

Entropy of randomized measurements and compression via neural networks

Juan Carrasquilla, ETH Zürich

Interplay between neural and tensor networks: a workshop on recent developments

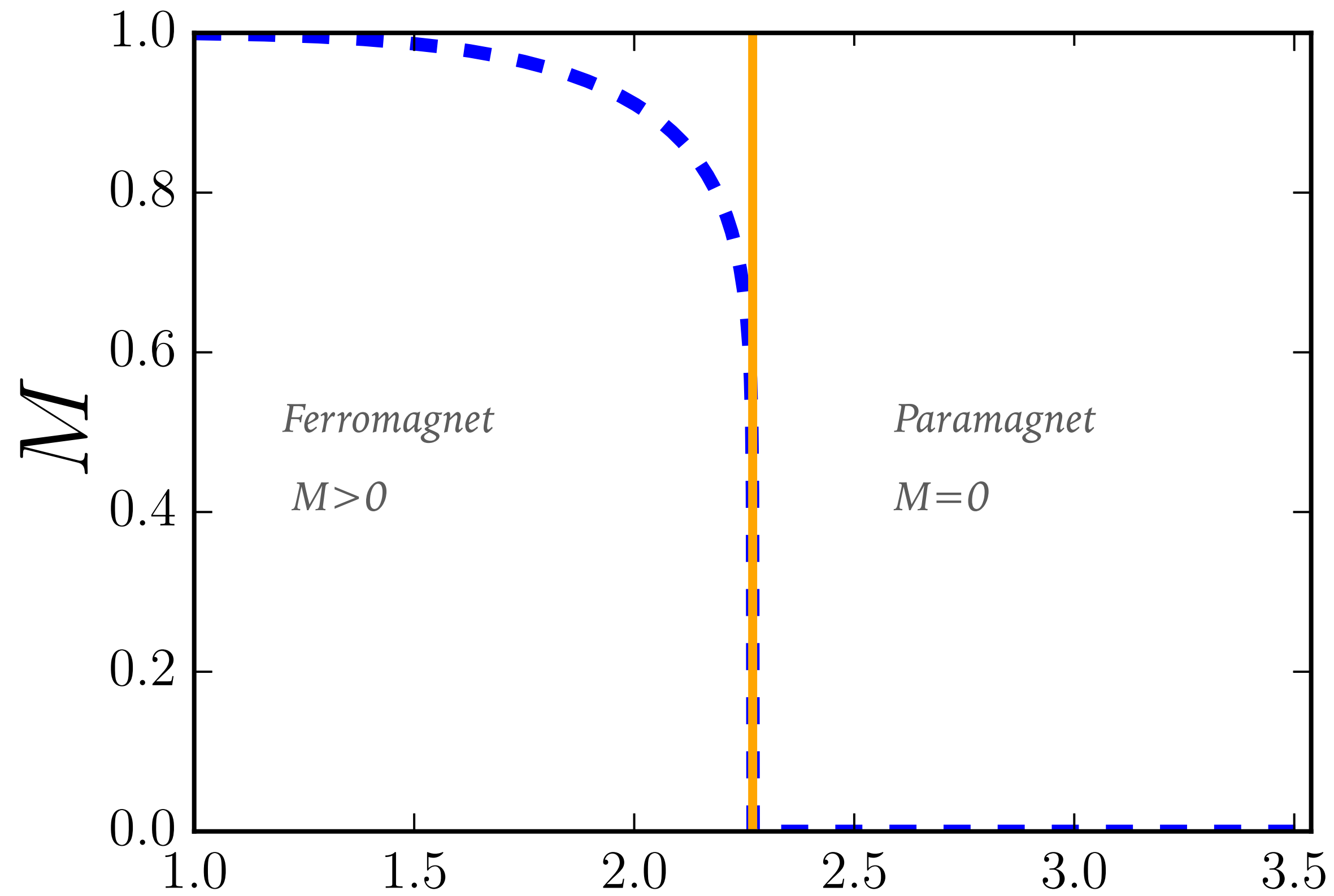
ETH zürich



Zahra Farahmand Nicola Togni

Ising model

Ferromagnetic transition



$$M = \frac{1}{N} \sum_i \langle \sigma_i \rangle, \quad T \quad \sigma_i = \pm 1$$

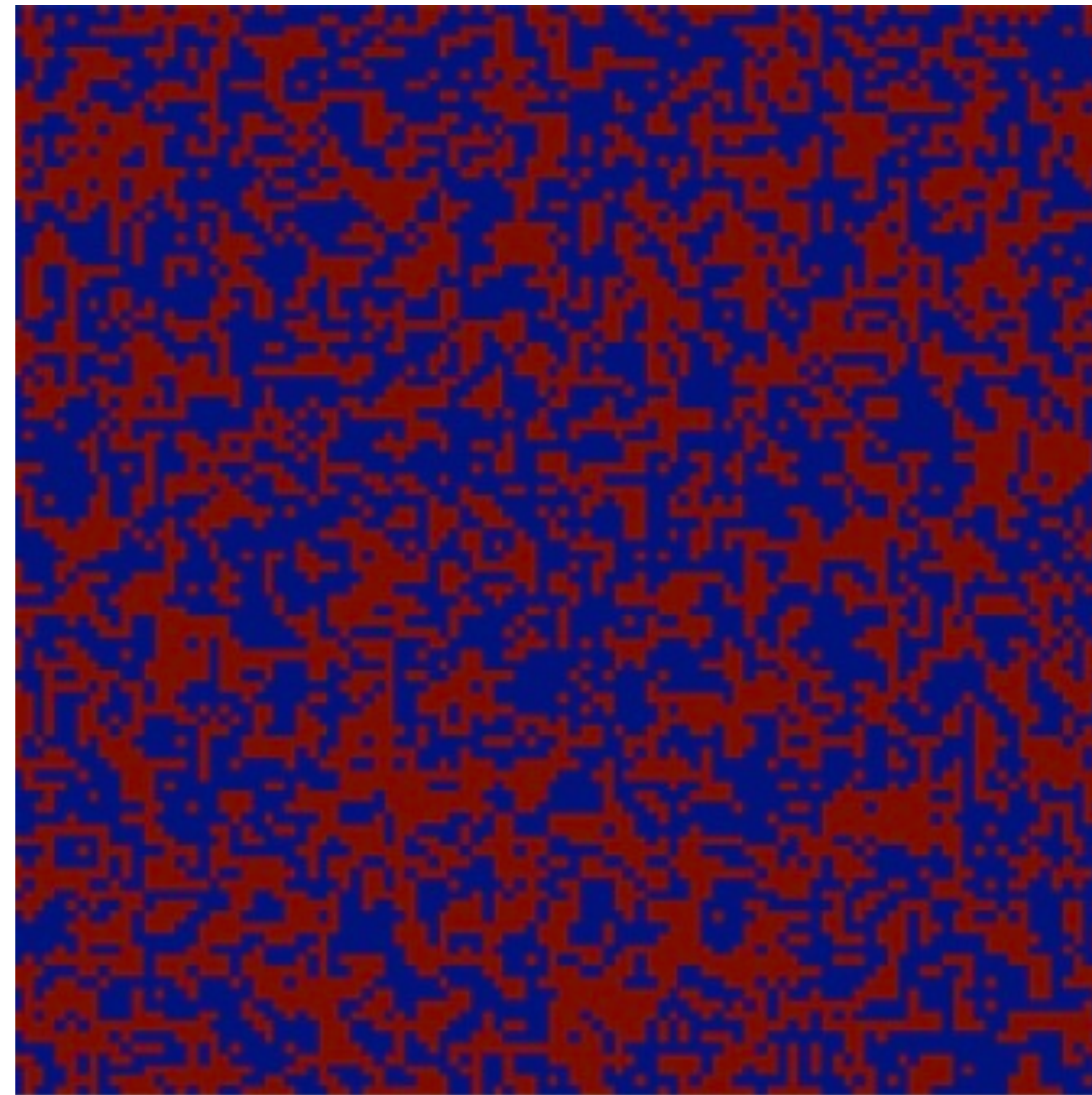
Lars Onsager Phys. Rev. 65, 117 (1944)

Classical Ising model

Uncompressed



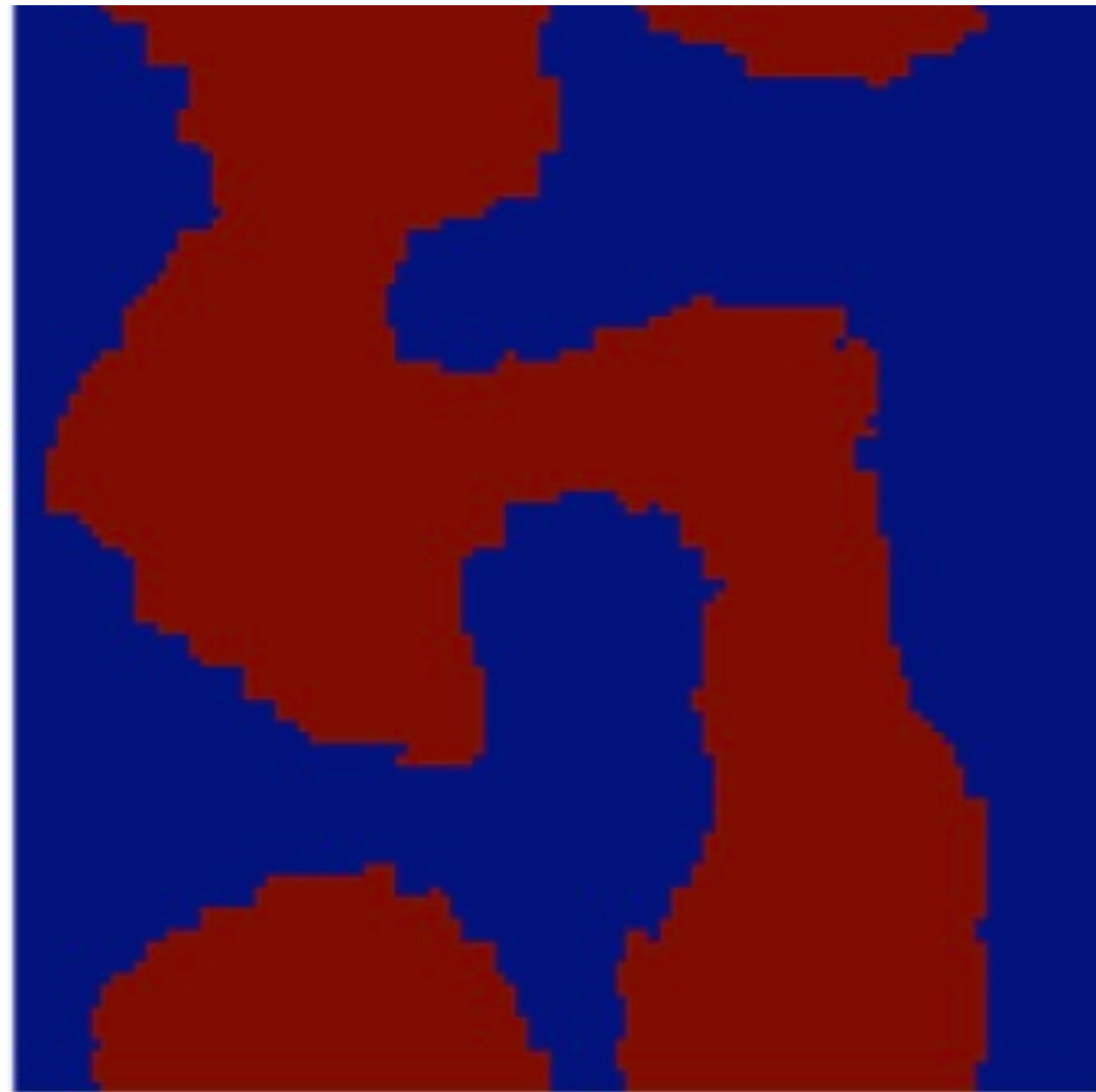
349 KB



345 KB

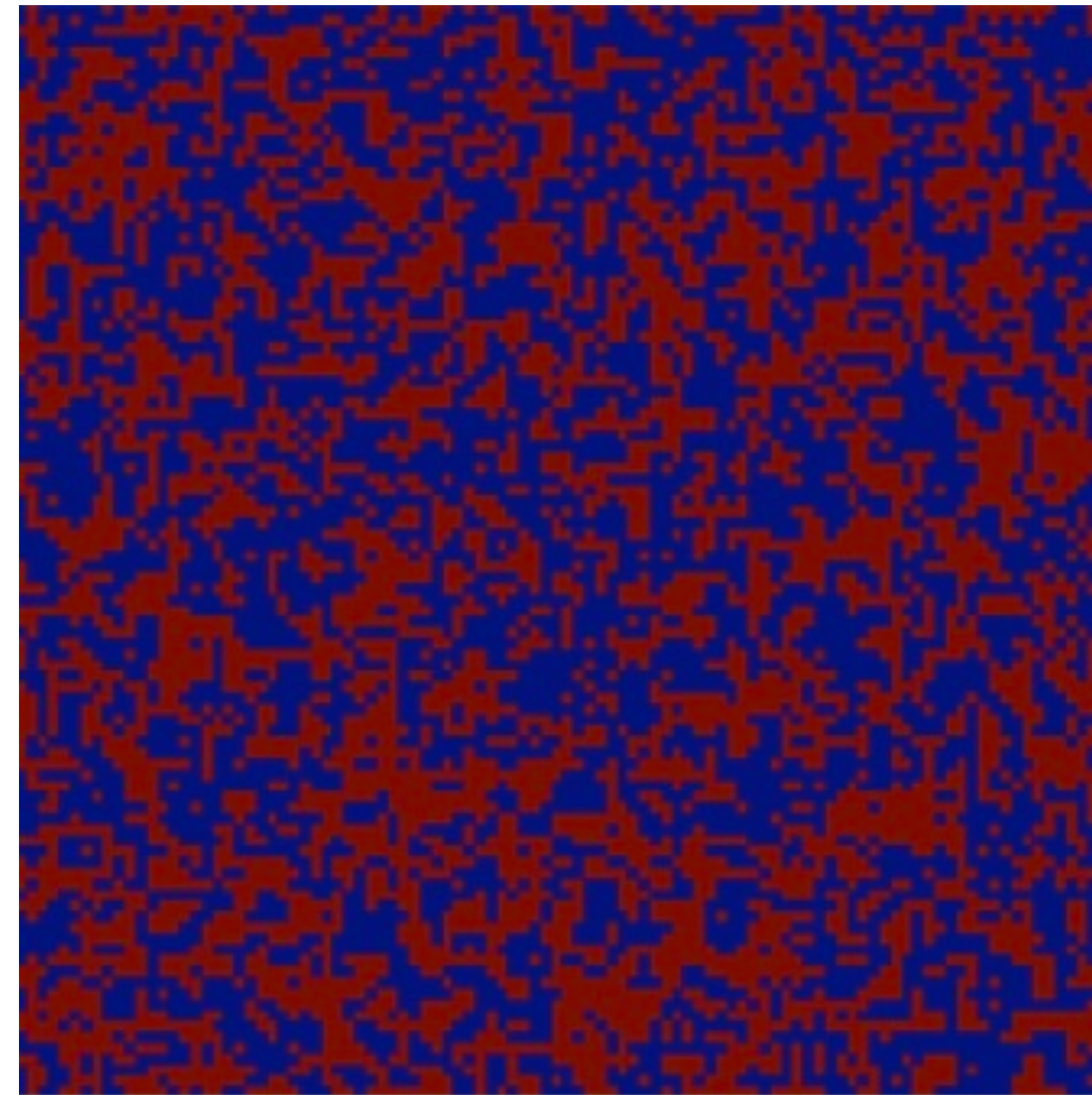
Classical Ising model

Compressed (ZIP)



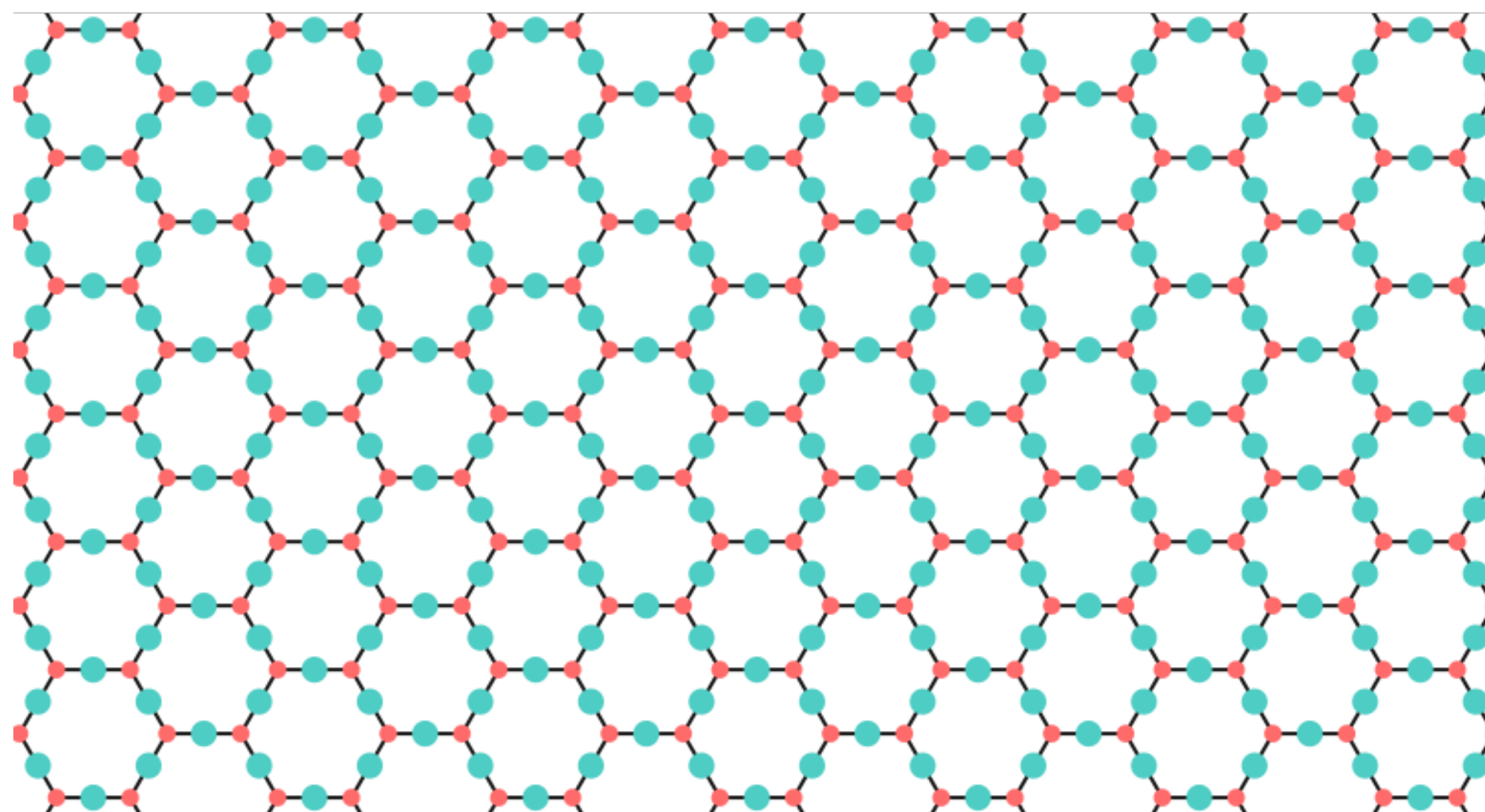
68 KB

<https://www.bdhammel.com/ising-model/>

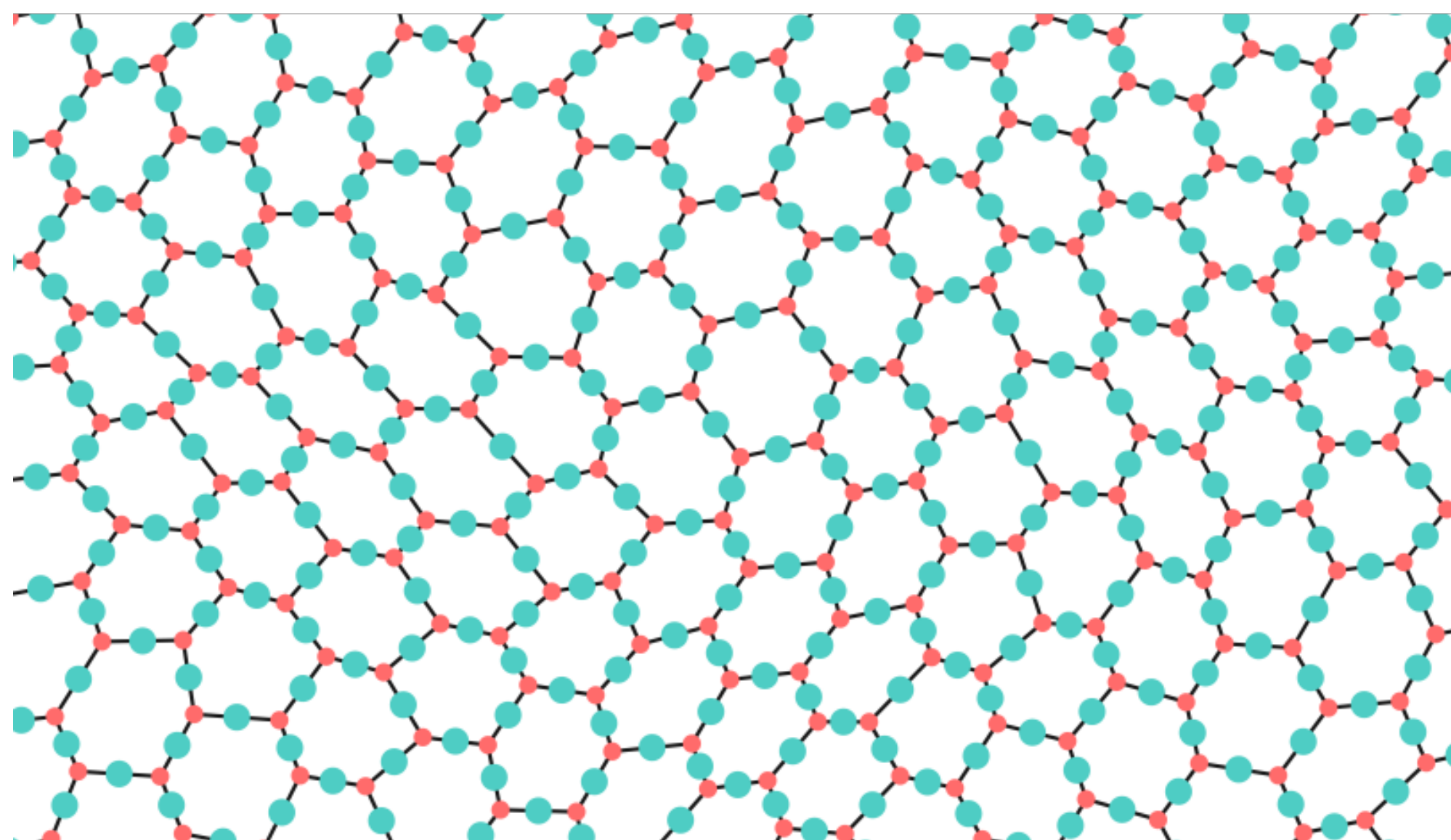


278 KB

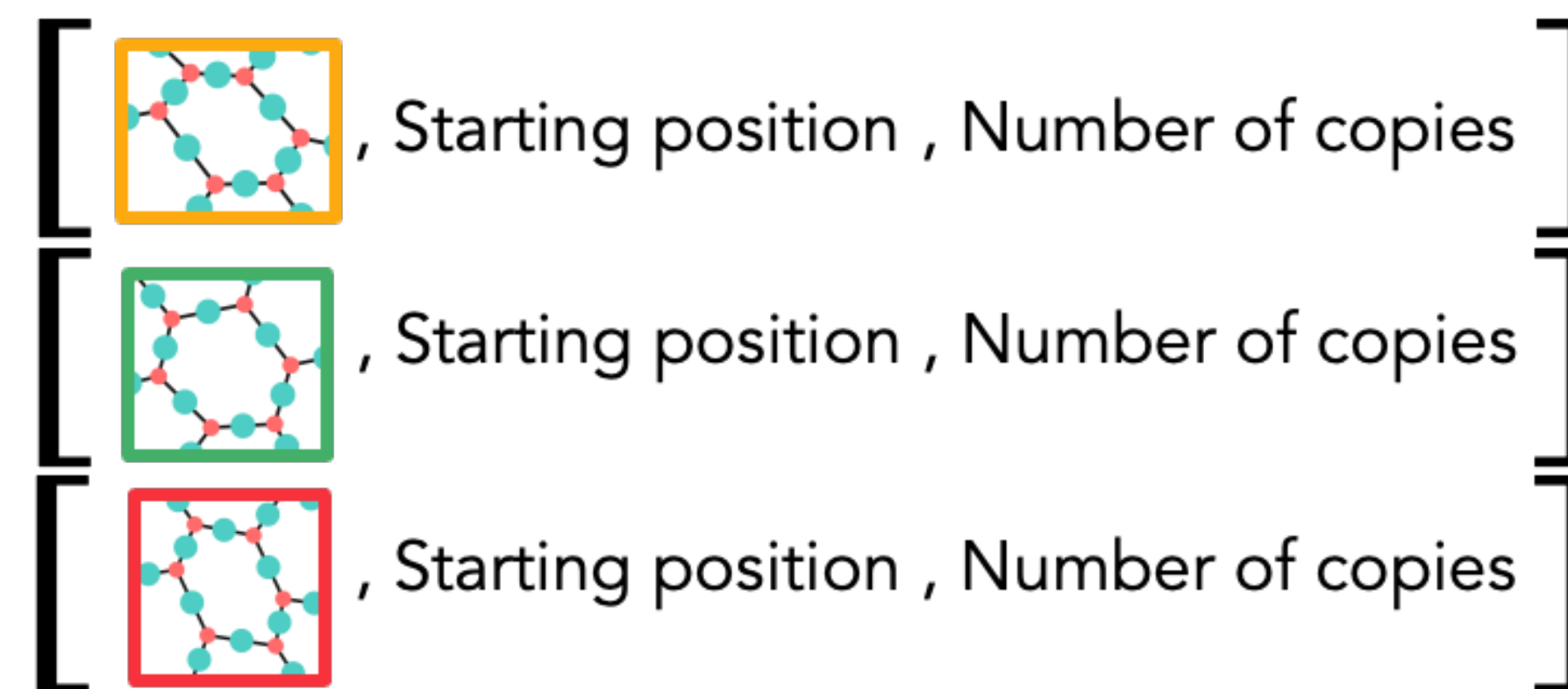
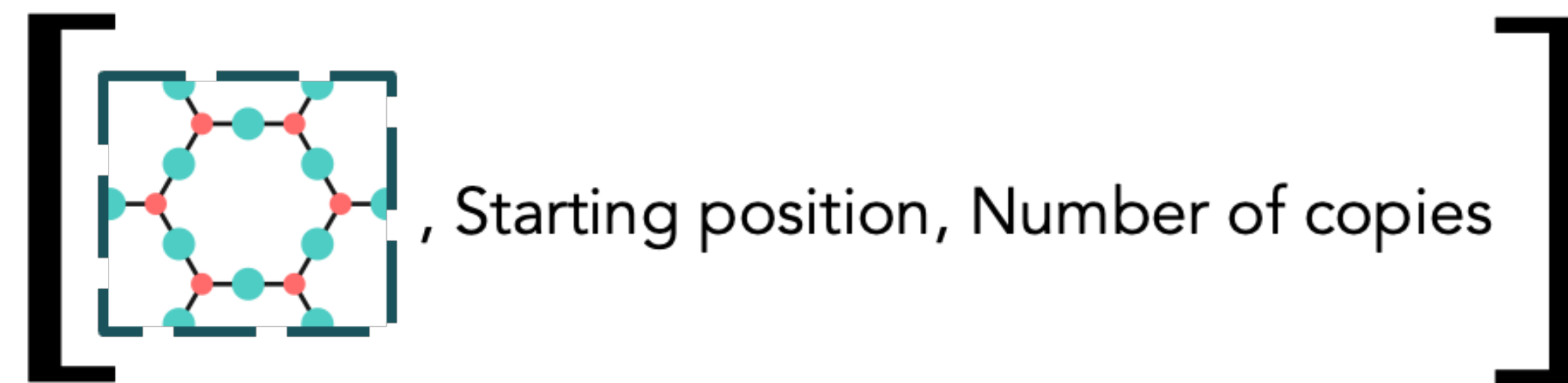
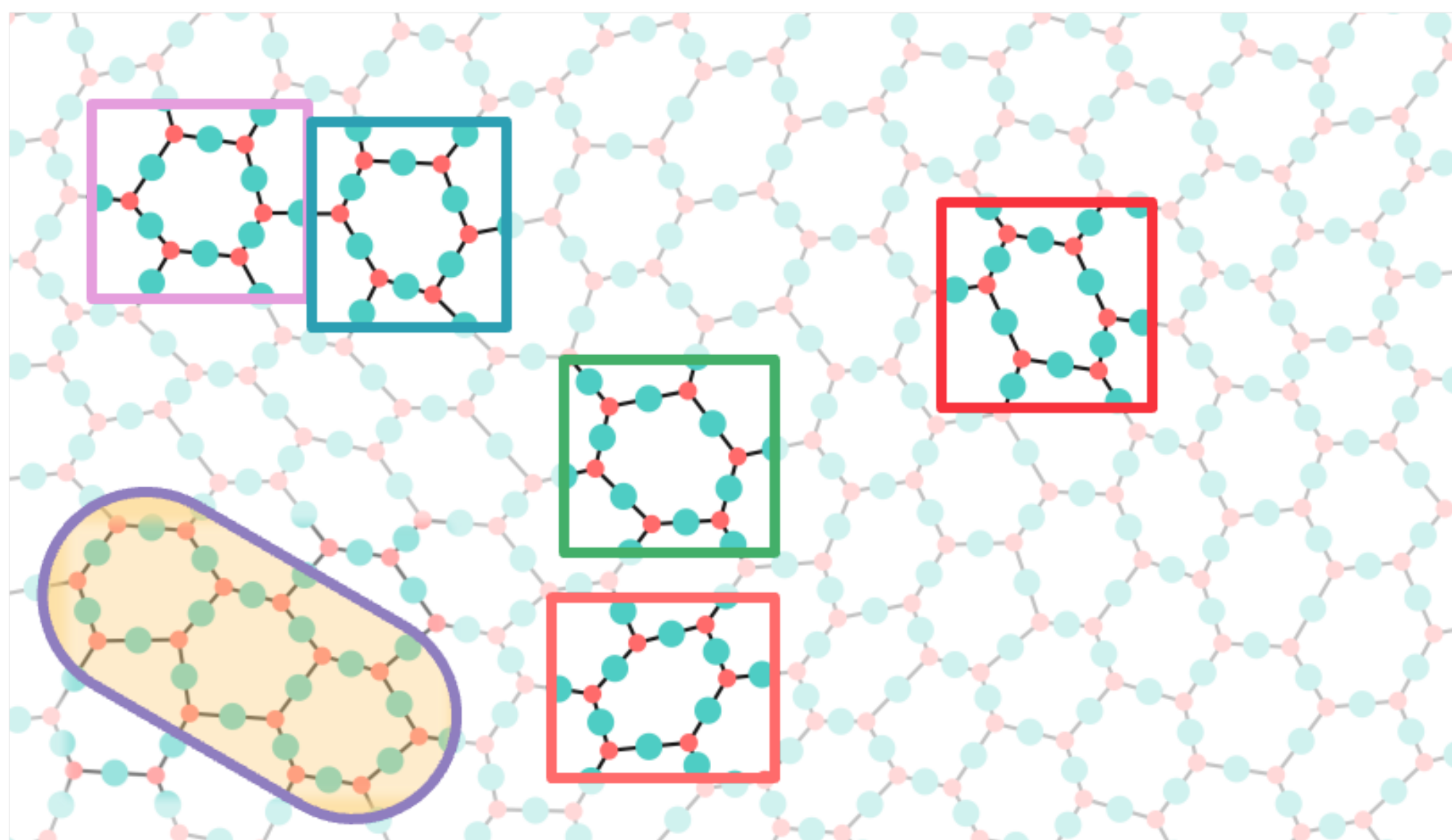
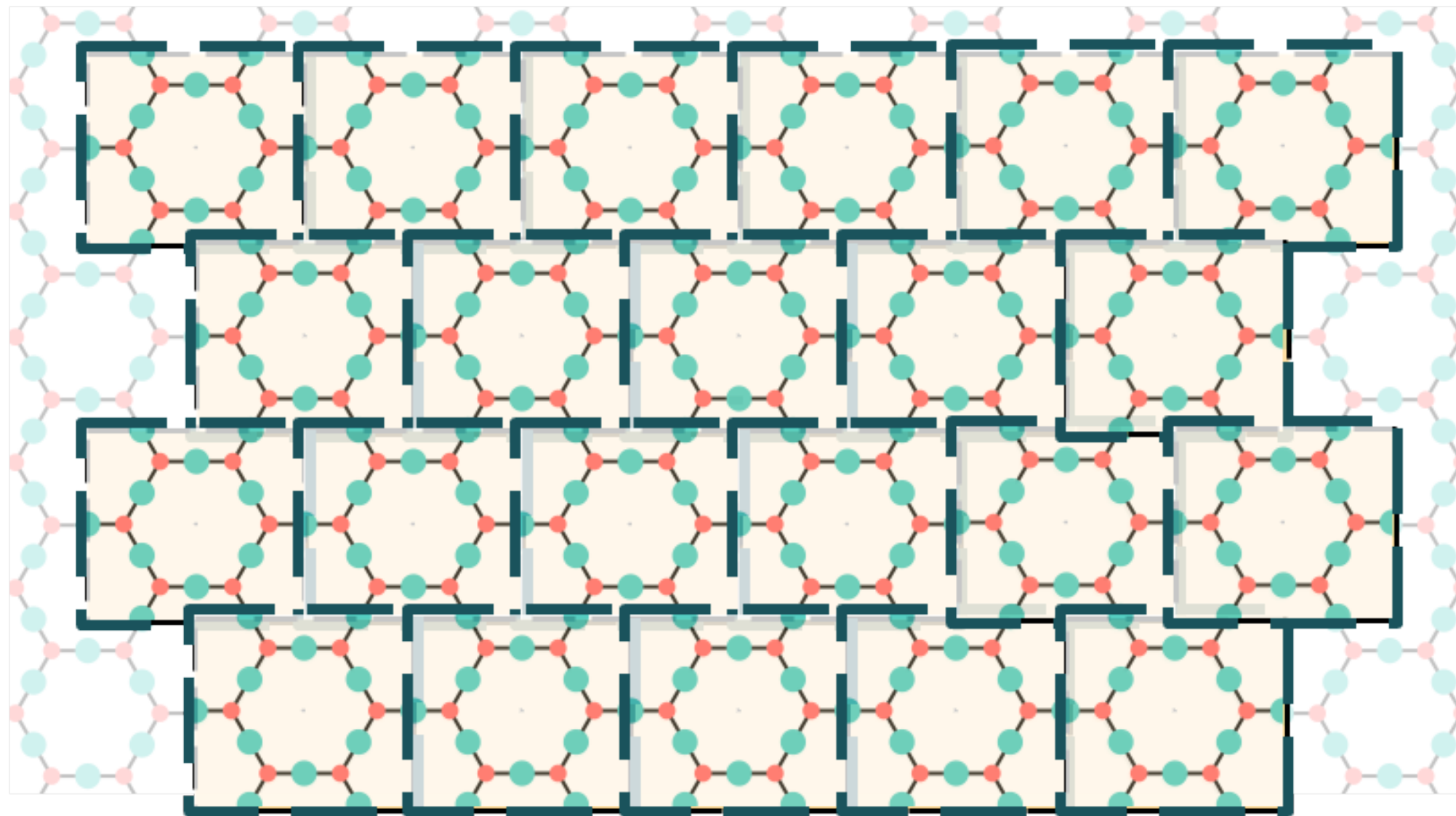
[Yu-Hsueh Chen, Tarun Grover https://arxiv.org/abs/2502.18898](https://arxiv.org/abs/2502.18898)



Ordered phase (Crystalline SiO₂)



Disordered phase (Amorphous SiO₂)



⋮

Compression detects entropy

Shannon's source coding theorem

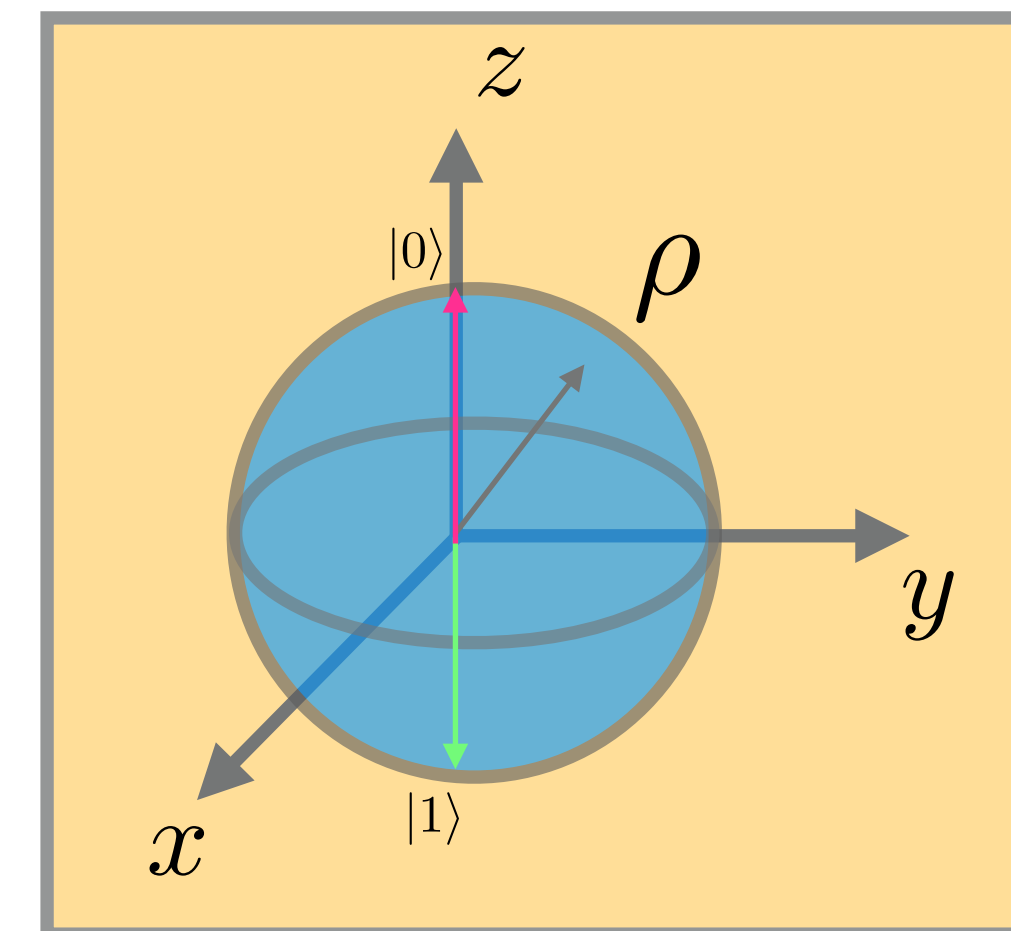
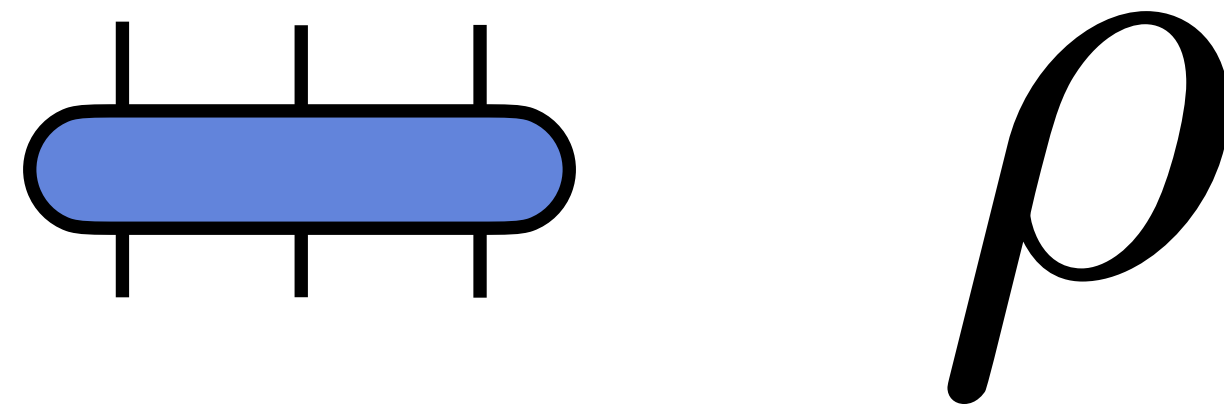
- A source of symbols (e.g. spin configurations) that looks random but has hidden statistical structure can be compressed down to its **entropy**, but not below it.
- Shannon's source coding theorem is basically the statement that **entropy is the ultimate compression limit**.
- Entropy of the distribution of outcomes knows about the structure of the data.
- $\mathbb{E} [|C(X)|] \geq H(X)$
- C is any encoding of symbol X , H is the entropy of the distribution producing X .
- $|C(X)|$ is the size/length of the the encoding of X .

**Are these observations useful for
quantum systems?**

QUANTUM STATES, MEASUREMENTS, AND PROBABILITY DISTRIBUTIONS

HOW IS A QUANTUM STATE TRADITIONALLY DESCRIBED?

- A **density matrix** describes the statistical state of a system in quantum mechanics. Everything we can possibly know about a quantum system is encoded in the density matrix.
- A quantum state is a positive semidefinite, Hermitian operator of trace 1 acting on the state space.
- The family of quantum states forms a convex set. For one qubit: Bloch sphere.



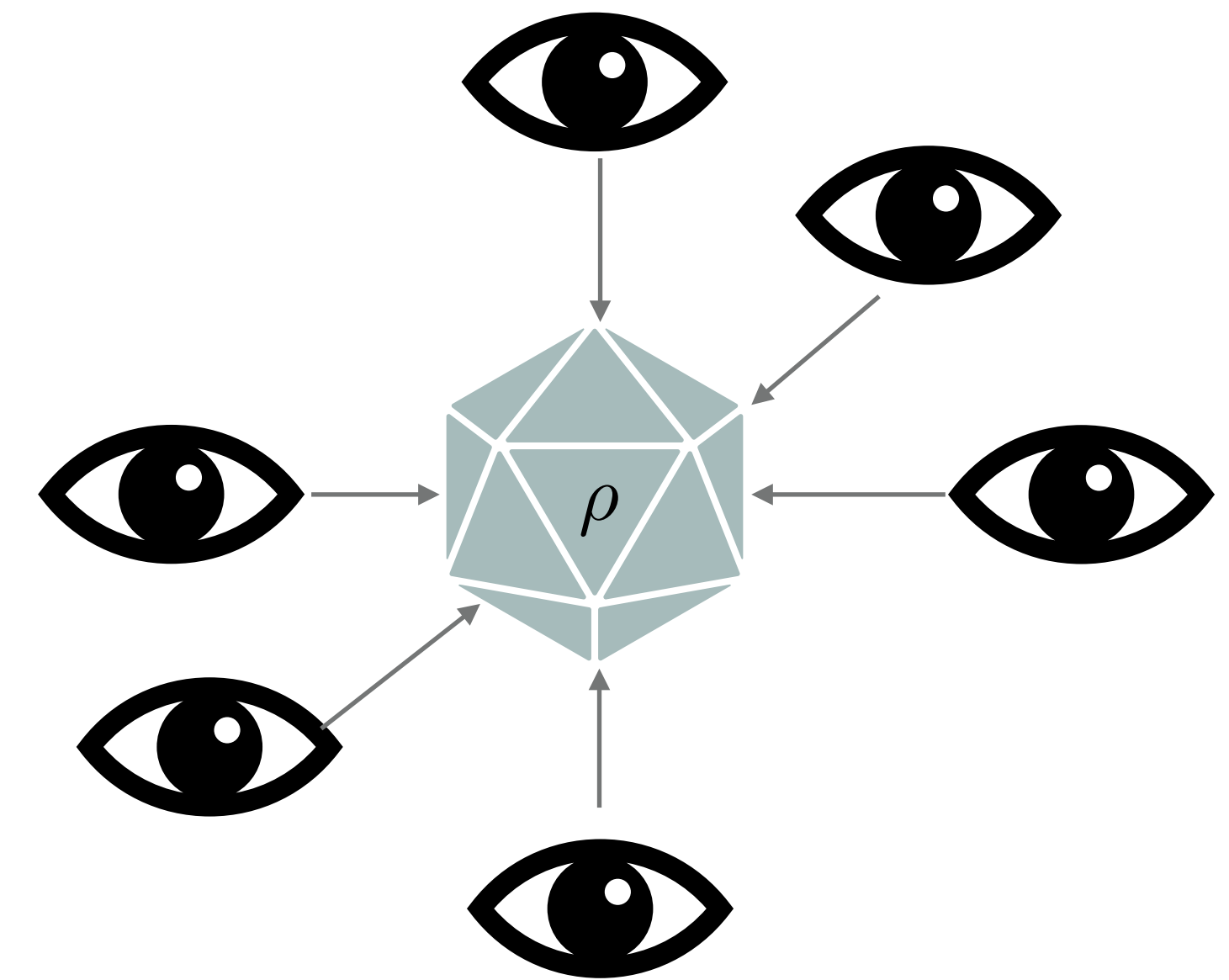
**HOW TO REPRESENT A
QUANTUM STATE WITH
ONLY PROBABILITY?**

MEASUREMENTS: POSITIVE OPERATOR-VALUED MEASURE (POVM)

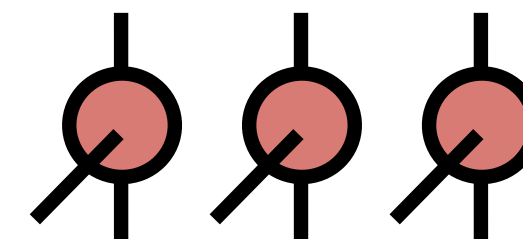
- POVM elements $\mathbf{M} = \{M^{(a)} \mid a \in \{1, \dots, m\}\}$
- Positive semidefinite operators $\sum_i M^{(a)} = \mathbb{1}$
- $\mathbf{M}$ describe the experimental measurements
- Born Rule $P(a) = \text{Tr} [\rho M^{(a)}]$ quantum theory $\longleftrightarrow$ experiment

INFORMATIONALLY COMPLETE MEASUREMENTS

- The measurement statistics $P(a)$ contains all of the information about the state.
- Relation between ρ and distribution $P(a)$ can be inverted



$$\mathbf{M} = \{M^{(a_1)} \otimes M^{(a_2)} \otimes \dots \otimes M^{(a_N)}\}_{a_1, \dots, a_N}$$



INVERTING BORN RULE

BORN RULE

$$P(\mathbf{a}) = \text{Tr} \rho M^{\mathbf{a}}$$

INFORMATIONALLY COMPLETENESS —> THIS RELATION
CAN BE INVERTED

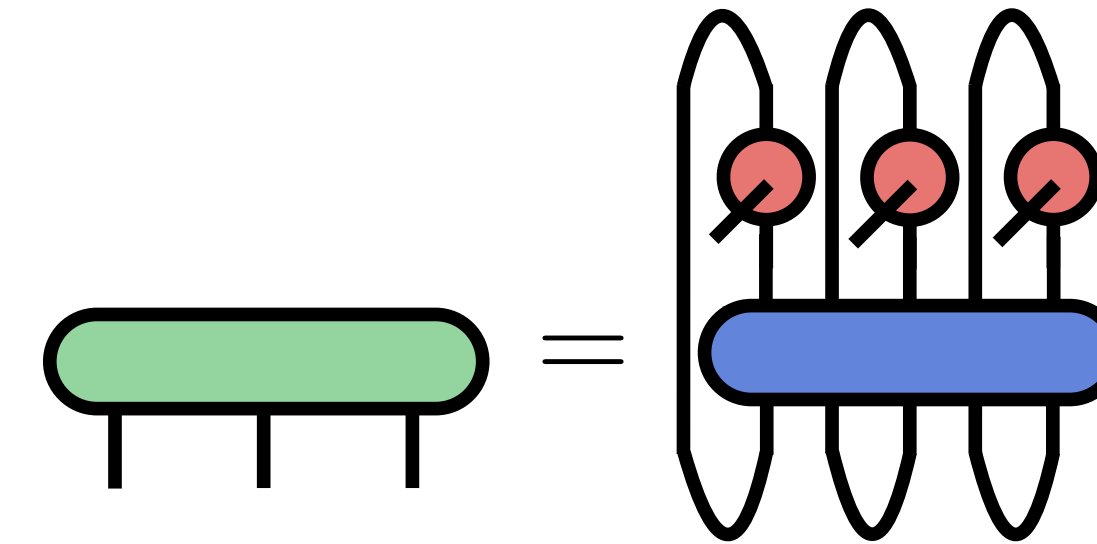
$$\rho = \sum_{a, a'} T_{a, a'}^{-1} P(a') M^{(a)}$$

$$T_{a, a'} = \text{Tr}[M^{(a)} M^{(a')}]$$

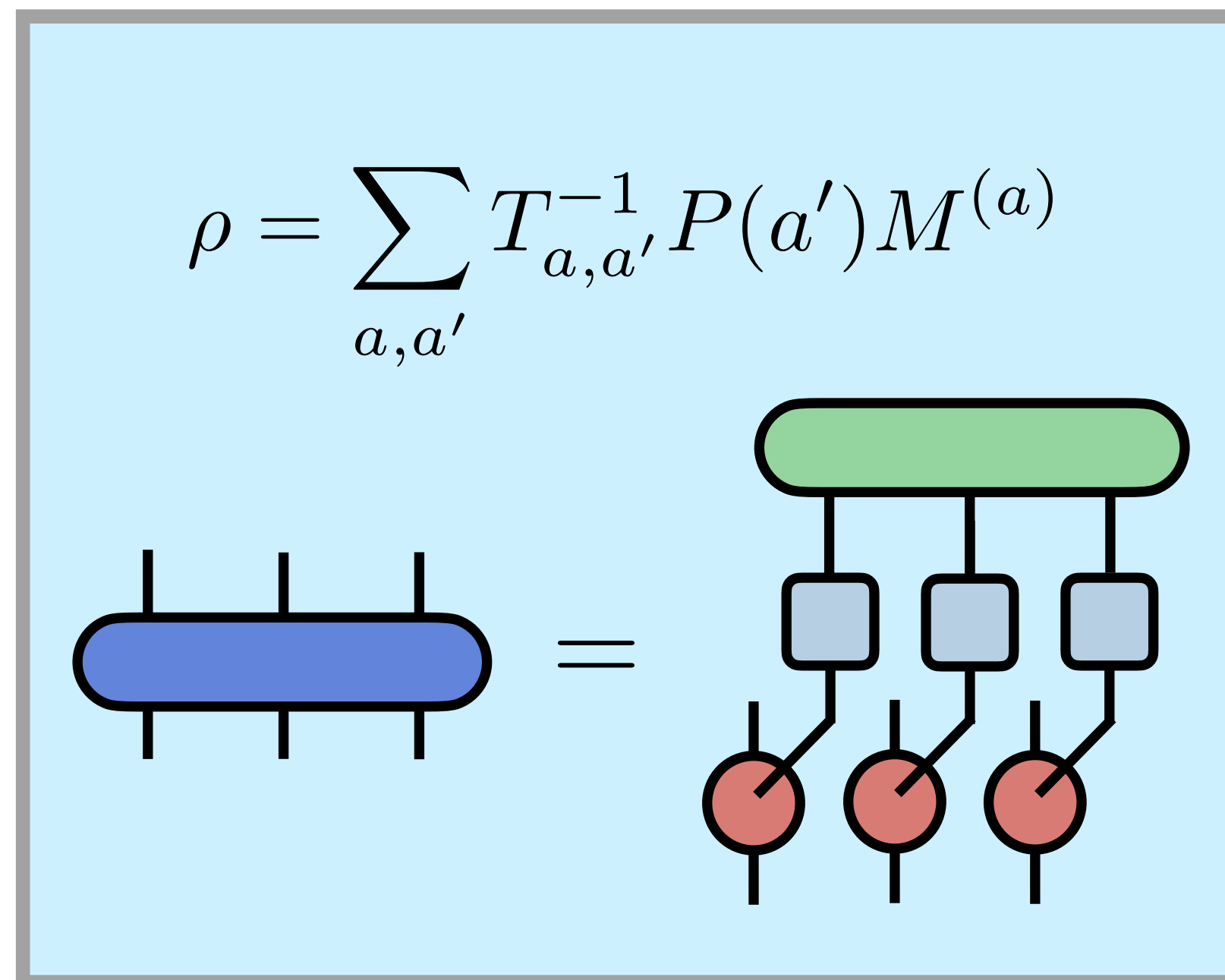
- Factorization of the state in terms of a probability distribution and a set of single-qubit tensors.
- All the entanglement and potential complexity of the state comes from the structure of the $P(a)$
- Tomography
- Real-time dynamics of close and open systems
- Measurements
- Etc

GRAPHICAL NOTATION AND INVERSE

BORN RULE $P(\mathbf{a}) = \text{Tr } \rho M^{\mathbf{a}}$



INFORMATIONALLY COMPLETENESS —> THIS RELATION CAN BE INVERTED

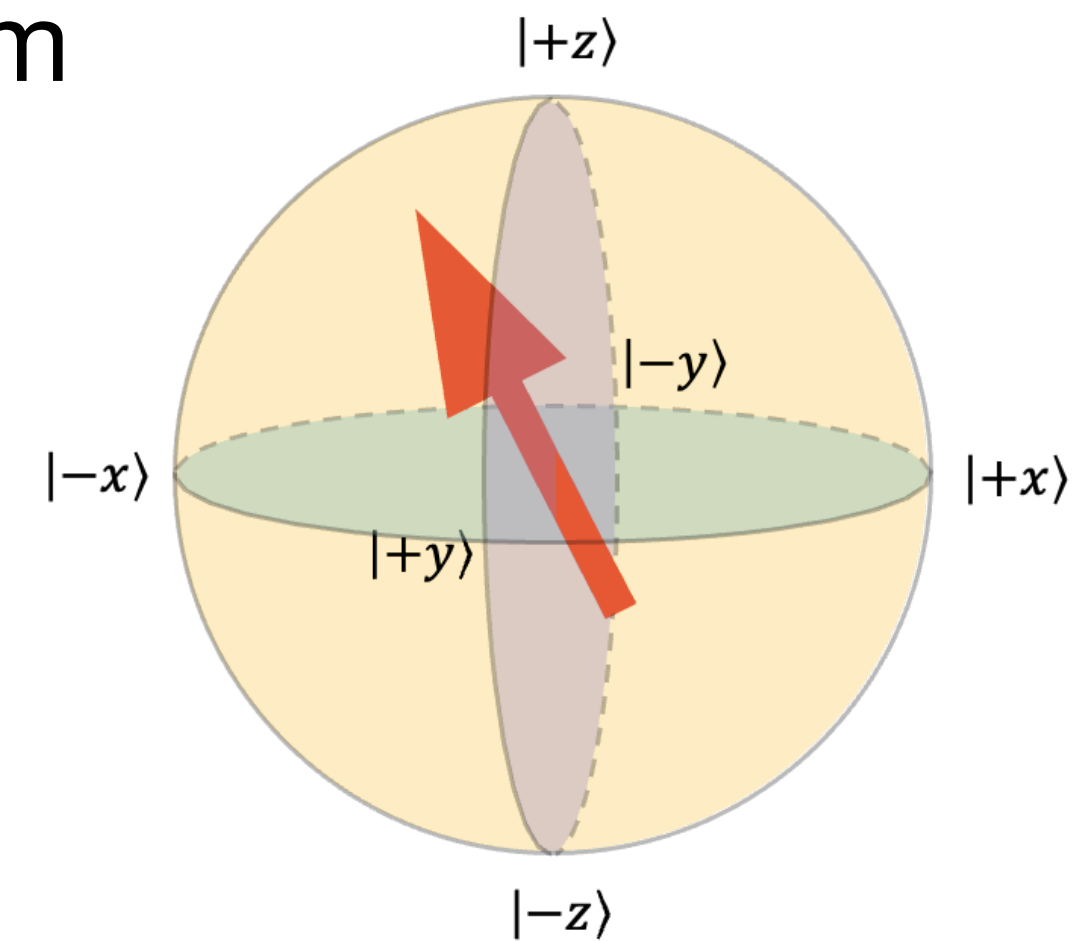


- Unitary evolution
- Schrodinger equation
- Linblad equation
- Measurements

Randomized measurements as IC-POVM

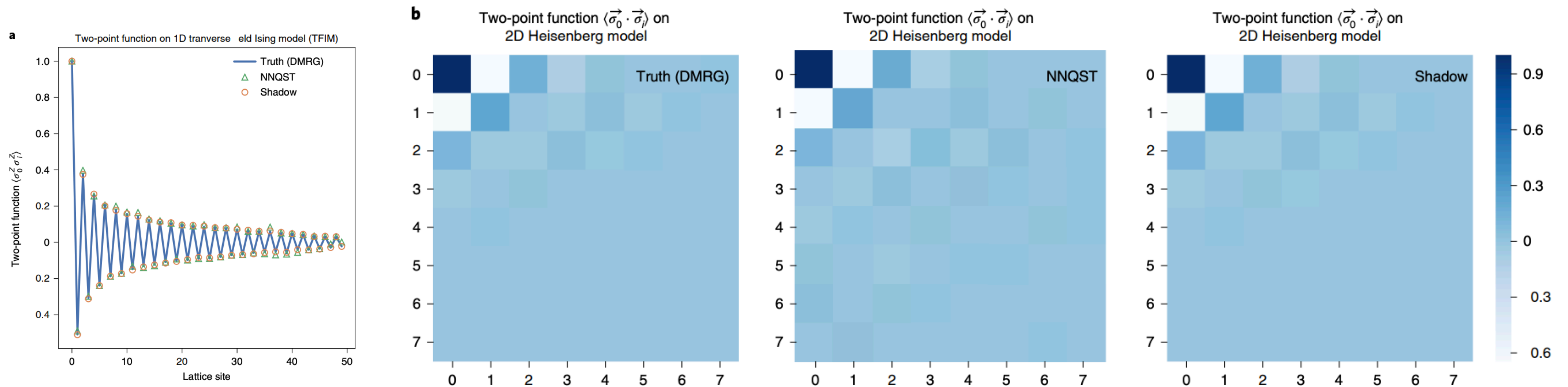
Randomized Pauli measurement

- On each qubit, choose a random direction to measure (x,y,z) with uniform probability.
- Apply unitary (Z basis: I, X basis: H, Y basis: $S^\dagger \rightarrow H$)
- Followed by measurement in computational basis.
- Equivalent POVM:
 - $M^{(0)} = \frac{1}{3} | +z \rangle \langle +z |$, $M^{(1)} = \frac{1}{3} | -z \rangle \langle -z |$, $M^{(2)} = \frac{1}{3} | +x \rangle \langle +x |$, $M^{(3)} = \frac{1}{3} | -x \rangle \langle -x |$, $M^{(4)} = \frac{1}{3} | +y \rangle \langle +y |$, $M^{(5)} = \frac{1}{3} | -y \rangle \langle -y |$,
 - These define a probability distribution $P(a_1, a_2, \dots, a_N)$, $a_i = 0, 1, 2, 3, 4, 5$
 - These were eventually called classical shadows.



Connection to classical shadows

- The data sets collected out of P have been theoretically analyzed in detail — classical shadows.



- “many properties of a quantum state can be accurately and efficiently predicted with a rigorous performance guarantee” In part explains the success of our approach. We are currently studying what generalization in ML model implications.

POVM/Randomized measurements

- All information about the quantum state is in the probability outcomes.
- From samples one can compute a lot of observables.
- Is there anything else can extract than observables on a quantum many-body system?

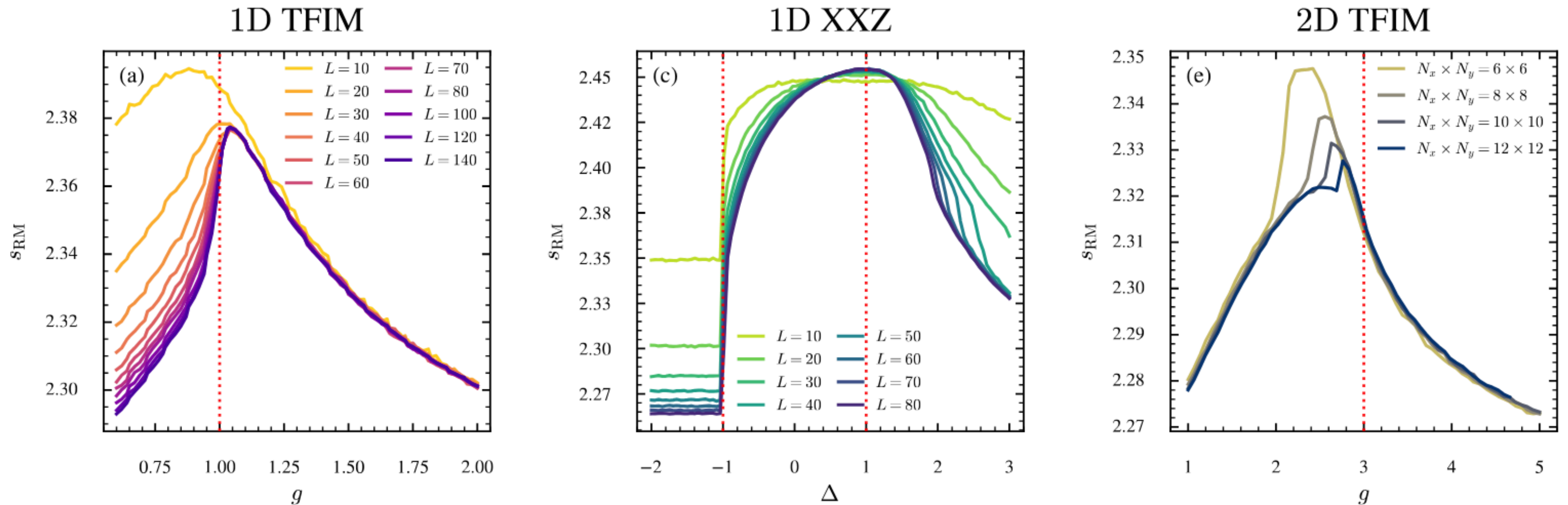
explore the distribution **P** itself

Shannon entropy of randomized outcomes

- The probability of observing a specific measurement outcome $\mathbf{a} = (a_1, \dots, a_L)$ for the state ρ is $P(\mathbf{a}) = \text{Tr}(\rho M^{(a_1)} \otimes \dots \otimes M^{(a_n)})$
- $S_{\text{RM}}(\rho) = - \sum_{\mathbf{a}} P(\mathbf{a}) \log P(\mathbf{a})$

Entropy of prototypical models

- Entropy peaks around the critical points. Entropy via DMRG/Sampling



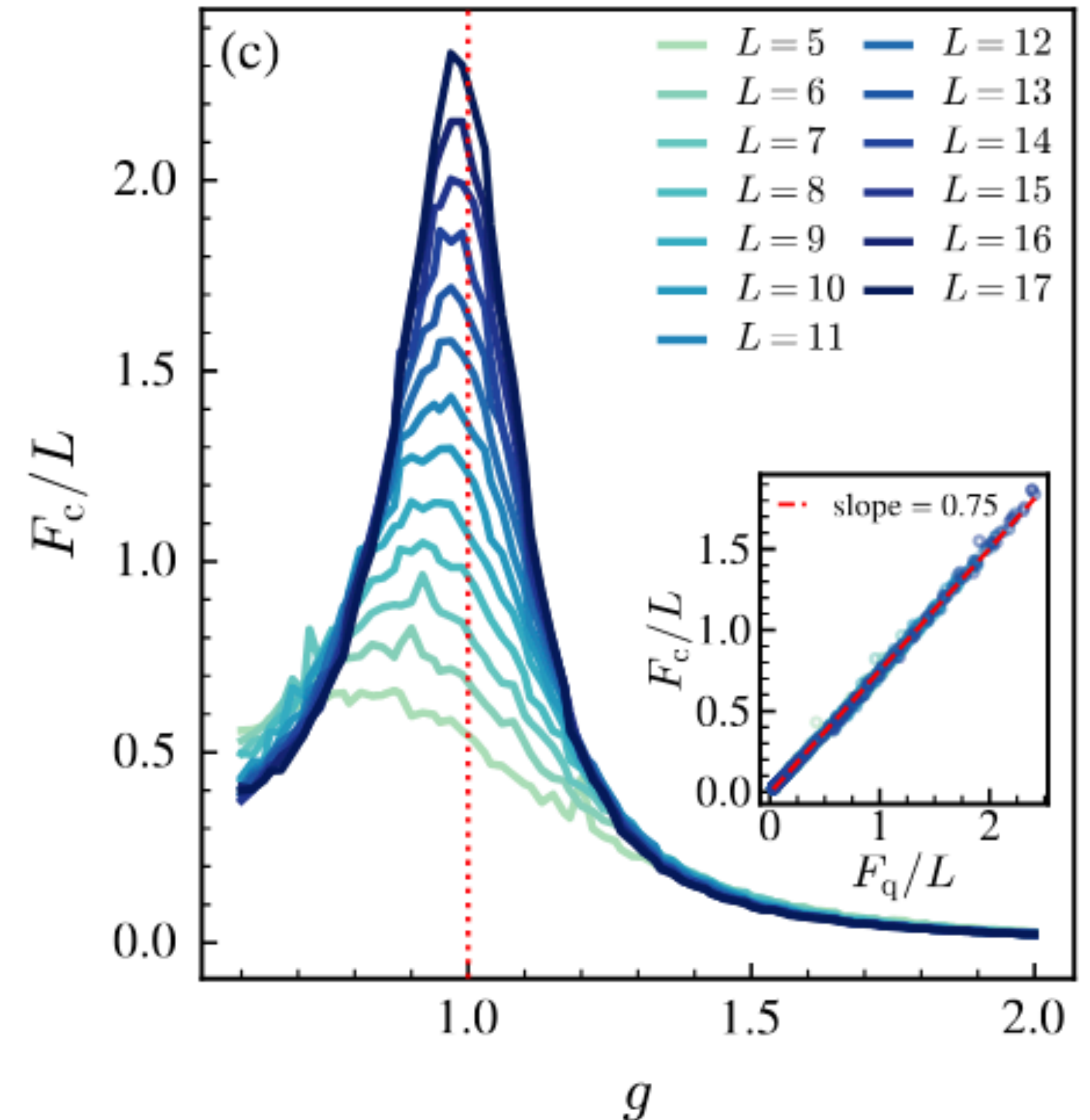
- Curvature gets sharper and shaper $\rightarrow$ Fisher information

Fisher information of P

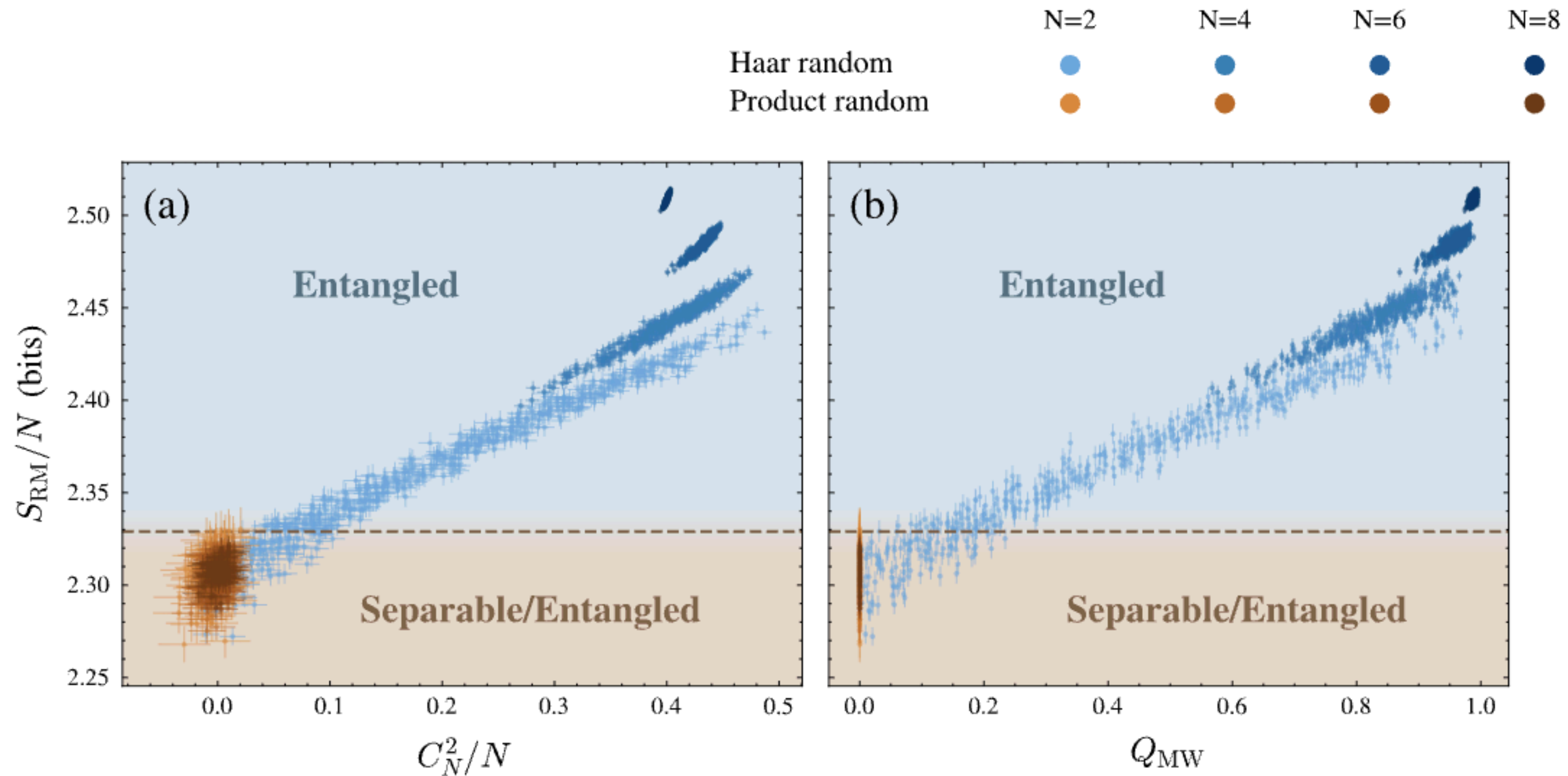
- Fisher information:

- $$F_c = \sum_{\mathbf{a}} P(\mathbf{a}) \left(\partial_g \log P(\mathbf{a}) \right)^2$$

- Behaves the same as fidelity susceptibility/quantum fisher information



Relation to entanglement

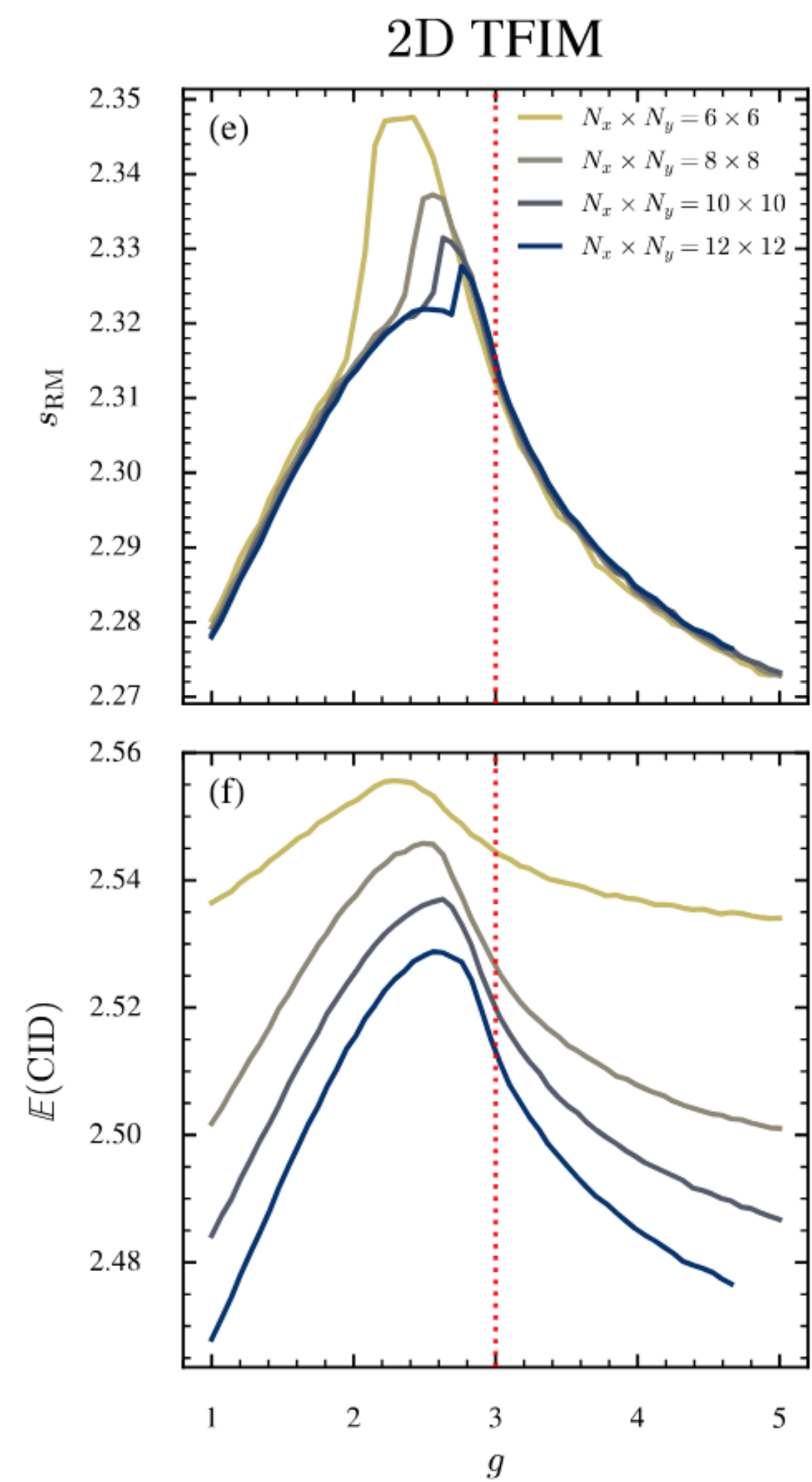
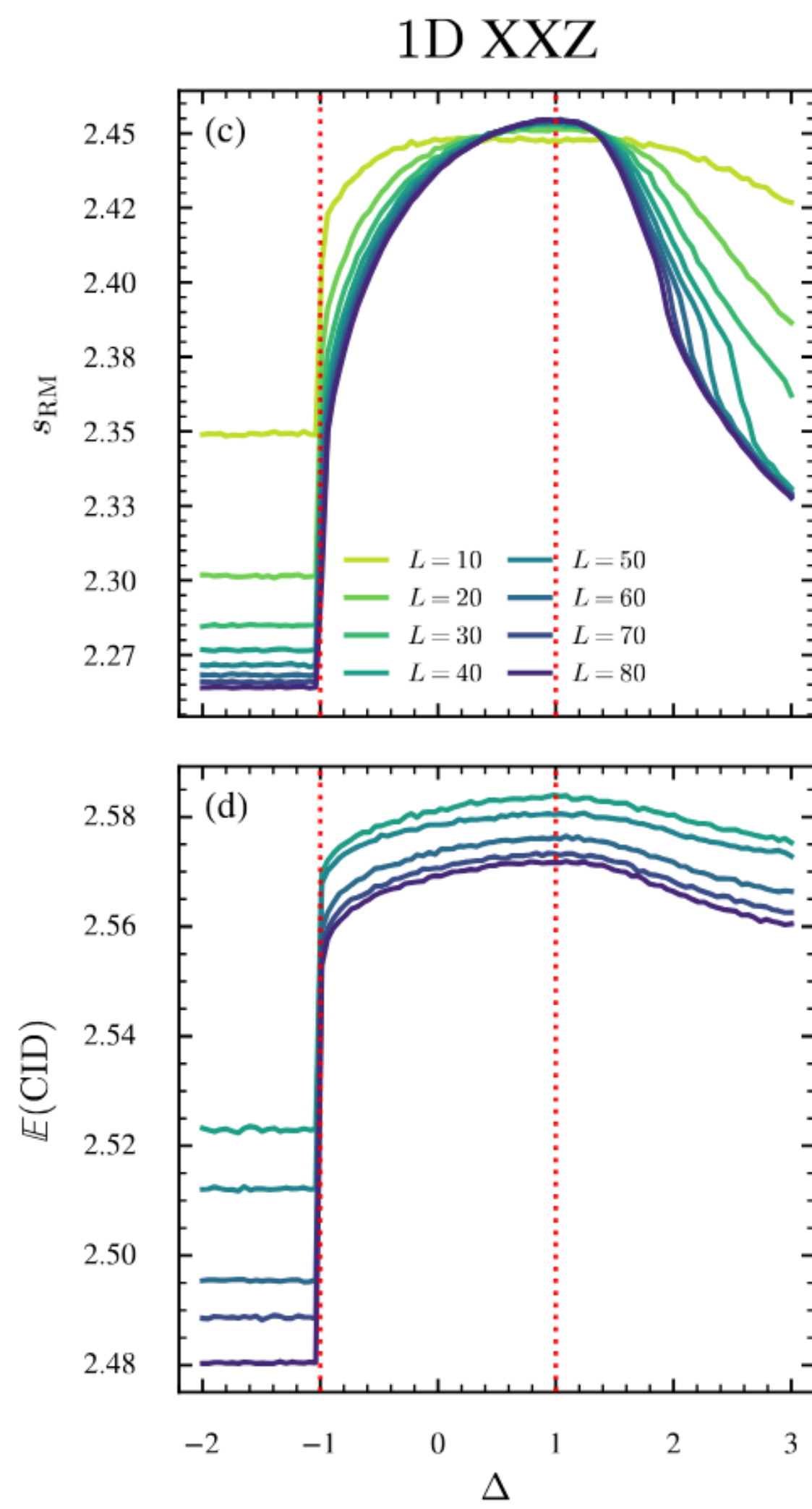
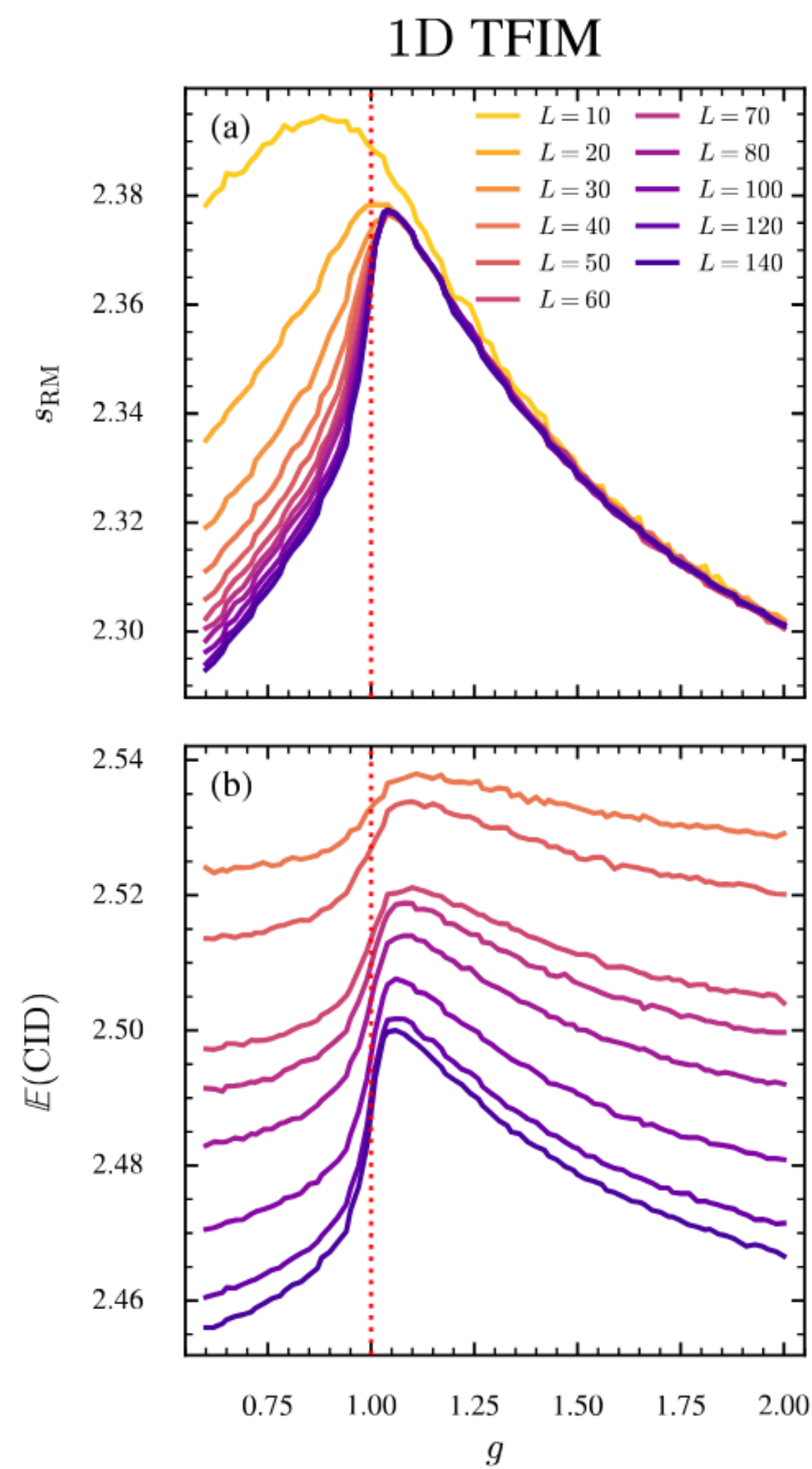


Randomized measurement entropy as a pure-state separability criterion

$$\frac{1}{n} S_{\text{RM}}(|\psi\rangle) \leq \log_2 3 + h\left(\frac{1 + 1/\sqrt{3}}{2}\right) \approx 2.33,$$

**Can we estimate these quantities from
experimental randomized measurements?**

Zip



Can we do better?

Autoregressive models

PROBABILISTIC AUTOREGRESSIVE MODELS

- The term *autoregressive* originates from time-series models: **observations from the previous time-steps are used to predict the value at the current time step.**
- Consider a probability distribution $P(\boldsymbol{\sigma}) = P(\sigma_1, \sigma_2, \dots, \sigma_N)$,

$$P(\sigma_1, \sigma_2, \dots, \sigma_N) = P(\sigma_1)P(\sigma_2|\sigma_1)P(\sigma_3|\sigma_1, \sigma_2) \dots P(\sigma_N|\sigma_1, \sigma_2, \dots, \sigma_{N-1})$$


- To specify P in a tabular form requires **exponential** resources
- To alleviate this exponential issue: parametrize the conditionals

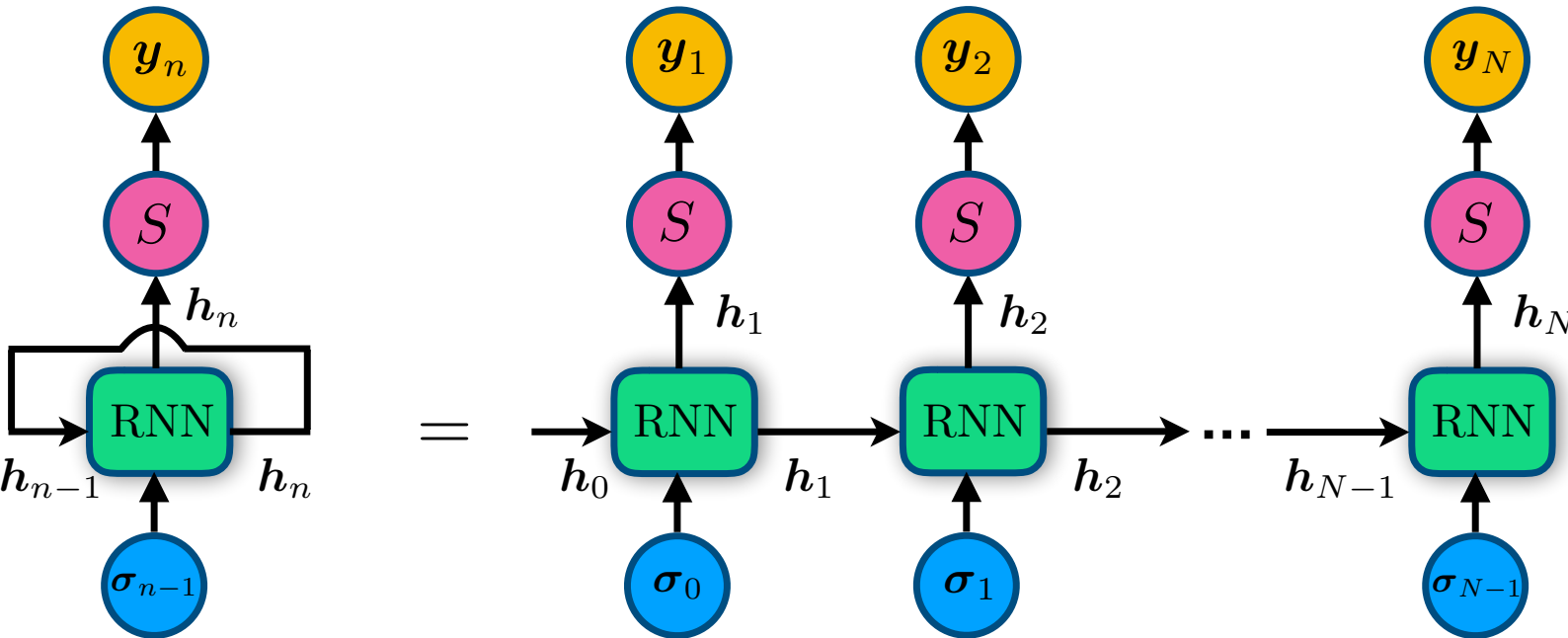
$$P(\sigma_i|\sigma_{<i}) = P_{\theta}(\sigma_i|\sigma_{<i})$$

Autoregressive models and their tractability advantages

- ✓ Computing $P(\boldsymbol{\sigma})$ or the amplitude $\Psi(\boldsymbol{\sigma})$ is efficient given $\boldsymbol{\sigma}$
- ✓ Can be exactly sampled efficiently in linear time in N
- ✓ Can be defined in **any** spatial dimension (e.g. 2d and 3d) — no tractability issues
- ✓ Their expressive power can be systematically improved
- ✓ Easy to encode product states and mean-field theories (e.g. Gutzwiller)
- ✓ We can impose some symmetries and conservation laws

RECURRENT NEURAL NETWORKS (RNN)

➤ The key building block of an RNN is a recurrent cell

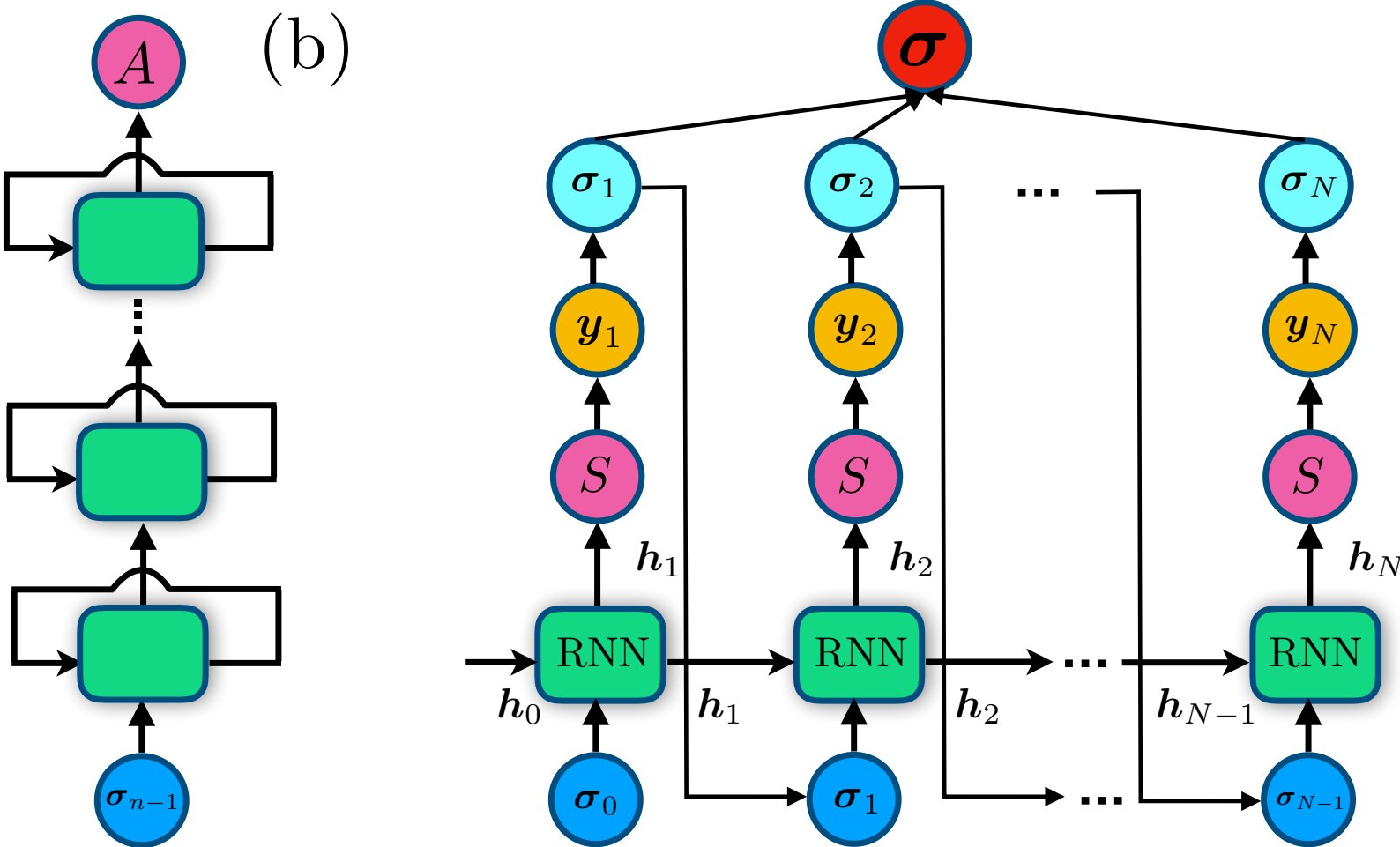


$$h_n = f(W[h_{n-1}; \sigma_{n-1}] + b)$$

$h_n \in \mathbb{R}^d$ — d measures the expressive power of the model

$$y_n \equiv S(Uh_n + c)$$

$$S(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \text{ — Softmax function}$$



$$y_n = P(\sigma_n | \sigma_1, \dots, \sigma_{n-1})$$

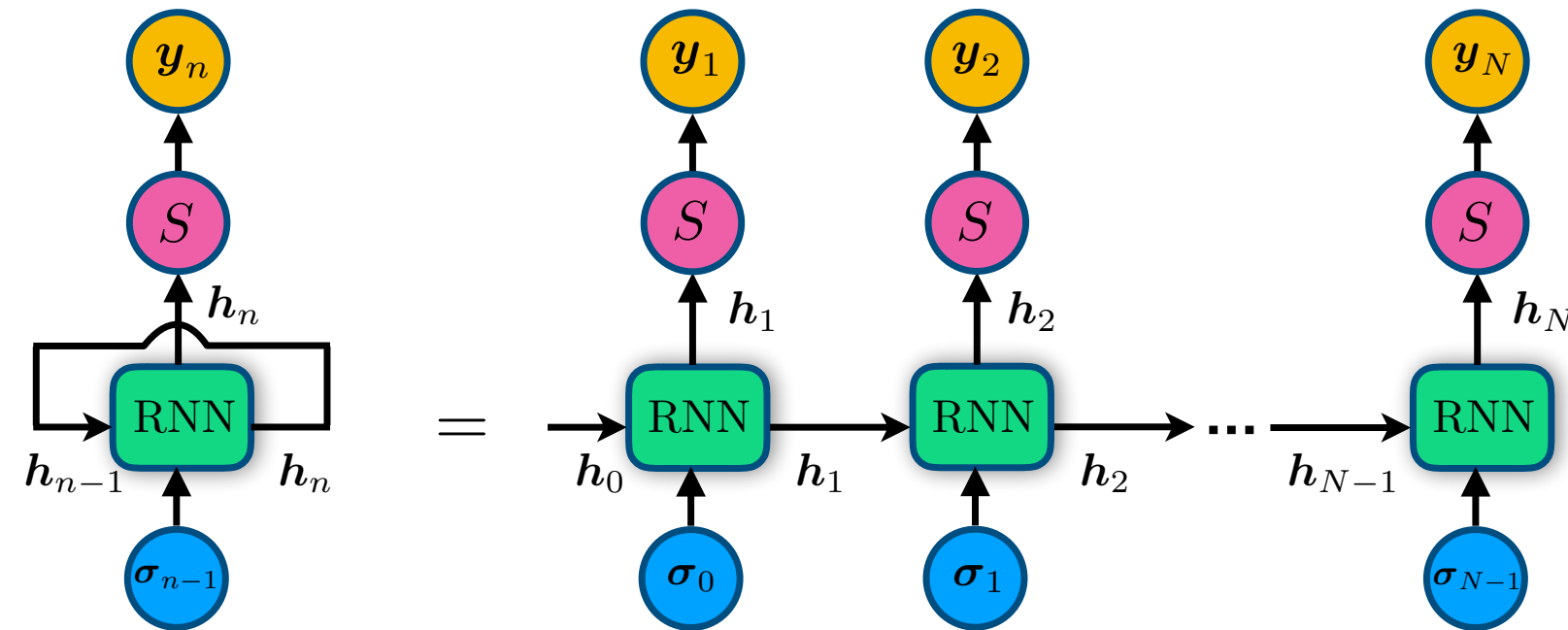
Sampling:

- Sample each conditional
- Input the sample to the next step

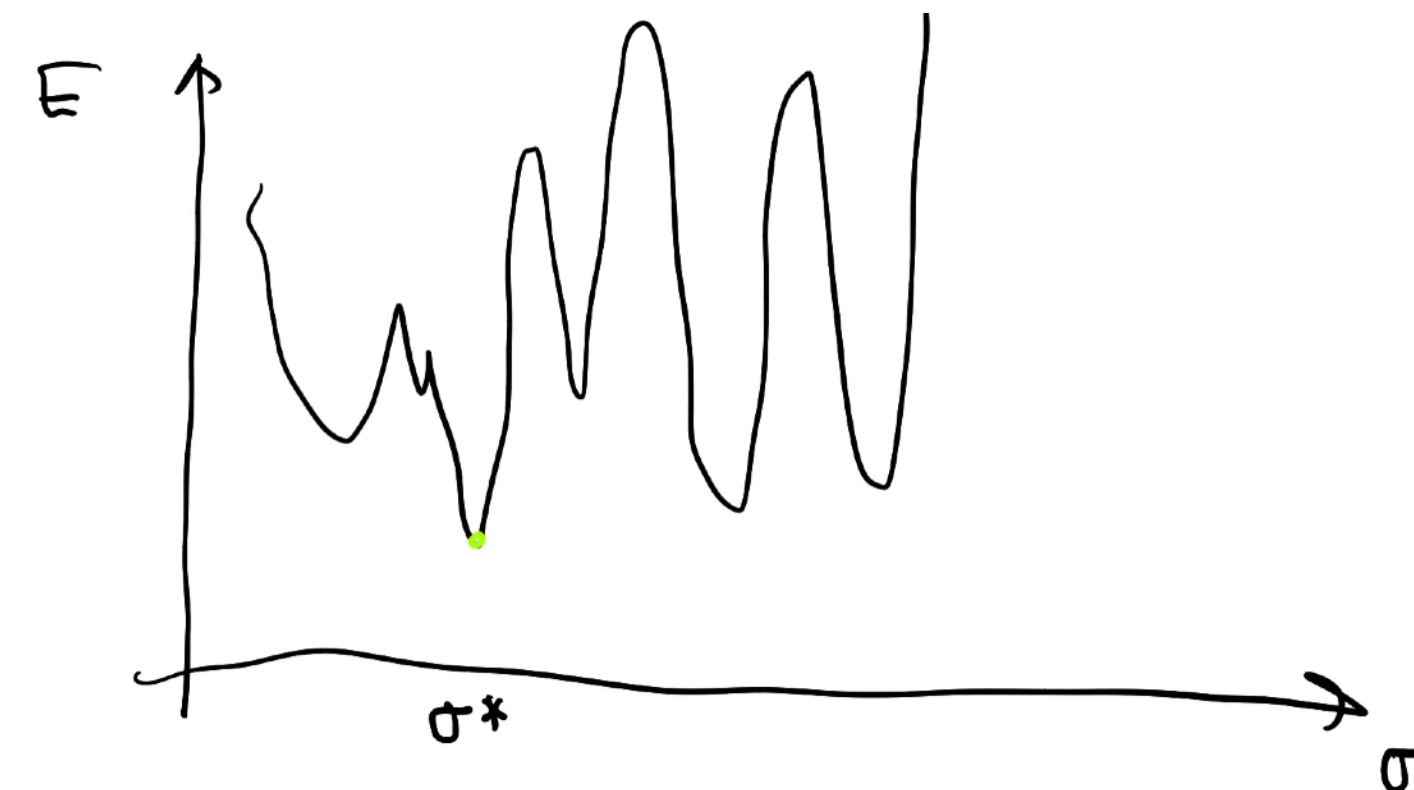
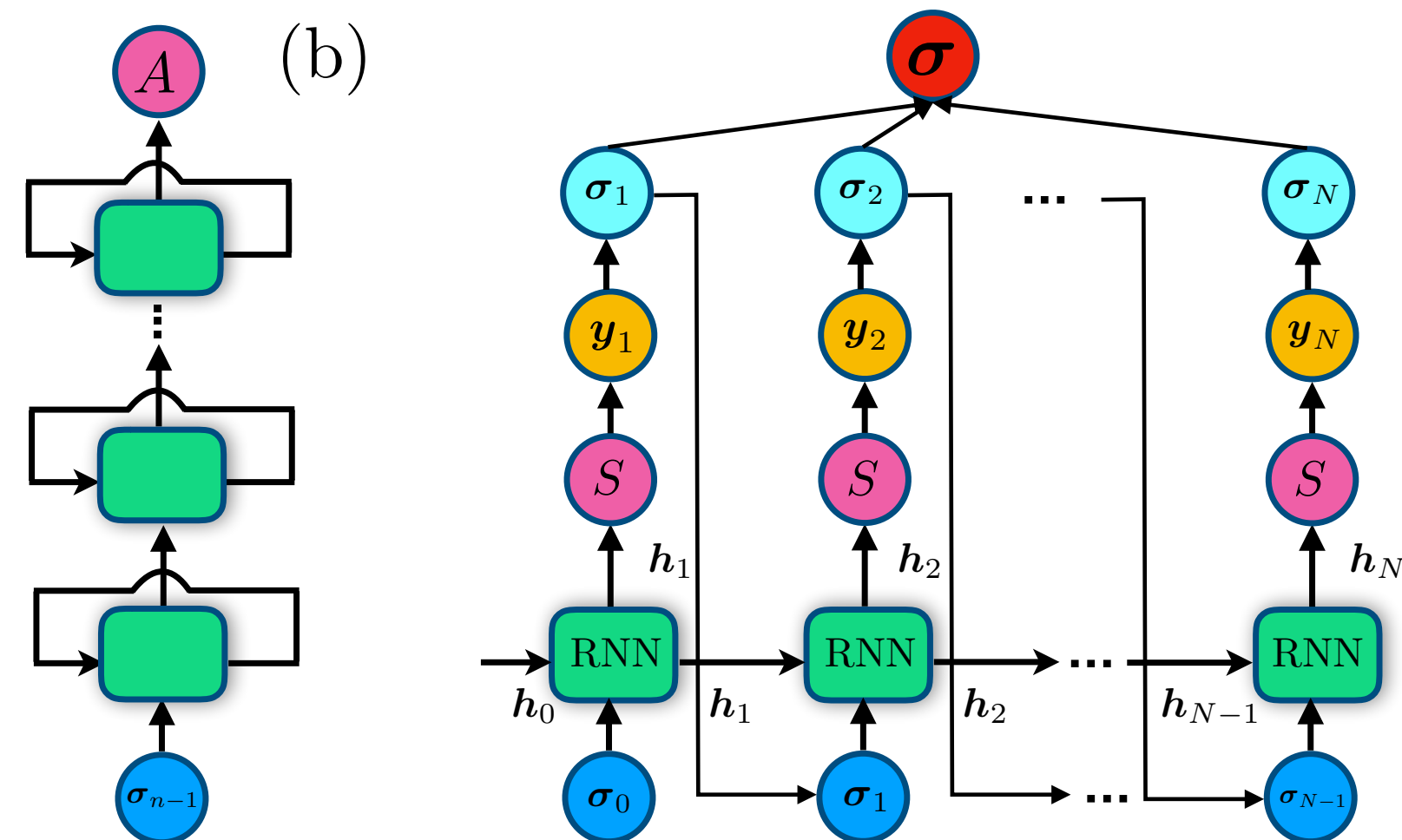
$$P(\sigma_1, \sigma_2, \dots, \sigma_N) = P(\sigma_1)P(\sigma_2|\sigma_1)P(\sigma_3|\sigma_1, \sigma_2) \dots P(\sigma_N|\sigma_1, \sigma_2, \dots, \sigma_{N-1})$$

RECURRENT NEURAL NETWORKS (RNN)

- The key building block of an RNN is a recurrent cell



- Key insight about this model choice:
- It can be sampled exactly even if it encodes a rough landscape.

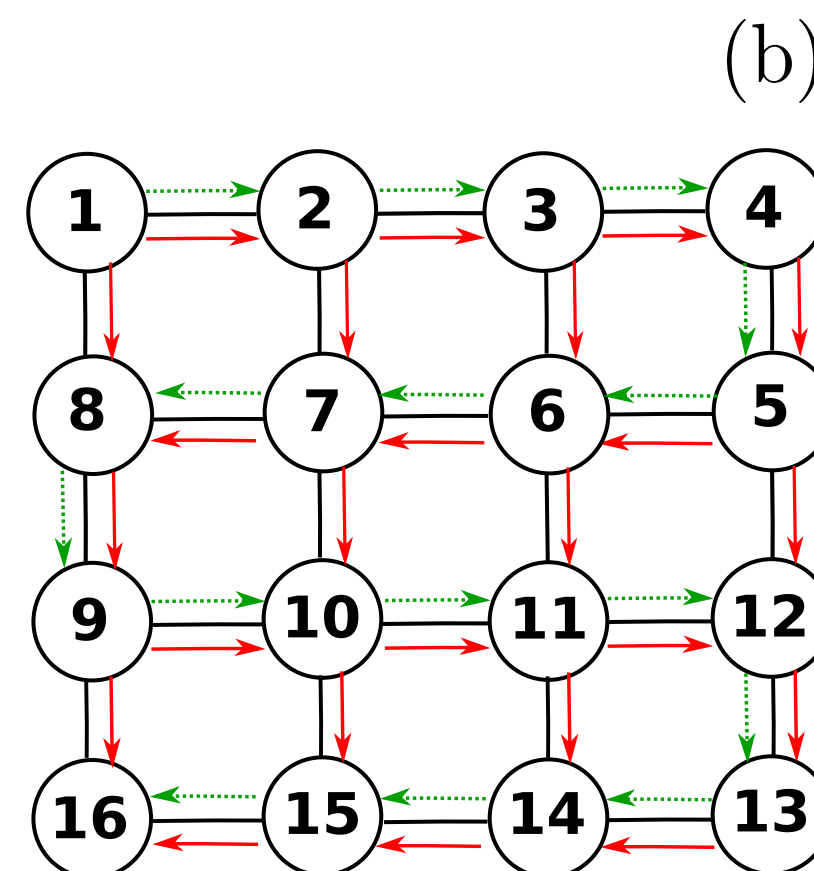
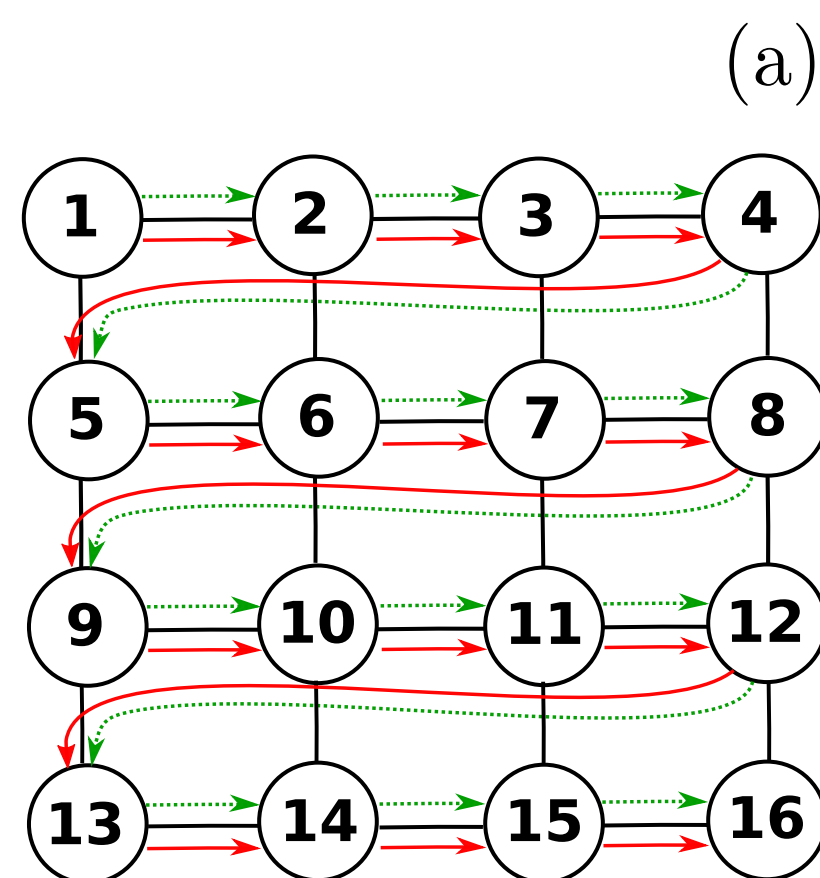


$$P(\sigma_1, \sigma_2, \dots, \sigma_N) = P(\sigma_1)P(\sigma_2|\sigma_1)P(\sigma_3|\sigma_1, \sigma_2) \dots P(\sigma_N|\sigma_1, \sigma_2, \dots, \sigma_{N-1})$$

Two dimensional RNNs

- Replace the single hidden state that is passed from one site to another by two hidden states, with each one corresponding to the state of a neighbouring site (vertical and horizontal) — 2D geometry of the problem

- $$\mathbf{h}_{i,j} = f\left(W^{(h)}[\mathbf{h}_{i-1,j}; \boldsymbol{\sigma}_{i-1,j}] + W^{(v)}[\mathbf{h}_{i,j-1}; \boldsymbol{\sigma}_{i,j-1}] + \mathbf{b}\right)$$



No tractability issues, can be defined in any spatial dimension

Parametrize statistics of measurements and invert

$$P(\mathbf{a}) = \text{Tr } \rho M^{\mathbf{a}}$$

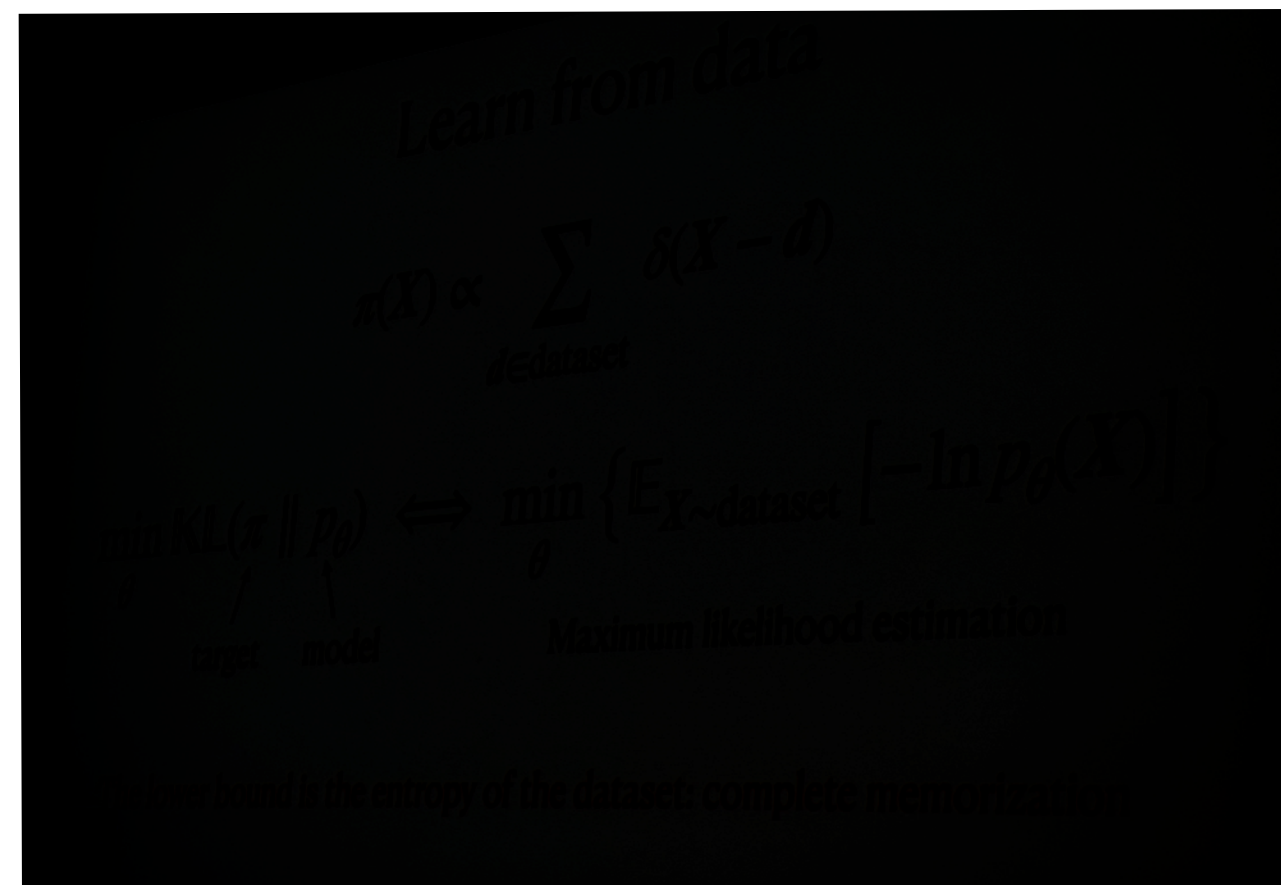
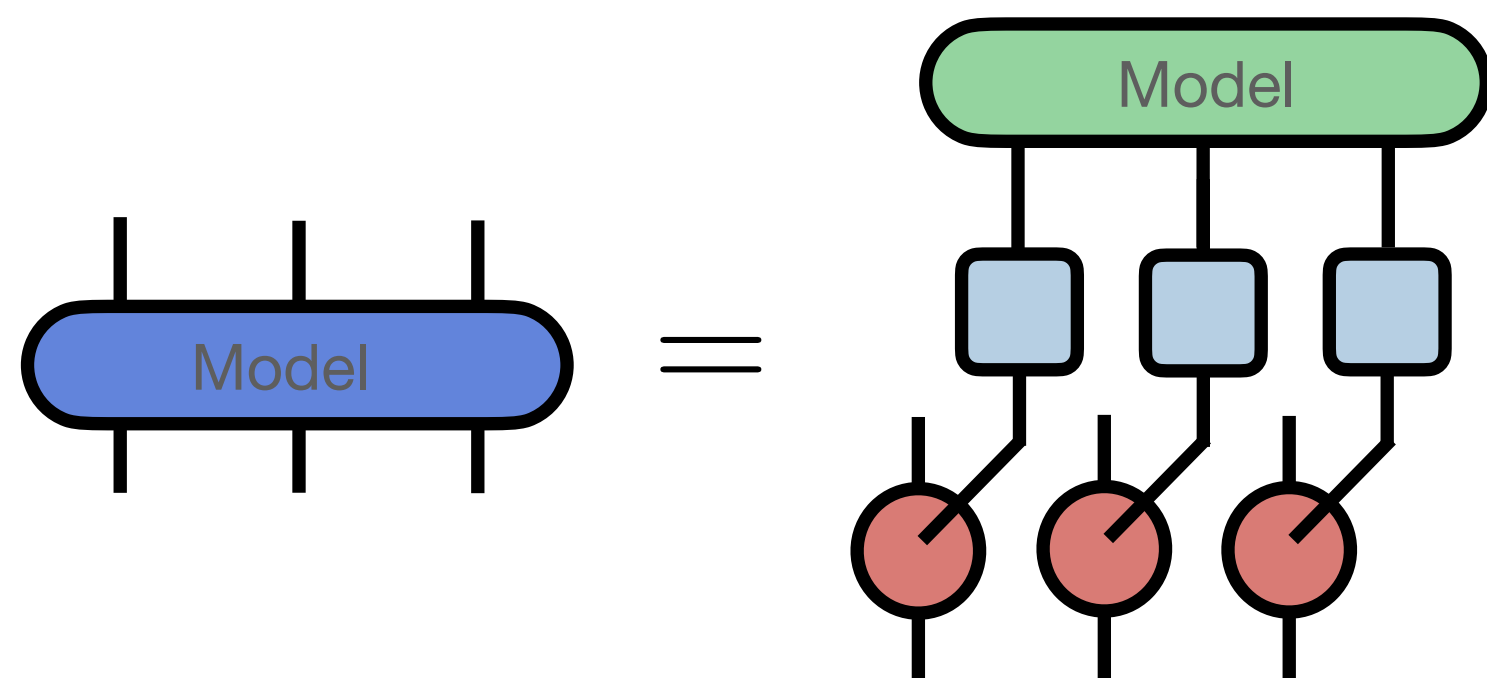
$$P_{\text{model}}(\mathbf{a})$$

Create an autoregressive model of $P(\mathbf{a})$

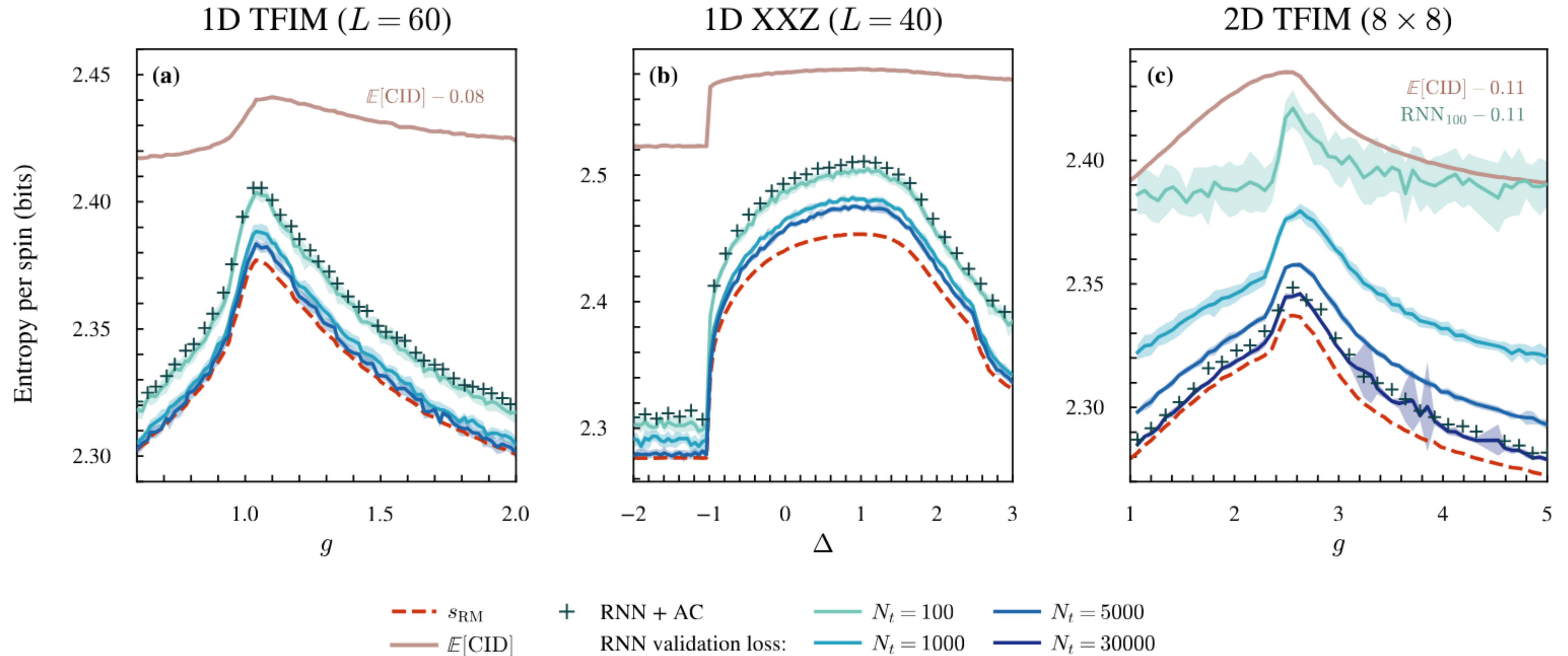
Autoregressive models (RNNs and transformer)

1. Allow for exact sampling
2. Tractable density $P_{\text{model}}(\mathbf{a})$
3. Use maximum likelihood estimation to infer the parameters of the model so that the distribution explains the dataset of measurements.

$$\rho_{\text{model}} = \sum_{a,a'} T_{a,a'}^{-1} P_{\text{model}}(a') M^{(a)}$$



Recurrent neural networks



Significant improvement in quality. It keeps improving with more and more training data.

Can we get fisher information?

One transformer model to learn everything

- Previous strategy learned a model for each system size L and hamiltonian parameter g .
- It is possible to create a single model that learns all of the g, L :

- $$p_{\theta}(\mathbf{a} \mid g, L) = \prod_{i=1}^L p_{\theta}(a_i \mid \mathbf{a}_{<i}, g, L)$$

- Architecture choice: transformer, but RNN is also possible. These models can process variable length input — variable system size (Recall Lei Wang's talk again)
- Take all the data and learn a single model.
- Apart from entropy, we estimate Fisher information: $F_{c,\theta}(g, L) = \mathbb{E}_{\mathbf{a} \sim p_{\theta}} \left[\partial_g \log p_{\theta}(\mathbf{a} \mid g, L)^2 \right] .$

Architecture

19

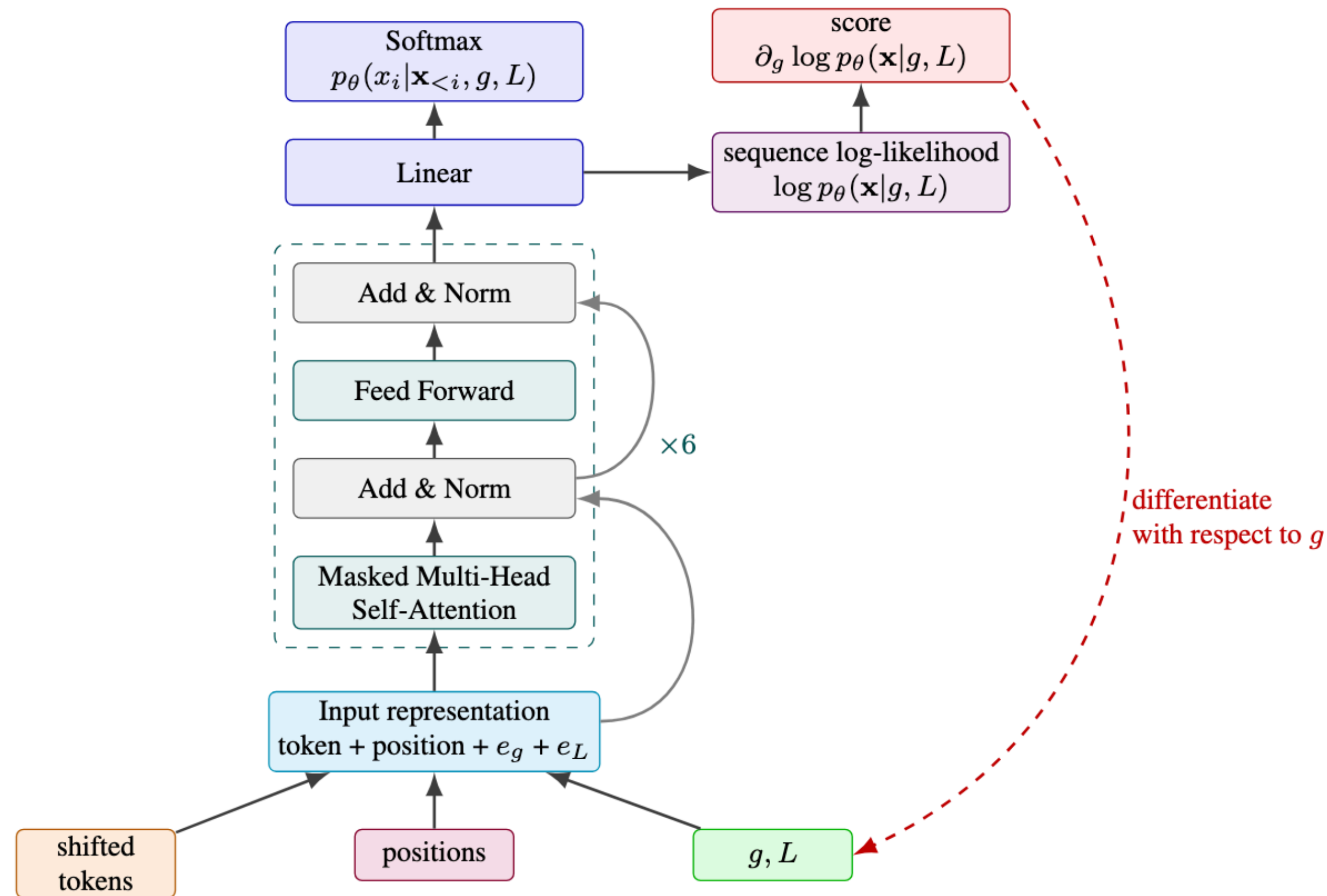
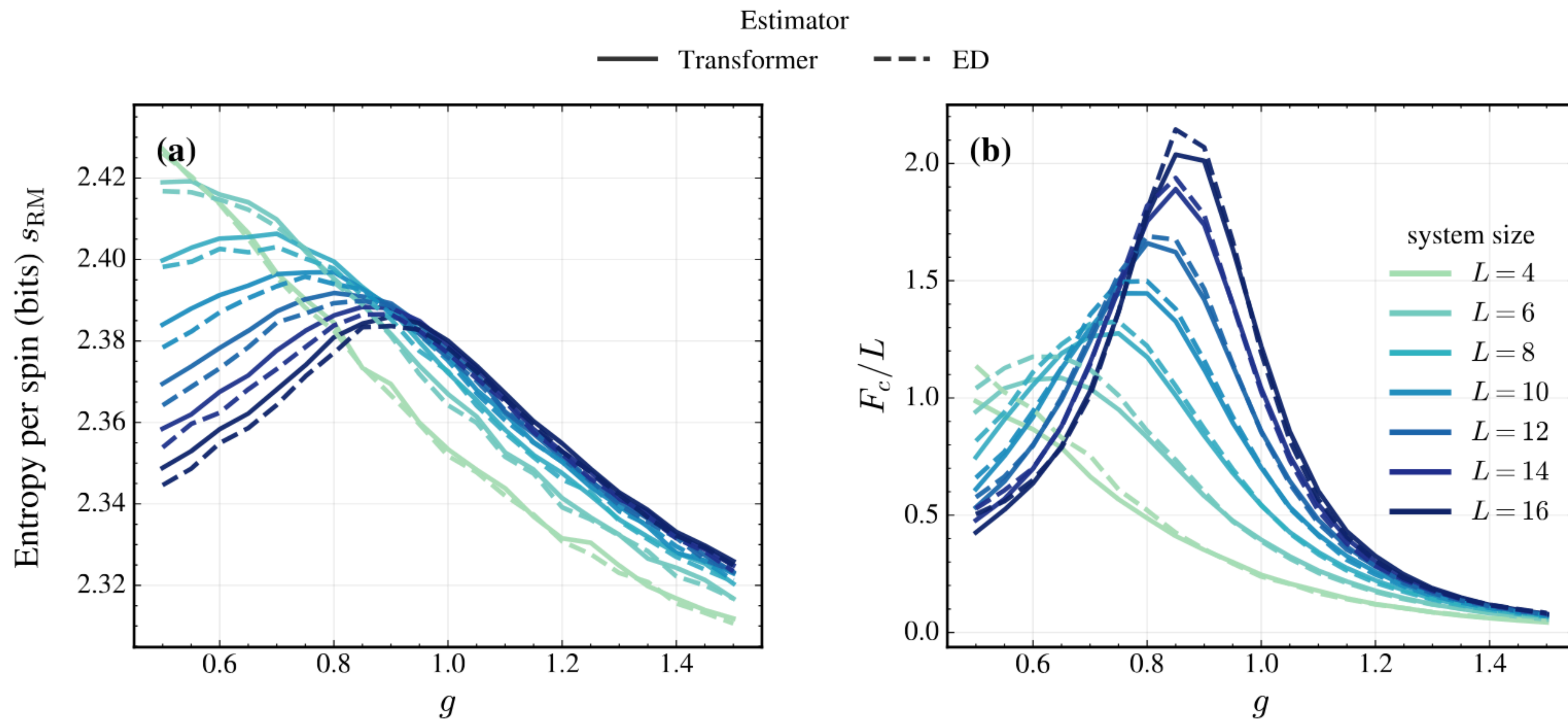


Figure 10. Conditional autoregressive transformer used to model the randomized-measurement distribution. Shifted outcome tokens, lattice positions, and the conditioning variables g, L are embedded and added before the stack of masked self-attention blocks. The entropy estimator uses the sequence log-likelihood, while the Fisher estimator differentiates the same scalar log-likelihood with respect to g .

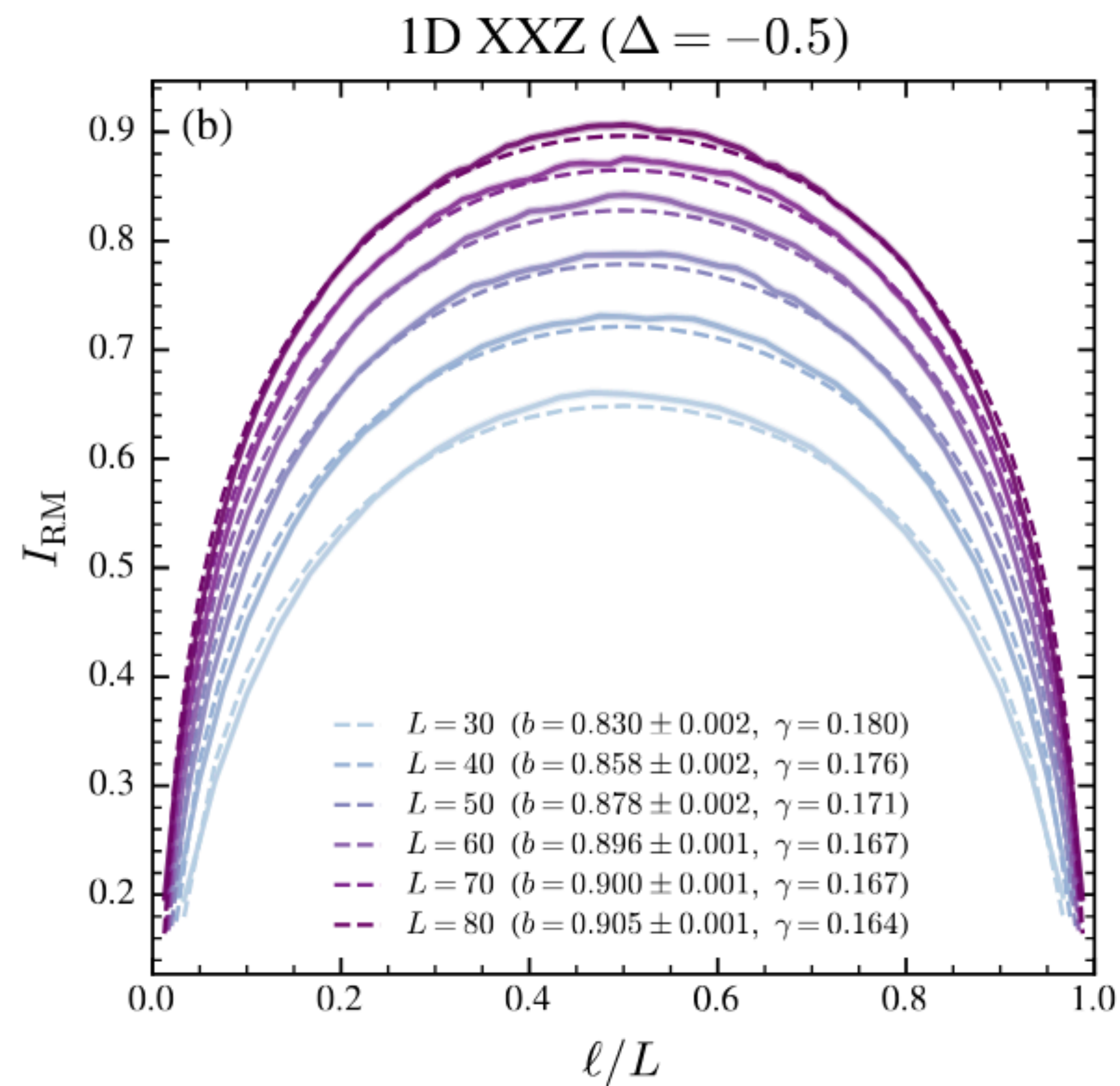
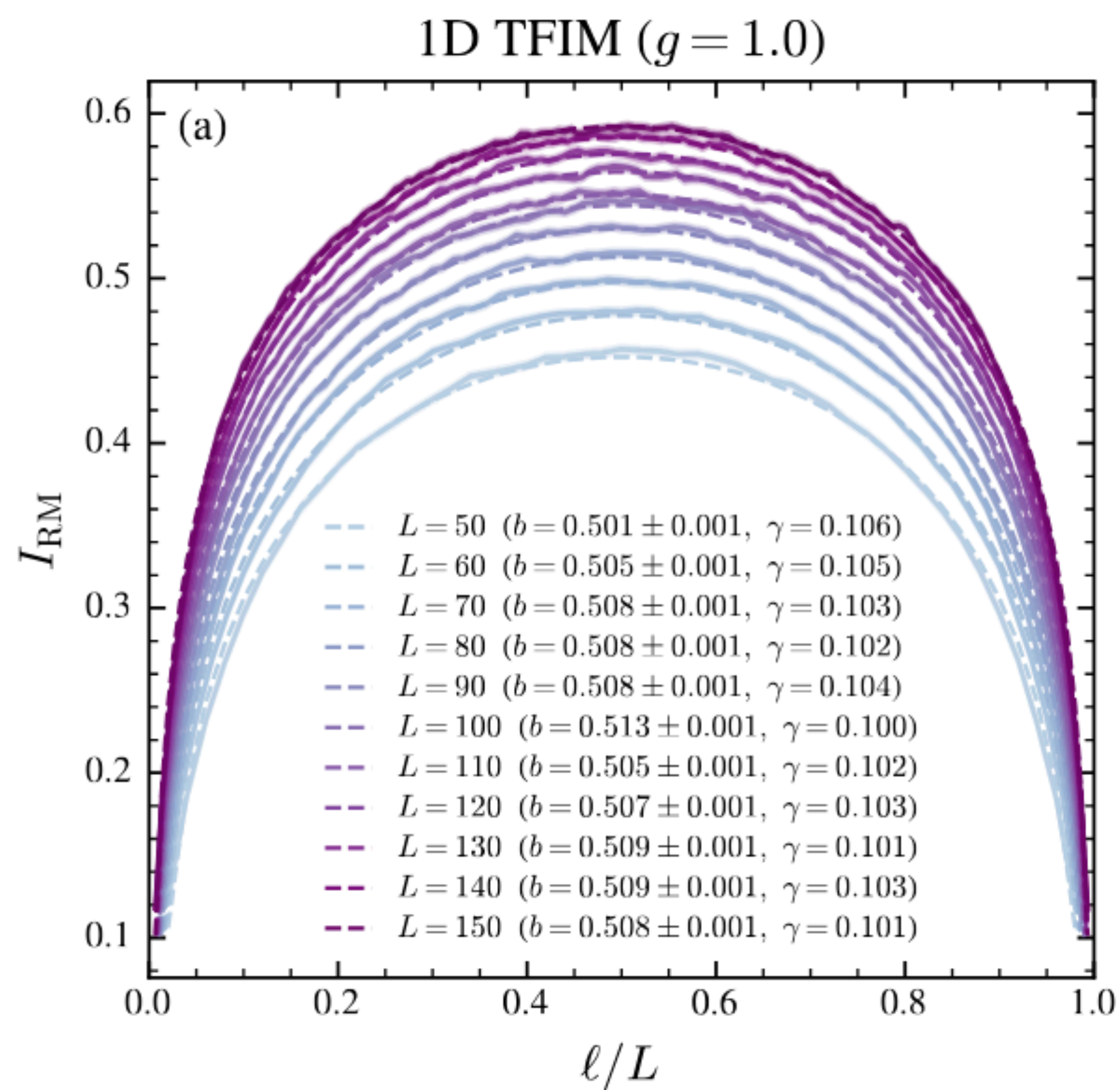
One transformer model to learn everything



Conclusions

- Best estimators of entropy of randomized measurements we have found so far are all neural network based.
- Connected the notion of randomized/POVM entropy to entanglement.
- Fisher information behaves similar to quantum fisher information/fidelity susceptibility.
- We hope this justifies learning from data beyond local observables already computable with only the data.
- We are looking into how this works experimentally.
- Compression is broadly applicable to all sort of data beyond the data treated here.

Mutual information



$$I_{\text{RM}} = \frac{b}{4} \ln \left(\frac{L}{\pi} \sin \left(\pi \frac{\ell}{L} \right) \right) + \gamma.$$

Entanglement measures

- Concurrence

$$C_N(|\psi\rangle) = 2 \sqrt{1 - \frac{1}{2^N} \sum_{A \subseteq \{1, \dots, N\}} \text{Tr}[\rho_A^2]},$$

- Meyer–Wallach global entanglement measure

$$Q(|\psi\rangle) = 2 \left(1 - \frac{1}{n} \sum_{k=1}^n \text{Tr} \rho_k^2 \right)$$