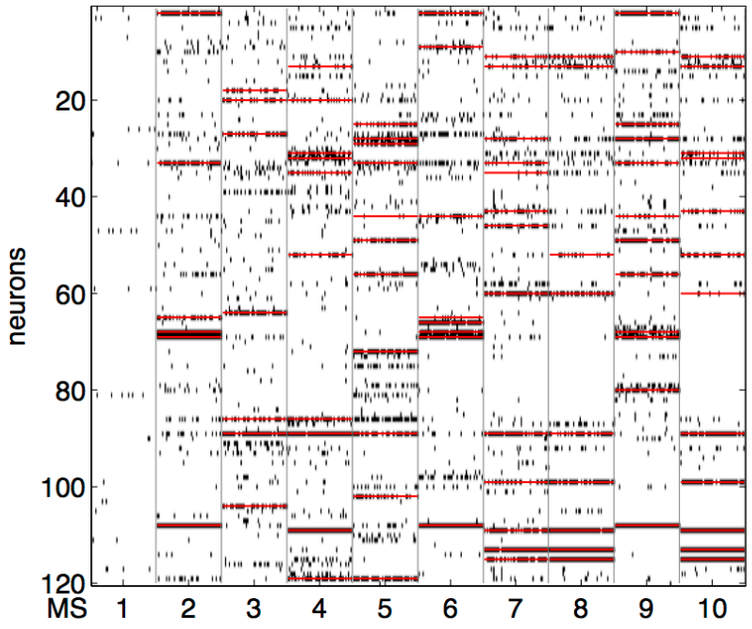
II: Energy landscape

- The energy landscape of the resulting models has many local minima (MS states)
- Stimuli could be encoded by the identity of the basin of attraction and not by the detailed micro-state





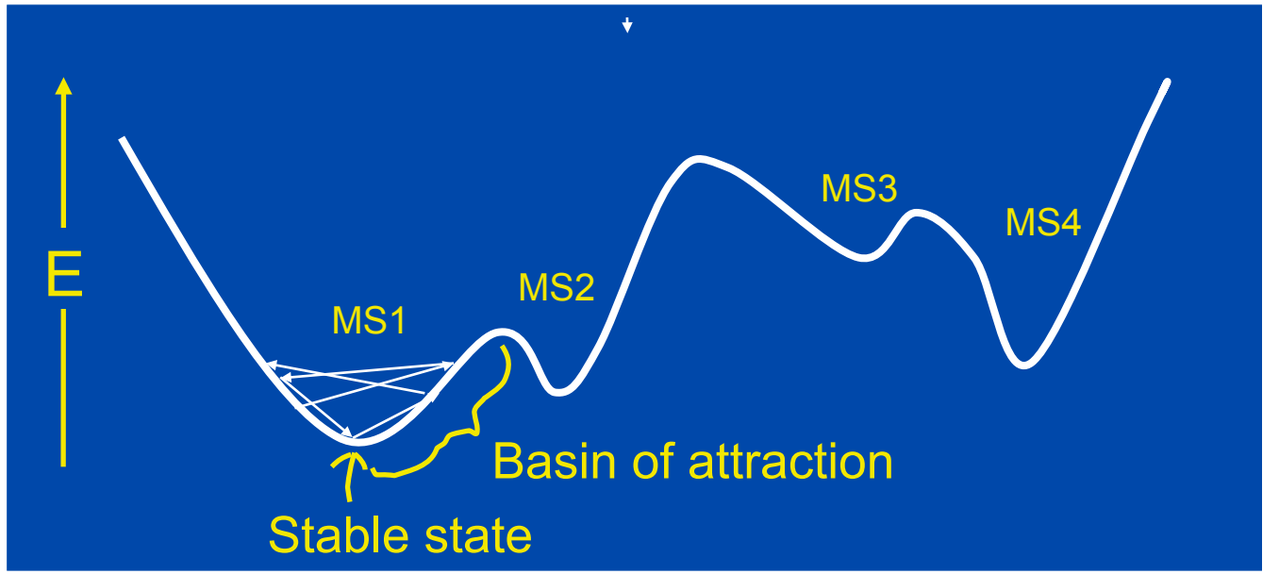
the number of MS states grows with the size of neural network N

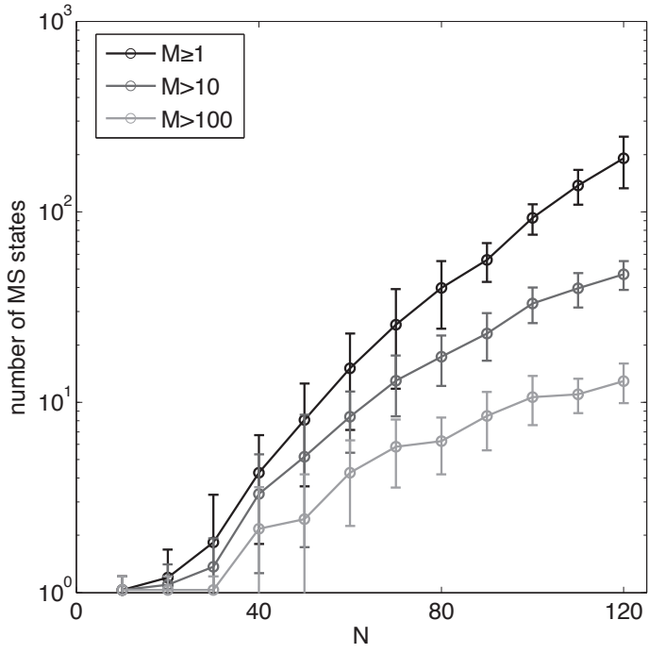


the mapping between micro-states and MS states, extracted from model *alone*

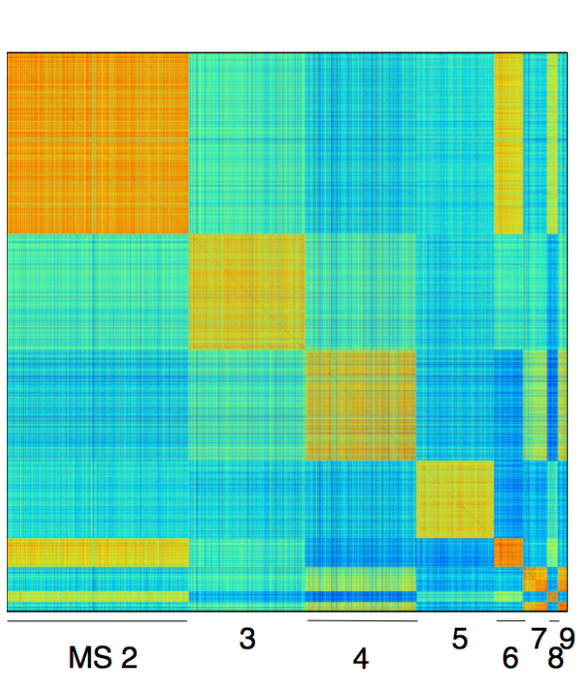
II: Energy landscape

- The energy landscape of the resulting models has many local minima (MS states)
- Stimuli could be encoded by the identity of the basin of attraction and not by the detailed micro-state



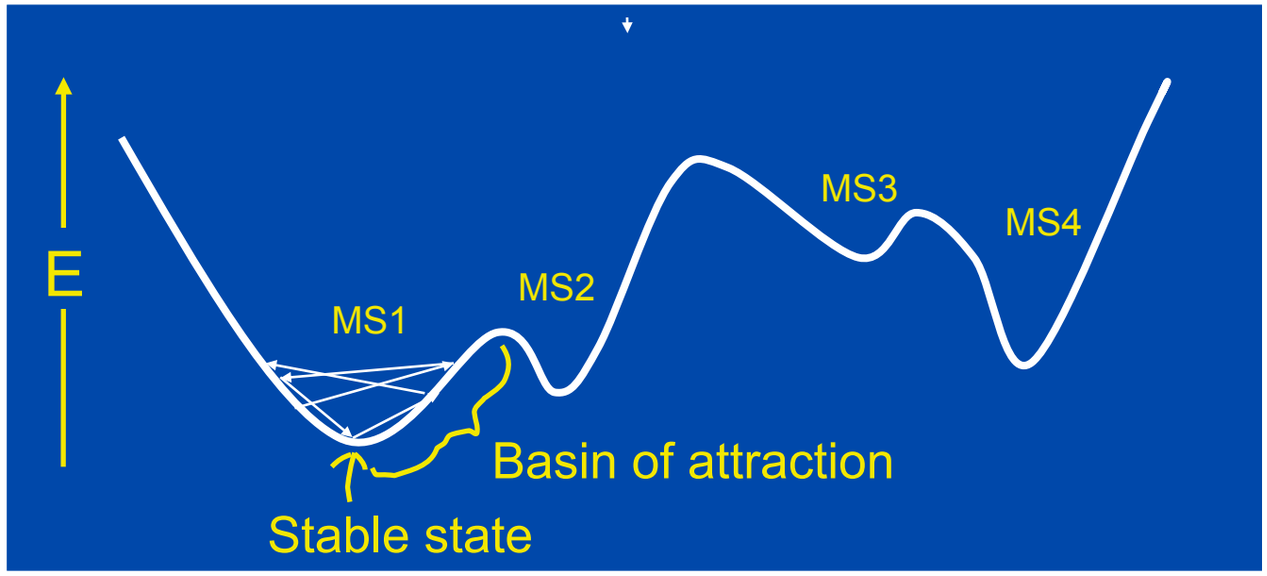


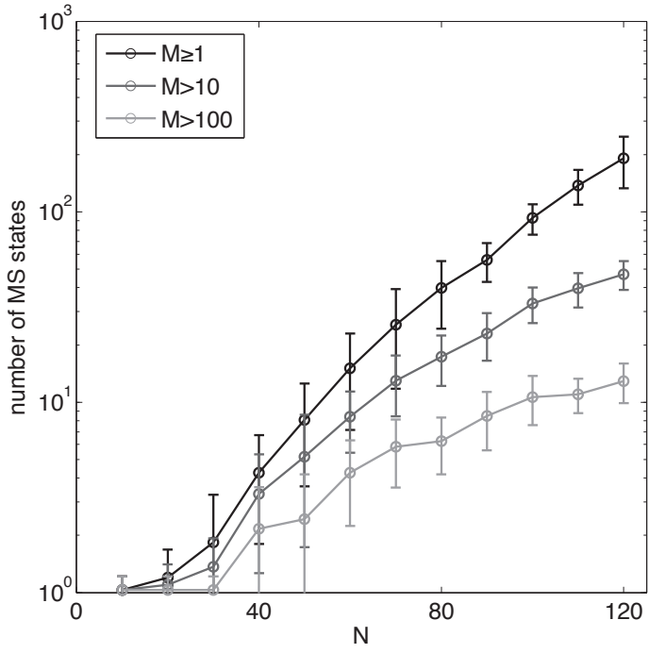
the number of MS states grows with the size of neural network N



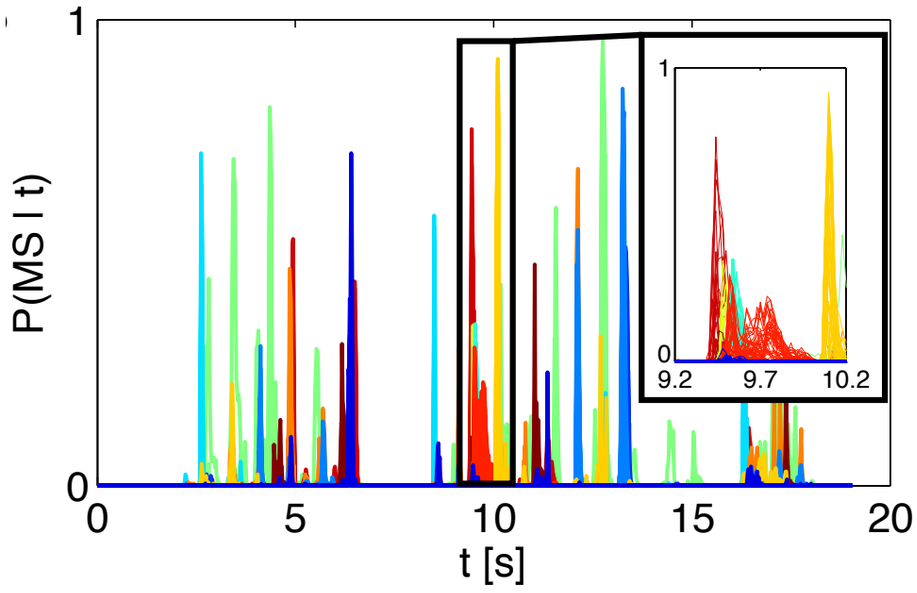
II: Energy landscape

- The energy landscape of the resulting models has many local minima (MS states)
- Stimuli could be encoded by the identity of the basin of attraction and not by the detailed micro-state





the number of MS states grows with the size of neural network N



across repeats of exactly the same stimulus, the retina reproducibly transitions between the same MS states

III: Scaling of entropy

We can compute (or upper-bound) the entropy
of the "vocabulary".

III: Scaling of entropy

We can compute (or upper-bound) the entropy
of the "vocabulary".

I. Run a series of MC samplings for Var E for T=0,..., I

$$S(T) = \int_0^T dT' \frac{C(T')}{T'}$$

II. Use Wang-Landau sampling.

III. P(0) = 1/Z by maxent constraint, P(0) can be sampled from data

III: Scaling of entropy

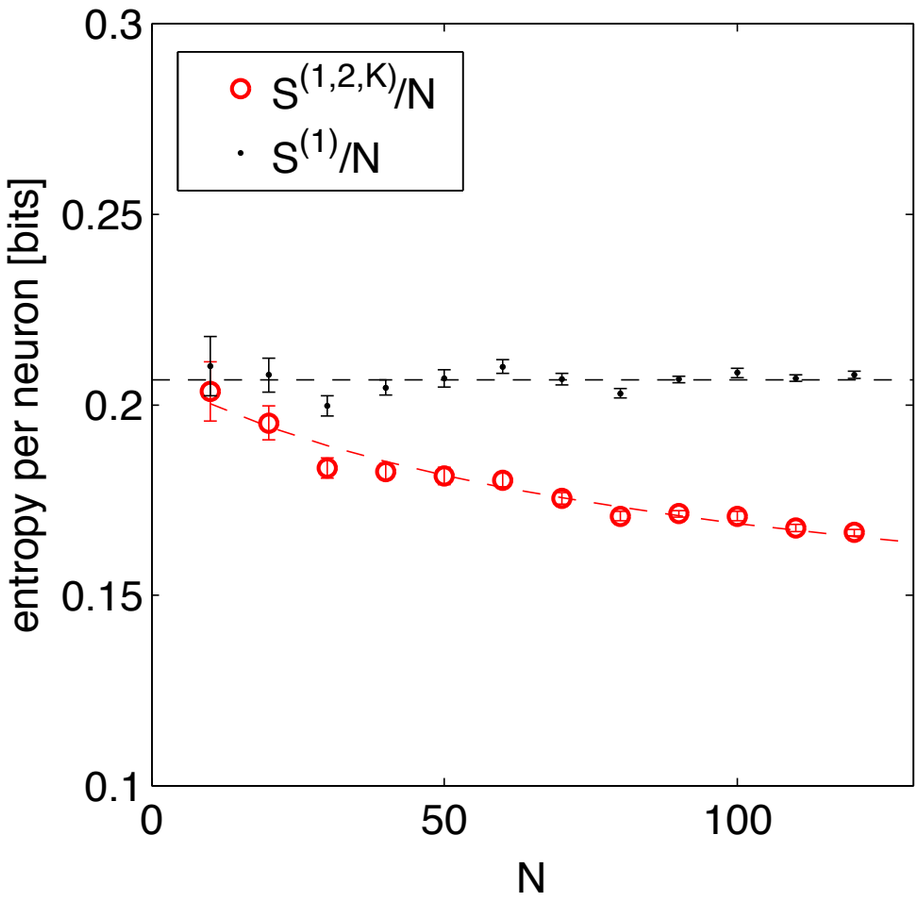
We can compute (or upper-bound) the entropy of the "vocabulary".

I. Run a series of MC samplings for Var E for T=0,..., I

$$S(T) = \int_0^T dT' \frac{C(T')}{T'}$$

II. Use Wang-Landau sampling.

III. P(0) = 1/Z by maxent constraint, P(0) can be sampled from data



III: Scaling of entropy

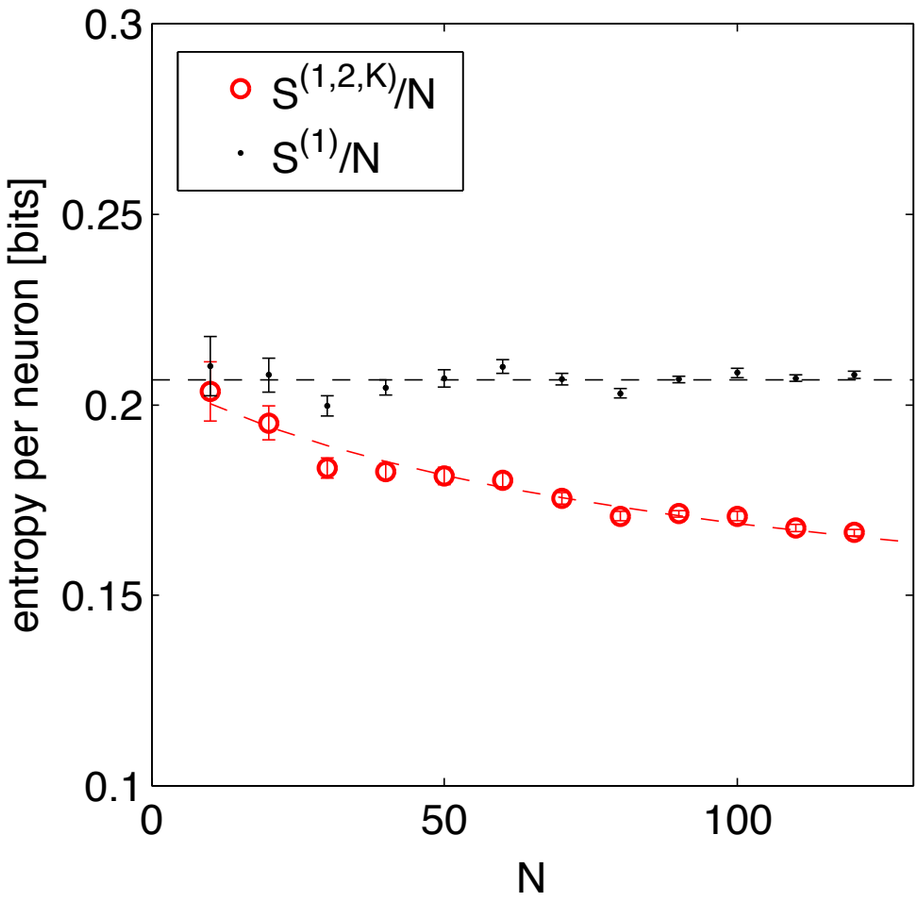
We can compute (or upper-bound) the entropy of the "vocabulary".

I. Run a series of MC samplings for Var E for T=0,..., I

$$S(T) = \int_0^T dT' \frac{C(T')}{T'}$$

II. Use Wang-Landau sampling.

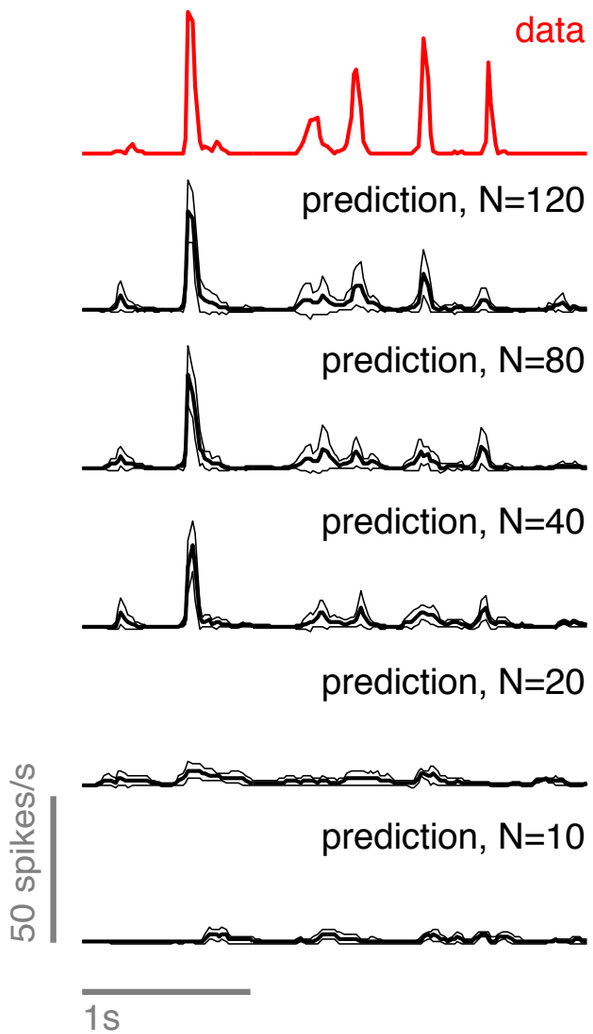
III. P(0) = 1/Z by maxent constraint, P(0) can be sampled from data



- This entropy is an upper bound for the information transmission.
- Even at 120 neurons, we are not yet in the extensive regime (consistent with the expectation of ~200 neuron correlated patch as a basic coding unit in the salamander retina).

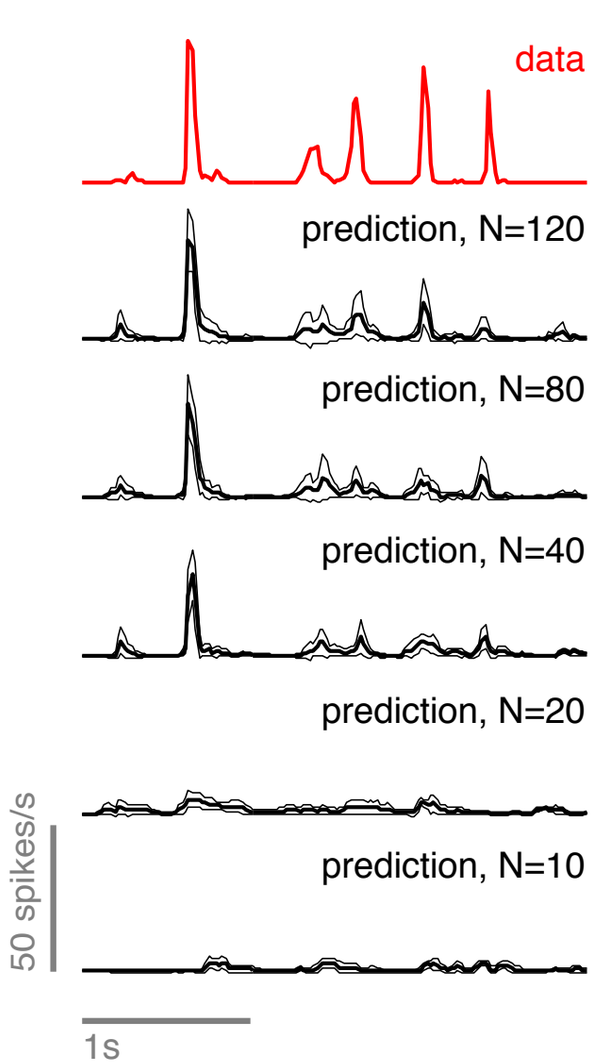
IV: Correlations and error correction

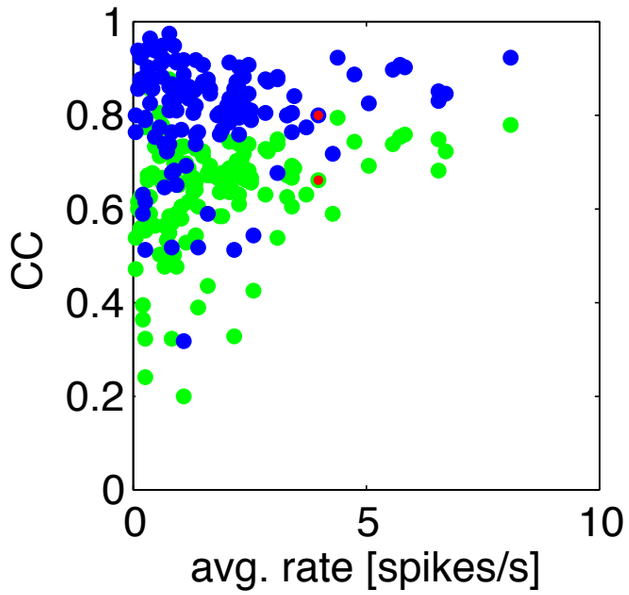
- The population is redundant in a way that permits error correction
- The state of single neuron is predictable from the rest of the network *even without knowing the stimulus*

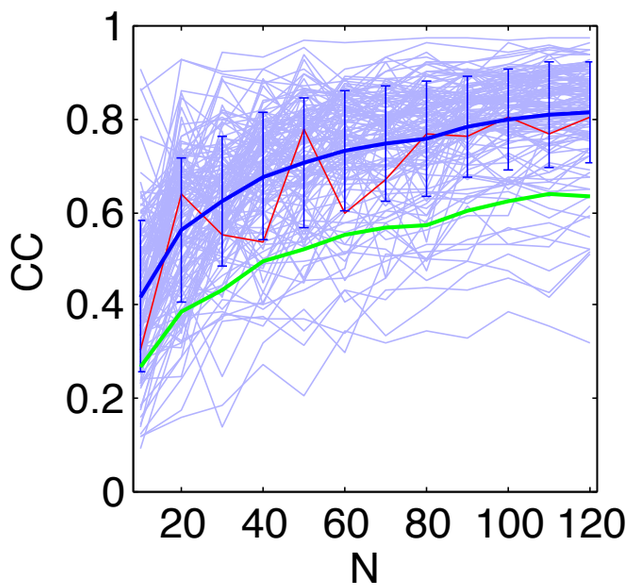


IV: Correlations and error correction

- The population is redundant in a way that permits error correction
- The state of single neuron is predictable from the rest of the network *even without knowing the stimulus*

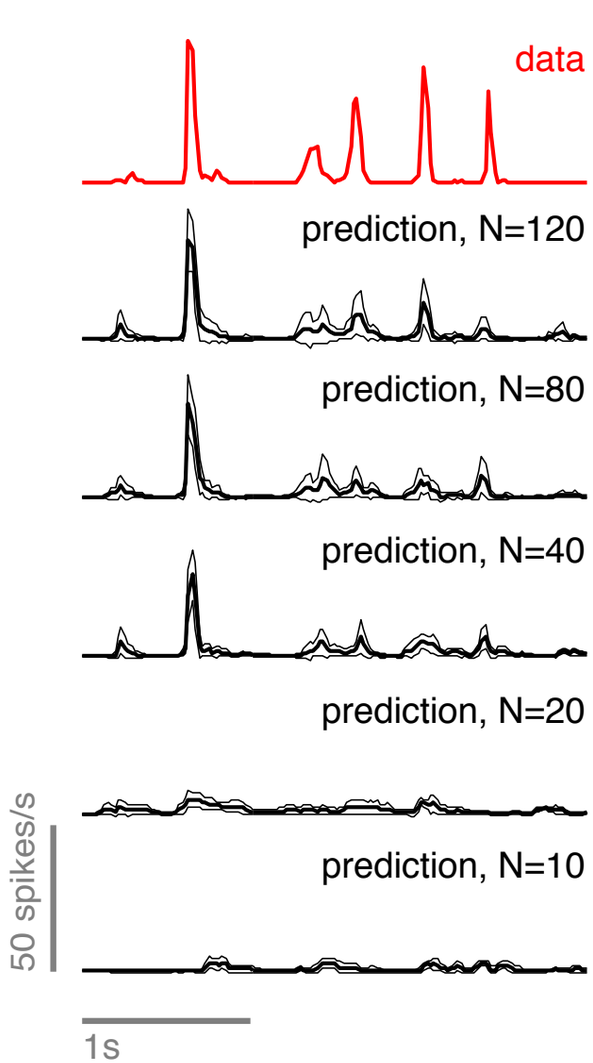


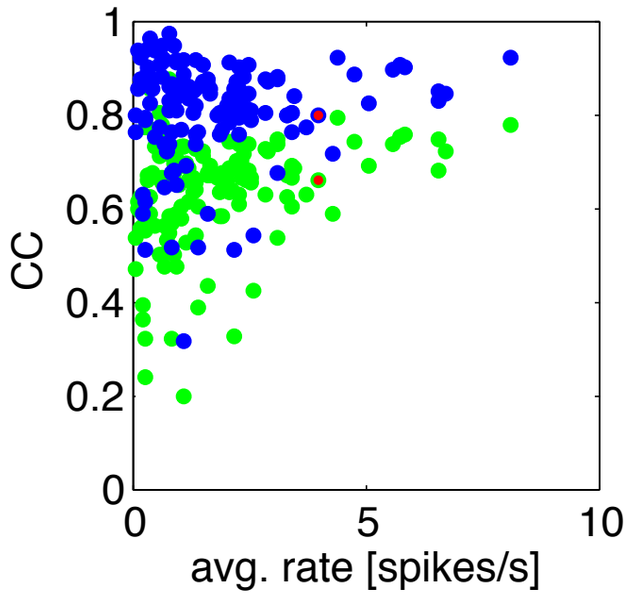


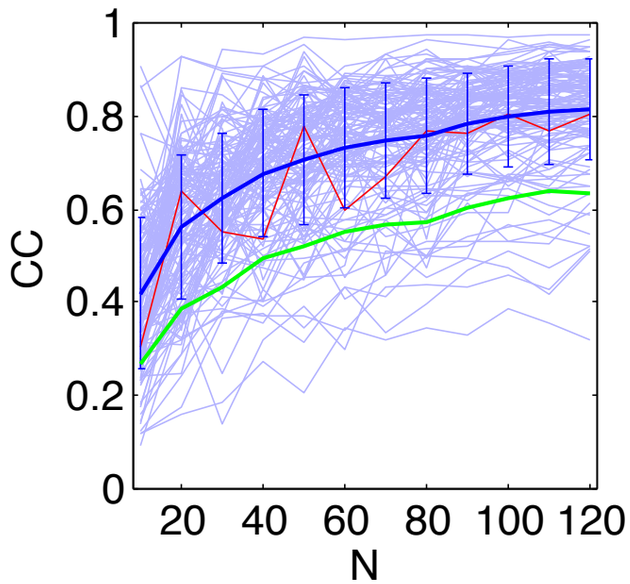


IV: Correlations and error correction

- The population is redundant in a way that permits error correction
- The state of single neuron is predictable from the rest of the network *even without knowing the stimulus*







One can predict the firing rate of a chosen neuron with ~80% correlation from the state of other neurons.

Thermodynamic behavior of the vocabulary P(σ)

- Is the vocabulary distribution, P(σ), "special"?
- How would one check if the system is close to critical...
  - long-range interactions among units?
  - what is the order parameter and how to couple to it?
- Try TD signatures (e.g., diverging heat capacity)!

Signatures of criticality: density of states

compute the microcanonical entropy
S(E) = log(# of microstates with energy ~E)...

Signatures of criticality: density of states

compute the microcanonical entropy
S(E) = log(# of microstates with energy ~E)...

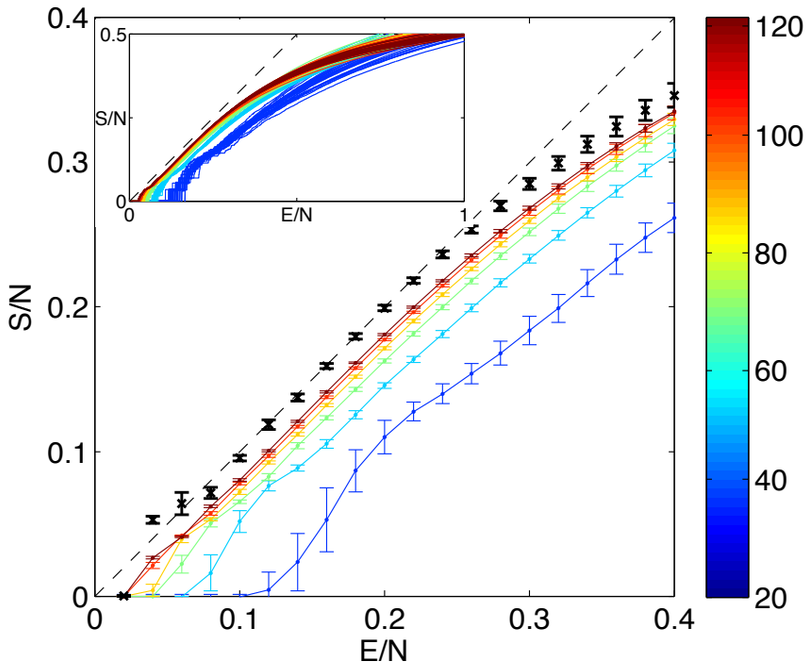
$d^2S/d^2E = 0$ ~ critical point

Signatures of criticality: density of states

compute the microcanonical entropy
S(E) = log(# of microstates with energy ~E)...

$d^2S/d^2E = 0$ ~ critical point

Use WL sampling to construct S/N vs E/N...

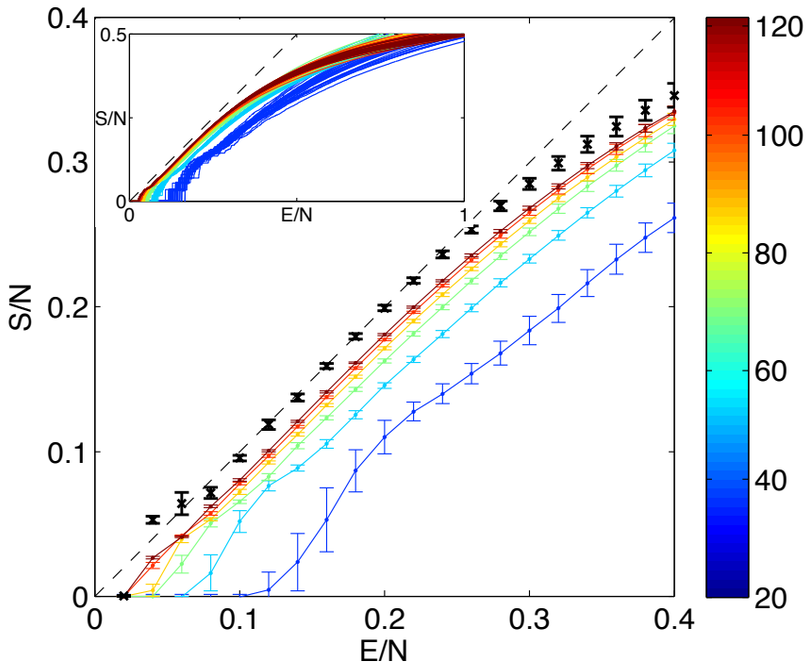


Signatures of criticality: density of states

compute the microcanonical entropy
S(E) = log(# of microstates with energy ~E)...

$d^2S/d^2E = 0$ ~ critical point

Use WL sampling to construct S/N vs E/N...



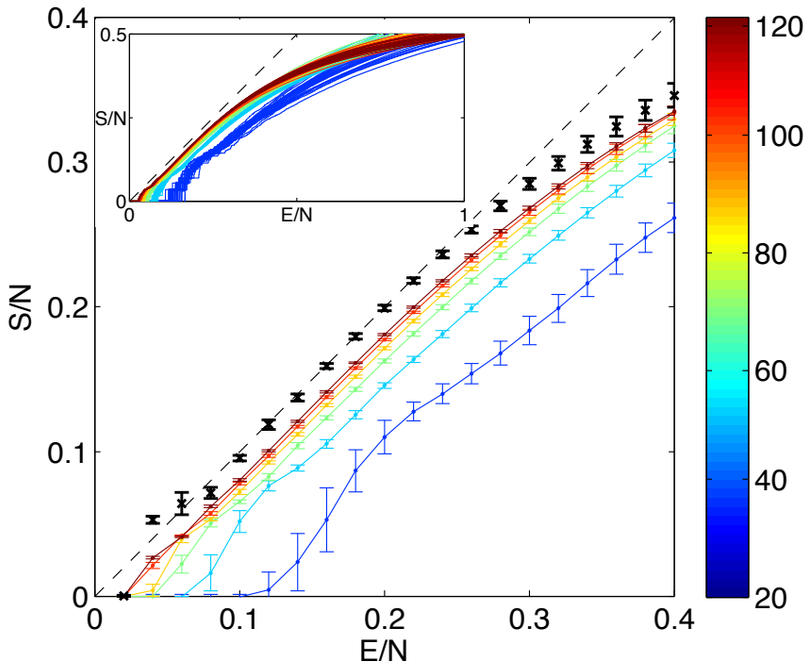
- as N increases, S/N = E/N to a good approximation

Signatures of criticality: density of states

compute the microcanonical entropy
S(E) = log(# of microstates with energy ~E)...

$d^2S/d^2E = 0$ ~ critical point

Use WL sampling to construct S/N vs E/N...



Estimate directly from data by histogram counts...

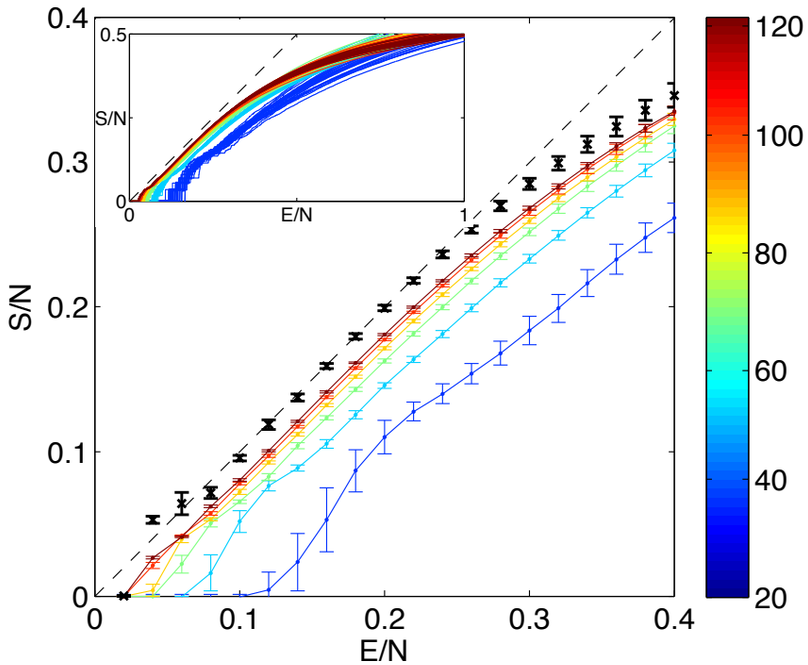
- as N increases, S/N = E/N to a good approximation

Signatures of criticality: density of states

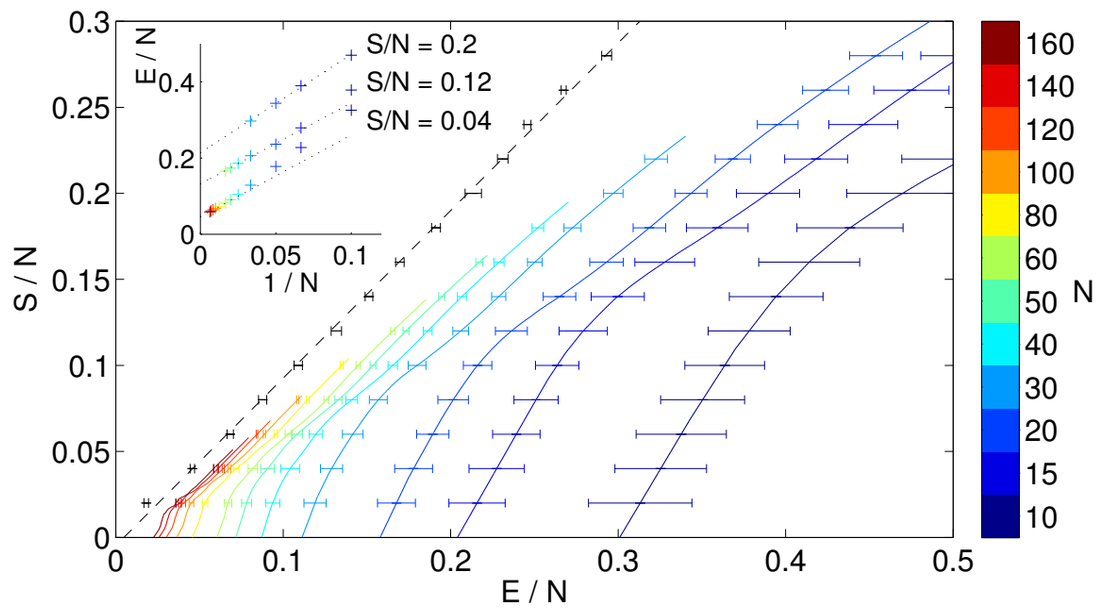
compute the microcanonical entropy
S(E) = log(# of microstates with energy ~E)...

$d^2S/d^2E = 0$ ~ critical point

Use WL sampling to construct S/N vs E/N...



Estimate directly from data by histogram counts...



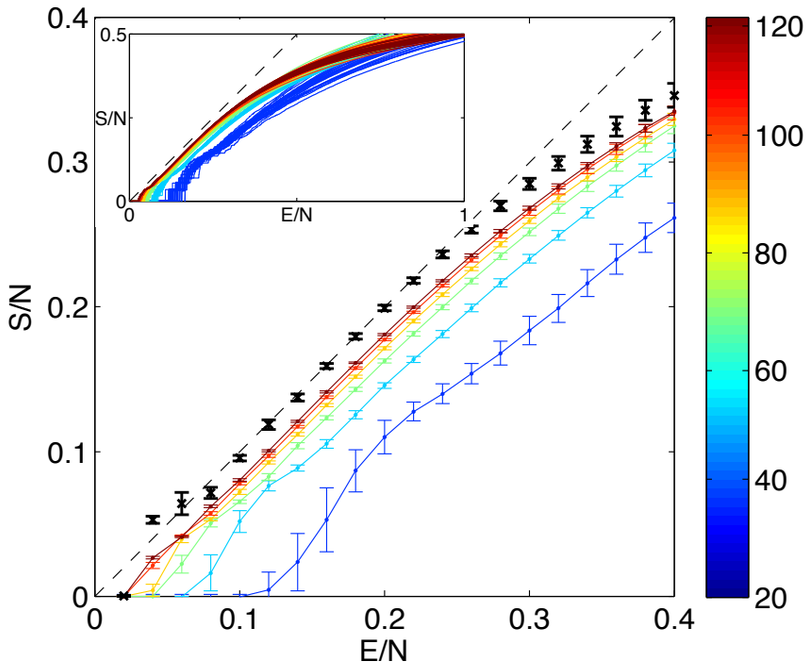
- as N increases, S/N = E/N to a good approximation

Signatures of criticality: density of states

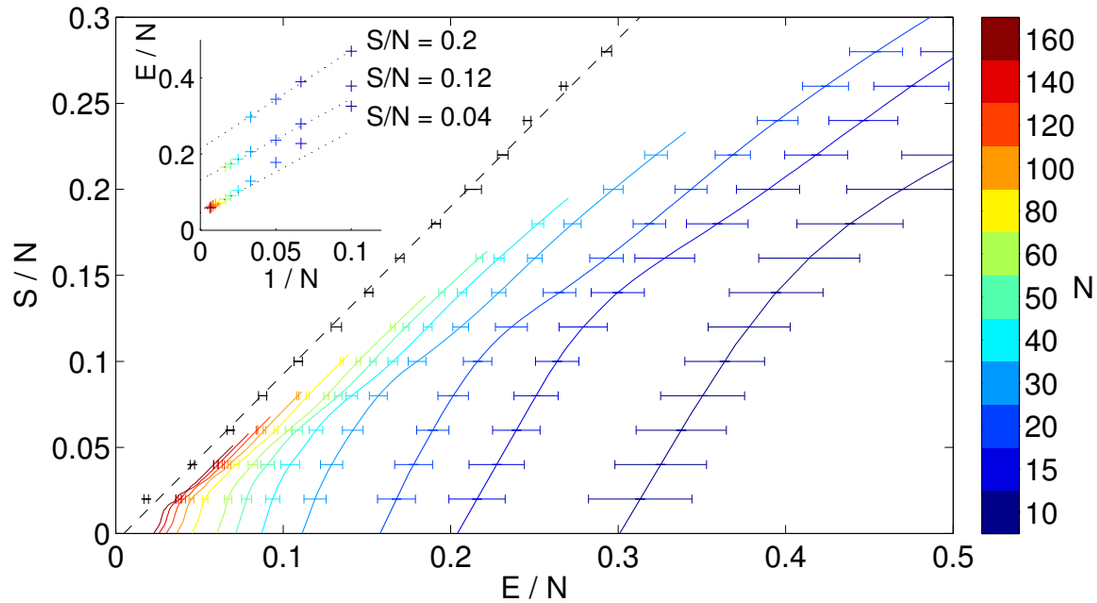
compute the microcanonical entropy
S(E) = log(# of microstates with energy ~E)...

$d^2S/d^2E = 0$ ~ critical point

Use WL sampling to construct S/N vs E/N...



Estimate directly from data by histogram counts...



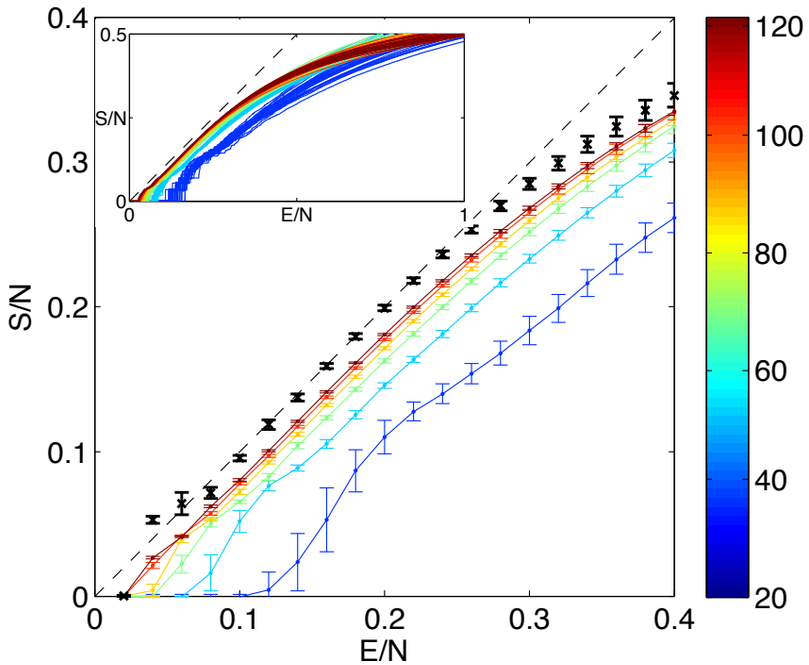
- as N increases, S/N = E/N to a good approximation
- the same behavior is observed when estimating S(E) directly from data

Signatures of criticality: density of states

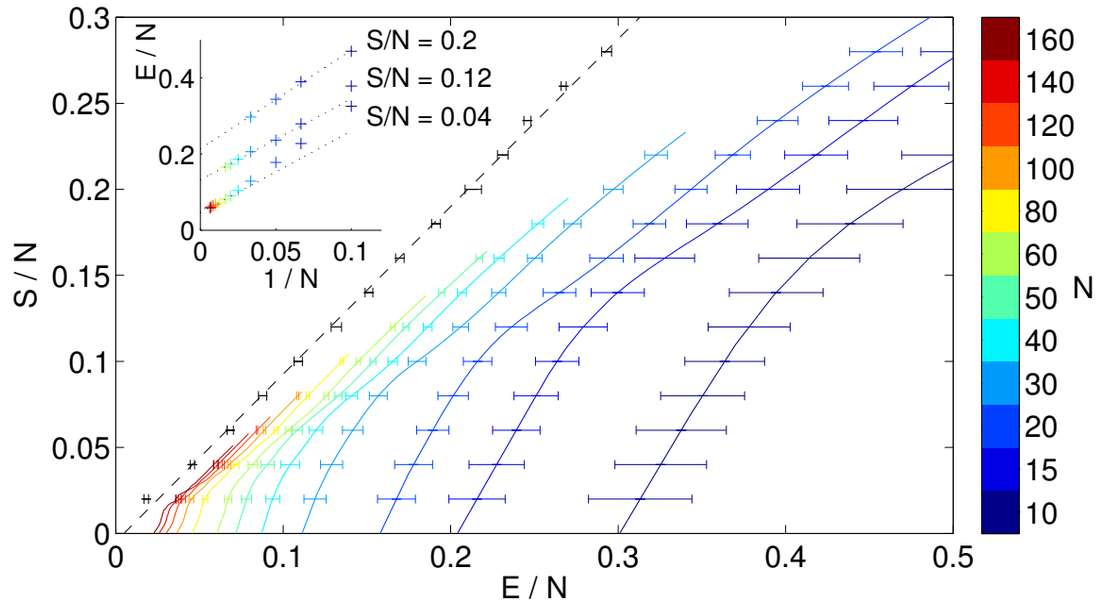
compute the microcanonical entropy
S(E) = log(# of microstates with energy ~E)...

$d^2S/d^2E = 0$ ~ critical point

Use WL sampling to construct S/N vs E/N...



Estimate directly from data by histogram counts...



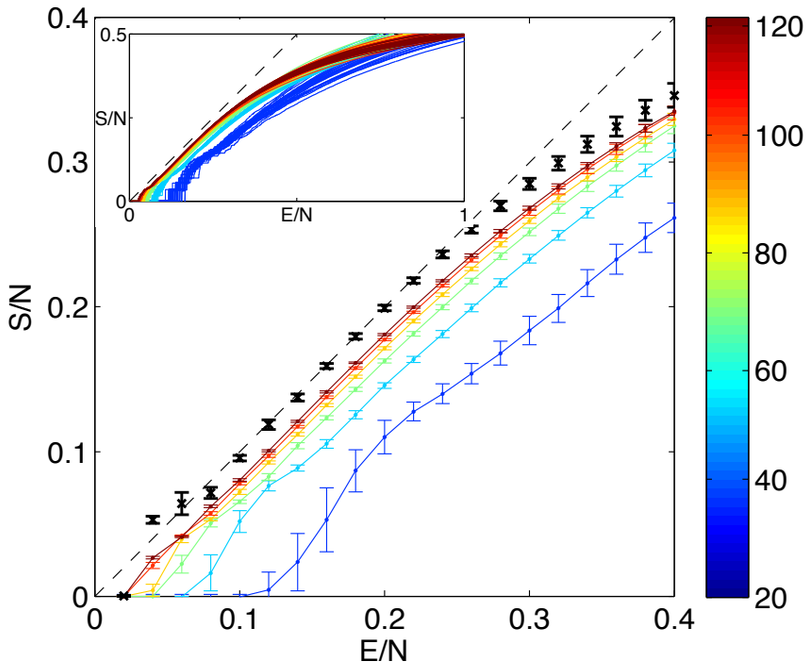
- as N increases, S/N = E/N to a good approximation
- the same behavior is observed when estimating S(E) directly from data
- This suggests a peculiar critical behavior, where $d^2S/d^2E = 0$ for a whole range of E!

Signatures of criticality: density of states

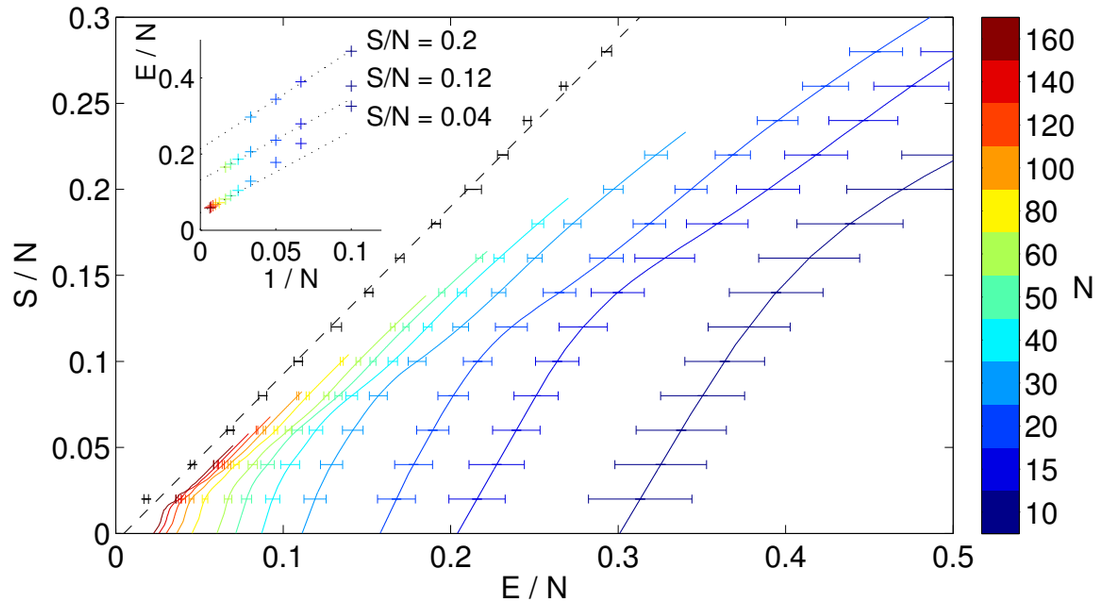
compute the microcanonical entropy
S(E) = log(# of microstates with energy ~E)...

$d^2S/d^2E = 0$ ~ critical point

Use WL sampling to construct S/N vs E/N...



Estimate directly from data by histogram counts...



- as N increases, S/N = E/N to a good approximation
- the same behavior is observed when estimating S(E) directly from data
- This suggests a peculiar critical behavior, where $d^2S/d^2E = 0$ for a whole range of E!
- Equivalent to Zipf law with -1 slope for the codewords.

Signatures of criticality: heat capacity

T rescales the reconstructed Hamiltonian to generate a family of distributions...

$$P(\{\sigma_i\};T) = \frac{1}{Z(T)} \exp\left[-\frac{1}{T}\mathcal{H}(\{\sigma_i\})\right]$$

(T=1 is the original model inferred from data)

Signatures of criticality: heat capacity

T rescales the reconstructed Hamiltonian to generate a family of distributions...

$$P(\{\sigma_i\};T) = \frac{1}{Z(T)} \exp\left[-\frac{1}{T}\mathcal{H}(\{\sigma_i\})\right]$$

(T=1 is the original model inferred from data)

... for each distribution, we compute C(T)

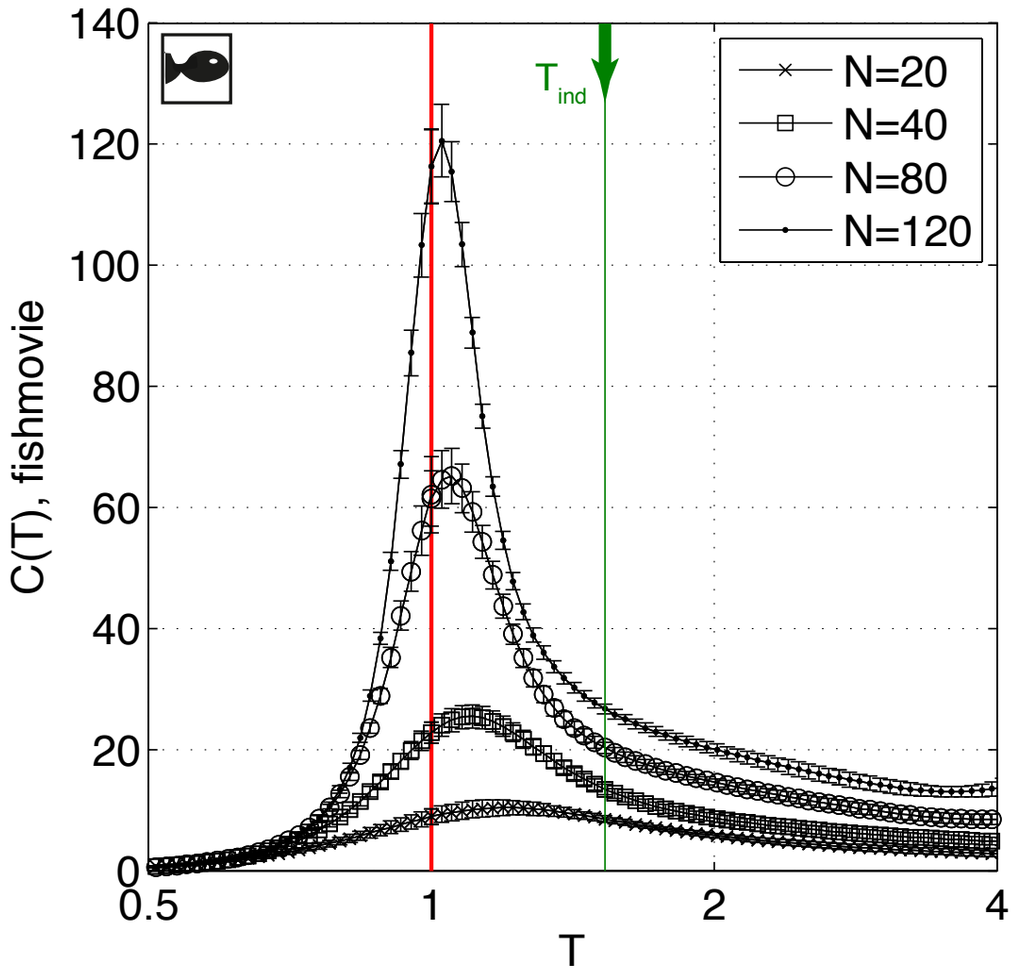
Signatures of criticality: heat capacity

T rescales the reconstructed Hamiltonian to generate a family of distributions...

$$P(\{\sigma_i\};T) = \frac{1}{Z(T)} \exp\left[-\frac{1}{T}\mathcal{H}(\{\sigma_i\})\right]$$

(T=1 is the original model inferred from data)

... for each distribution, we compute C(T)



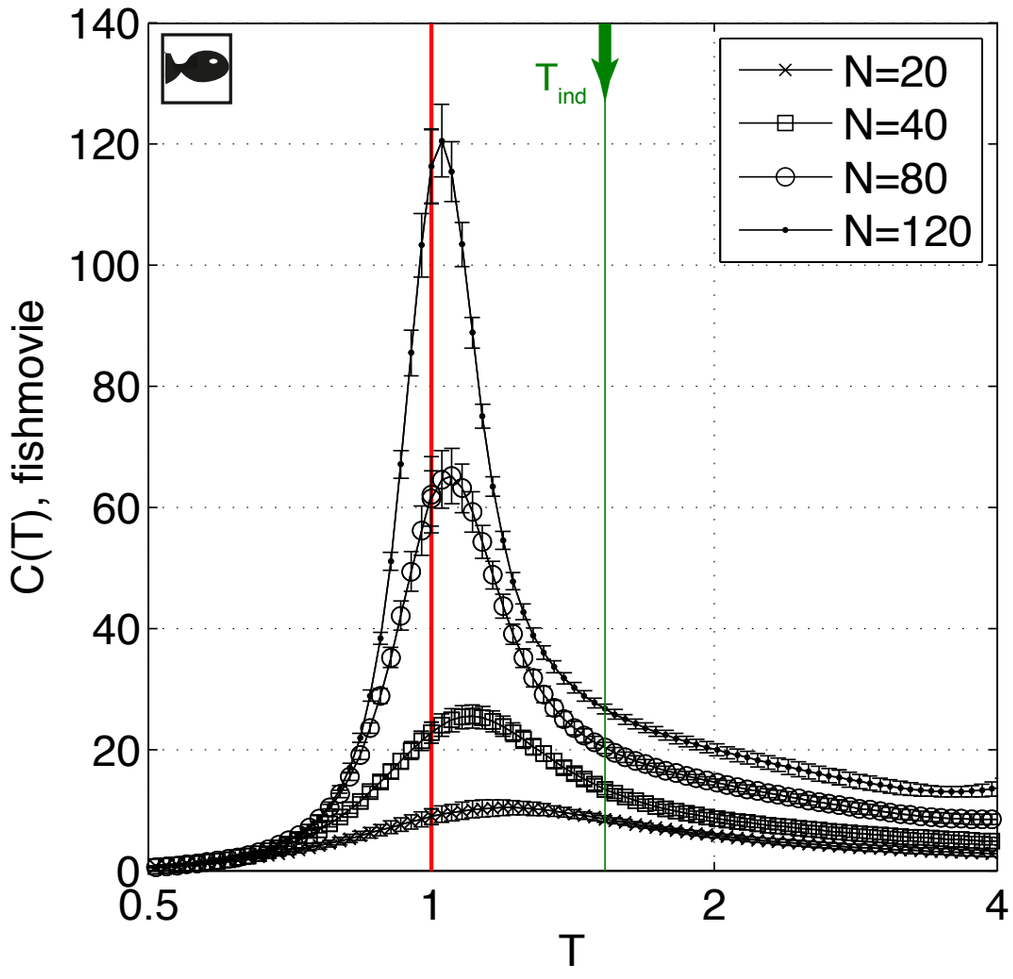
Signatures of criticality: heat capacity

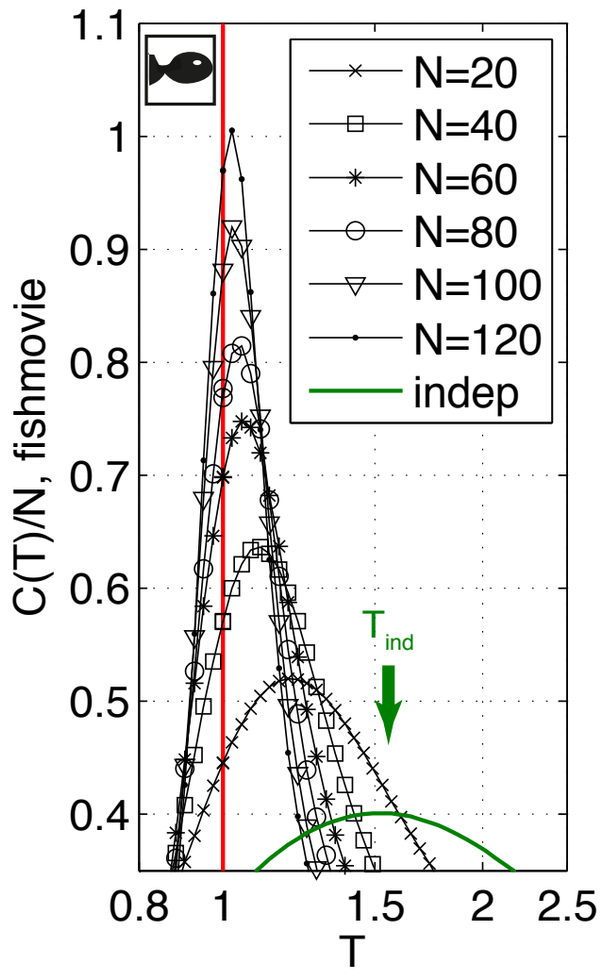
T rescales the reconstructed Hamiltonian to generate a family of distributions...

$$P(\{\sigma_i\};T) = \frac{1}{Z(T)} \exp\left[-\frac{1}{T}\mathcal{H}(\{\sigma_i\})\right]$$

(T=1 is the original model inferred from data)

... for each distribution, we compute C(T)





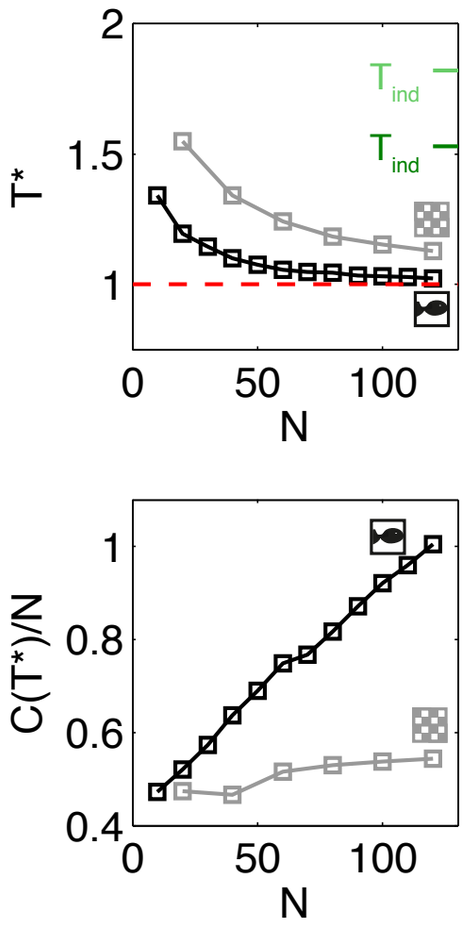
Signatures of criticality: heat capacity

T rescales the reconstructed Hamiltonian to generate a family of distributions...

$$P(\{\sigma_i\};T) = \frac{1}{Z(T)} \exp\left[-\frac{1}{T}\mathcal{H}(\{\sigma_i\})\right]$$

(T=1 is the original model inferred from data)

... for each distribution, we compute C(T)



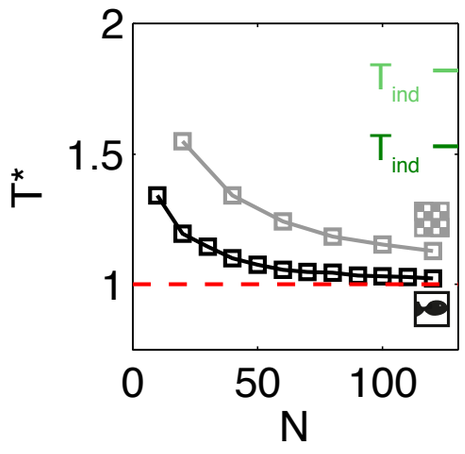
Signatures of criticality: heat capacity

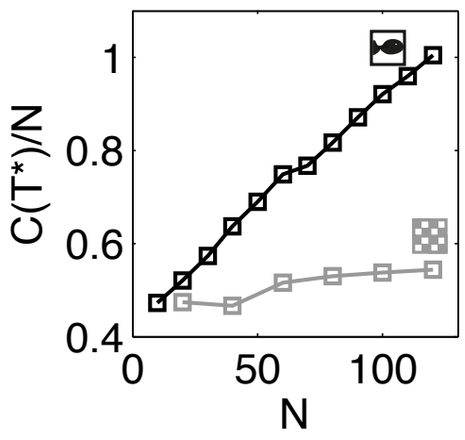
T rescales the reconstructed Hamiltonian to generate a family of distributions...

$$P(\{\sigma_i\};T) = \frac{1}{Z(T)} \exp\left[-\frac{1}{T}\mathcal{H}(\{\sigma_i\})\right]$$

(T=1 is the original model inferred from data)

... for each distribution, we compute C(T)





- as N increases, T* approaches 1
- C*/N grows linearly with N
- Emerging peak of C(T) is a signature of criticality

Signatures of criticality: correlation scaling

α rescales the coupling terms (while maintaining <σ> fixed to data)
to generate a family of distributions...

$$P(\{\sigma_i\};\alpha) = \frac{1}{Z(\alpha)} \exp\left(\sum h'(\alpha)\sigma_i + \alpha\left[\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j + V(K)\right]\right)$$

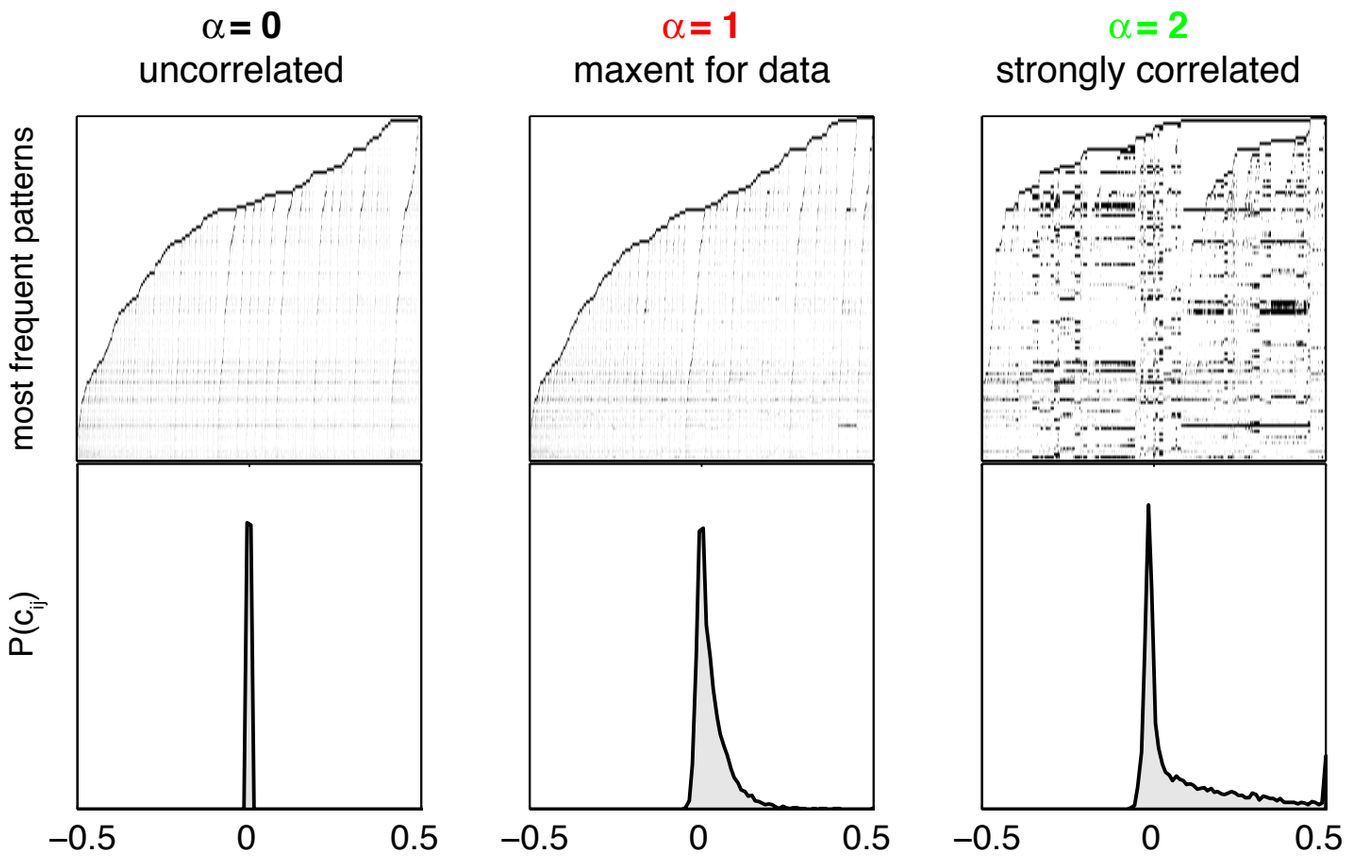
(α = 1 is the original model inferred from data)

Signatures of criticality: correlation scaling

α rescales the coupling terms (while maintaining <σ> fixed to data)
to generate a family of distributions...

$$P(\{\sigma_i\};\alpha) = \frac{1}{Z(\alpha)} \exp\left(\sum h'(\alpha)\sigma_i + \alpha\left[\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j + V(K)\right]\right)$$

(α = 1 is the original model inferred from data)

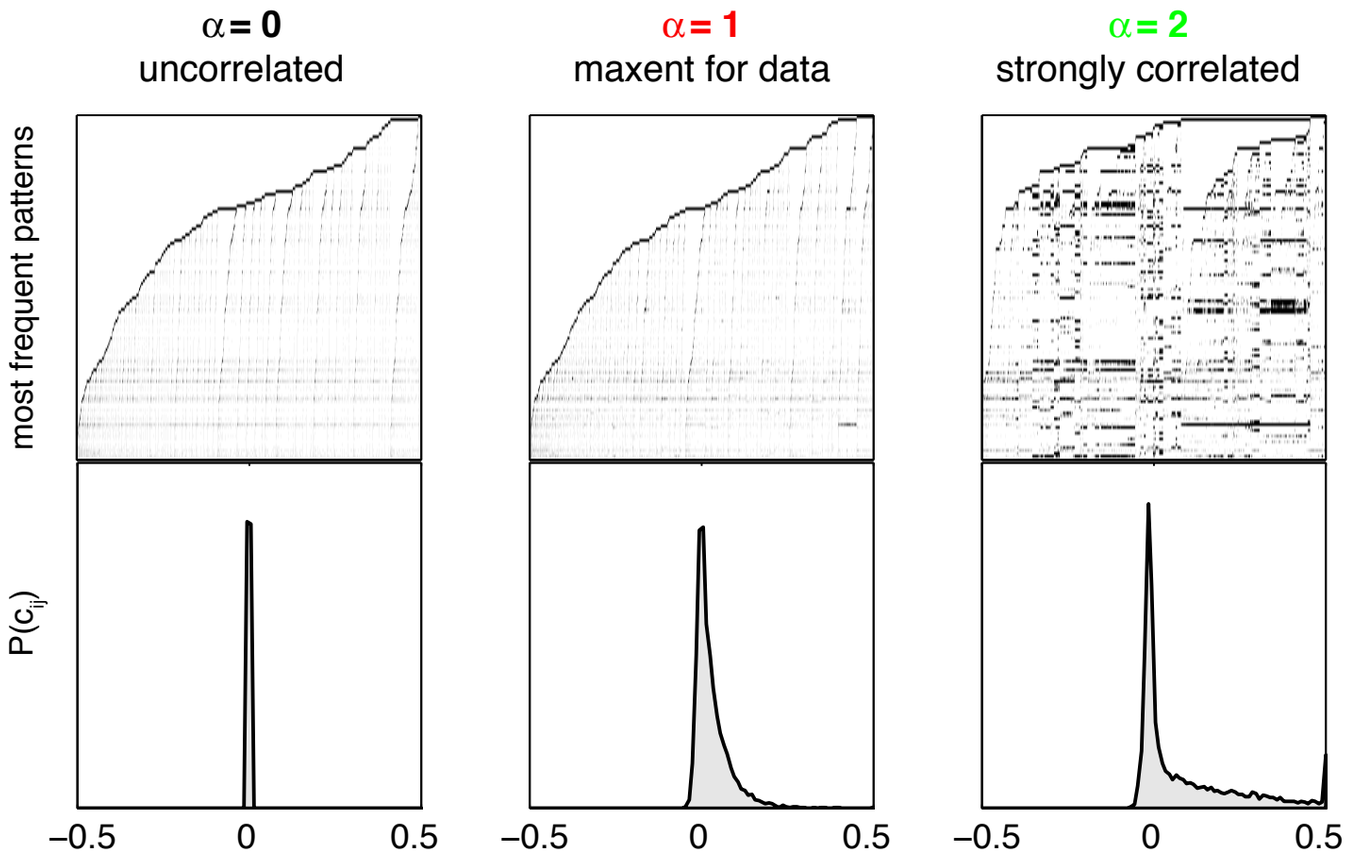


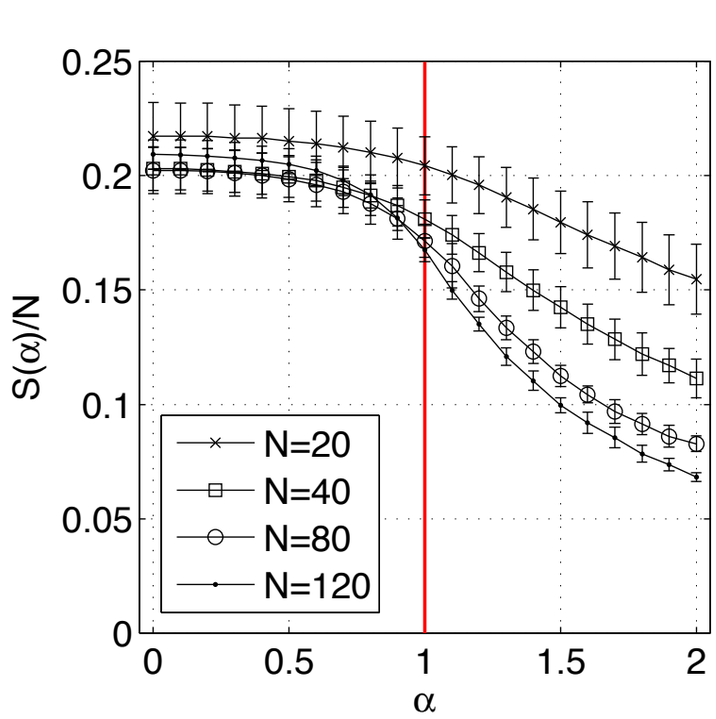
Signatures of criticality: correlation scaling

α rescales the coupling terms (while maintaining <σ> fixed to data)
to generate a family of distributions...

$$P(\{\sigma_i\};\alpha) = \frac{1}{Z(\alpha)} \exp\left(\sum h'(\alpha)\sigma_i + \alpha\left[\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j + V(K)\right]\right)$$

(α = 1 is the original model inferred from data)



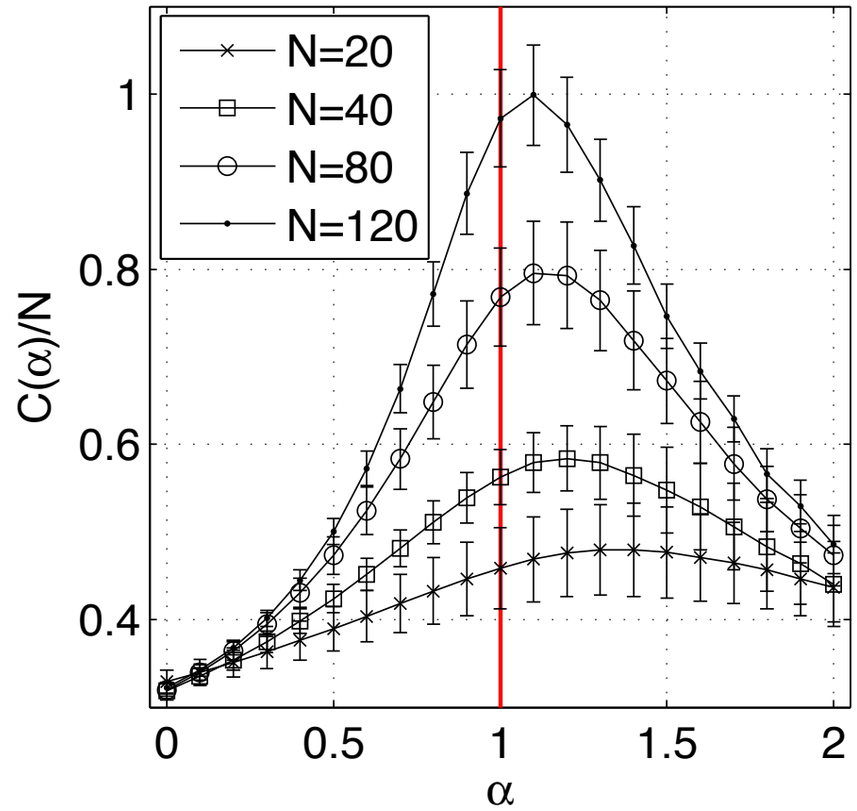


Signatures of criticality: correlation scaling

α rescales the coupling terms (while maintaining <σ> fixed to data)
to generate a family of distributions...

$$P(\{\sigma_i\};\alpha) = \frac{1}{Z(\alpha)} \exp\left(\sum h'(\alpha)\sigma_i + \alpha\left[\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j + V(K)\right]\right)$$

(α = 1 is the original model inferred from data)



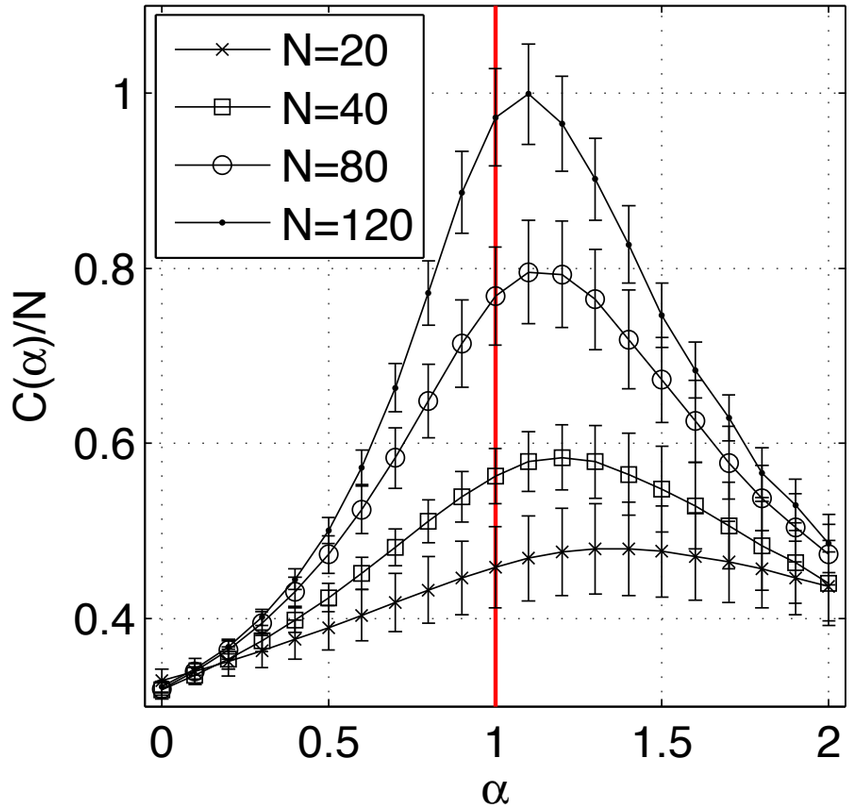
Signatures of criticality: correlation scaling

α rescales the coupling terms (while maintaining <σ> fixed to data)
to generate a family of distributions...

$$P(\{\sigma_i\};\alpha) = \frac{1}{Z(\alpha)} \exp\left(\sum h'(\alpha)\sigma_i + \alpha\left[\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j + V(K)\right]\right)$$

(α = 1 is the original model inferred from data)

- ensembles of codes with same mean rates but tunable correlations
- observed correlation strength is close to critical



Critical codes, why and what for?

Critical codes, why and what for?

We don't know.

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?

Mastromatteo + Marsili

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?
Mastromatteo + Marsili

NO.

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?
Mastromatteo + Marsili

NO.

Maybe it is an "automatic" consequence of modeling a driven system?
Schwab + Nemenman + Mehta

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?
Mastromatteo + Marsili

NO.

Maybe it is an "automatic" consequence of modeling a driven system?
Schwab + Nemenman + Mehta

MAYBE, BUT...

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?
Mastromatteo + Marsili

NO.

Maybe it is an "automatic" consequence of modeling a driven system?
Schwab + Nemenman + Mehta

MAYBE, BUT...

Maybe criticality has a functional significance....

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?
Mastromatteo + Marsili

NO.

Maybe it is an "automatic" consequence of modeling a driven system?
Schwab + Nemenman + Mehta

MAYBE, BUT...

Maybe criticality has a functional significance....

1. For encoding a wide range of surprise quickly (cf. "usual" optimal codes).

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?
Mastromatteo + Marsili

NO.

Maybe it is an "automatic" consequence of modeling a driven system?
Schwab + Nemenman + Mehta

MAYBE, BUT...

Maybe criticality has a functional significance....

1. For encoding a wide range of surprise quickly (cf. "usual" optimal codes).
2. A consequence of the need for long-timescale dynamics.

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?
Mastromatteo + Marsili

NO.

Maybe it is an "automatic" consequence of modeling a driven system?
Schwab + Nemenman + Mehta

MAYBE, BUT...

Maybe criticality has a functional significance....

1. For encoding a wide range of surprise quickly (cf. "usual" optimal codes).
2. A consequence of the need for long-timescale dynamics.
3. A side-effect of adaptation to maximize information transmission?

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?
Mastromatteo + Marsili

NO.

Maybe it is an "automatic" consequence of modeling a driven system?
Schwab + Nemenman + Mehta

MAYBE, BUT...

Maybe criticality has a functional significance....

1. For encoding a wide range of surprise quickly (cf. "usual" optimal codes).
2. A consequence of the need for long-timescale dynamics.
3. A side-effect of adaptation to maximize information transmission?
4. Learnability downstream?

Critical codes, why and what for?

We don't know.

Maybe it is due to (over)fitting a maxent model?
Mastromatteo + Marsili

NO.

Maybe it is an "automatic" consequence of modeling a driven system?
Schwab + Nemenman + Mehta

MAYBE, BUT...

Maybe criticality has a functional significance....

1. For encoding a wide range of surprise quickly (cf. "usual" optimal codes).
2. A consequence of the need for long-timescale dynamics.
3. A side-effect of adaptation to maximize information transmission?
4. Learnability downstream?

Whatever the answer might be, this is the distribution of activity patterns seen by the circuit downstream of the retina (remember, these circuits don't have direct access to the stimulus!)

Conclusions

Conclusions

- Inverse statistical mechanics using **maximum entropy principle** is a powerful tool to study the neural codeword vocabulary

Conclusions

- Inverse statistical mechanics using **maximum entropy principle** is a powerful tool to study the neural codeword vocabulary
- **Interesting coding properties:** strongly collective behavior (despite small pairwise correlations), MS states, non-extensive entropy scaling, redundancy for error correction, high coincidence probability

Conclusions

- Inverse statistical mechanics using **maximum entropy principle** is a powerful tool to study the neural codeword vocabulary

- **Interesting coding properties:** strongly collective behavior (despite small pairwise correlations), MS states, non-extensive entropy scaling, redundancy for error correction, high coincidence probability

- **Signatures of criticality:** S vs E curve, emerging peak in C(T) and $C(\alpha)$ at $T, \alpha = 1$

Conclusions

- Inverse statistical mechanics using **maximum entropy principle** is a powerful tool to study the neural codeword vocabulary
- **Interesting coding properties:** strongly collective behavior (despite small pairwise correlations), MS states, non-extensive entropy scaling, redundancy for error correction, high coincidence probability
- **Signatures of criticality:** S vs E curve, emerging peak in C(T) and $C(\alpha)$ at $T, \alpha = 1$
- **Is the system finely tuned (or adapting to) the critical point, or is this result somehow generic?**