Radboud Universiteit

NWO

Complexity, machine learning and evolution

Mikhail Katsnelson

Outline

(1) Multiscale structural complexity of natural and unnatural patterns

(2) Statistical mechanics of machine learning: various aspects

(3) Complexity of quantum spin systems

(4) Machine learning and biological evolution

Main collaborators

Andrey Bagrov, Tom Westerhout, Askar Iliasov, Anna Kravchenko, Ilya Schurov, Andrea Russo
Radboud University, Nijmegen

Vladimir Mazurenko, Ilia Iakovlev, Oleg Sotnikov
Ural Federal University, Ekaterinburg

Vitaly Vanchurin
University of Minnesota

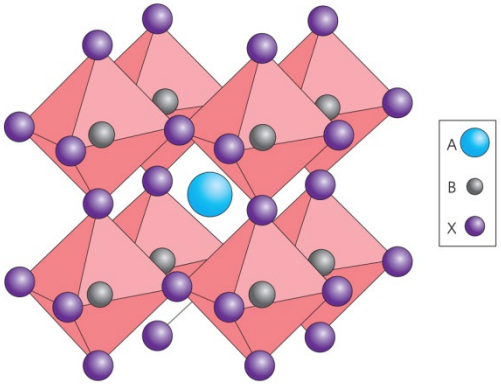
Eugene Koonin, Yuri Wolf
National Center for Biotechnology Information, Bethesda

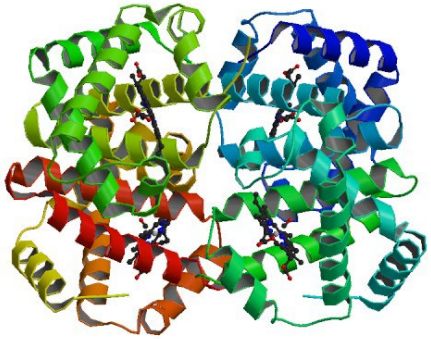
Veronica Dudarev
University of British Columbia, Vancouver

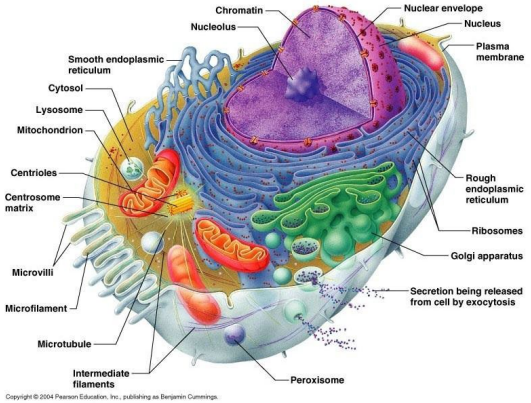
Complexity

Schrödinger: life substance is "aperiodic crystal" (modern formulation – Laughlin, Pines and others – glass)

Intuitive feeling: crystals are simple, biological structures are complex

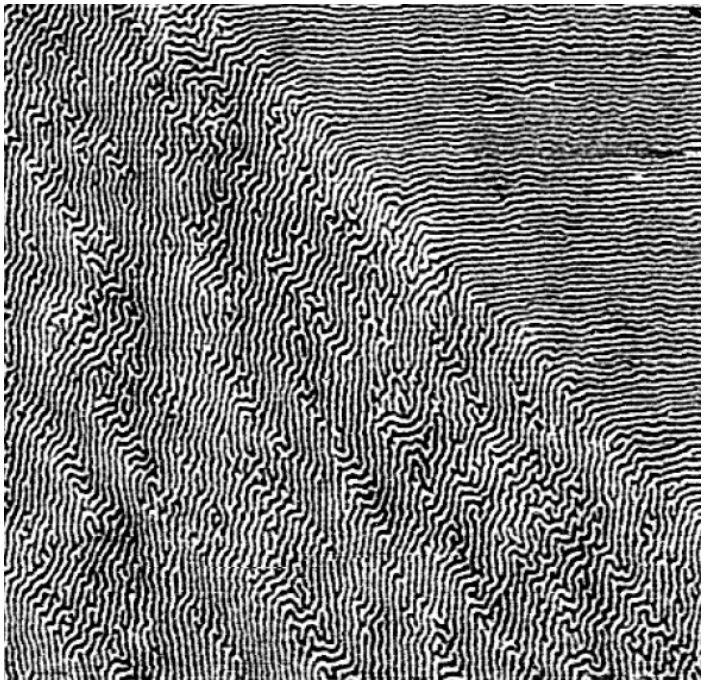






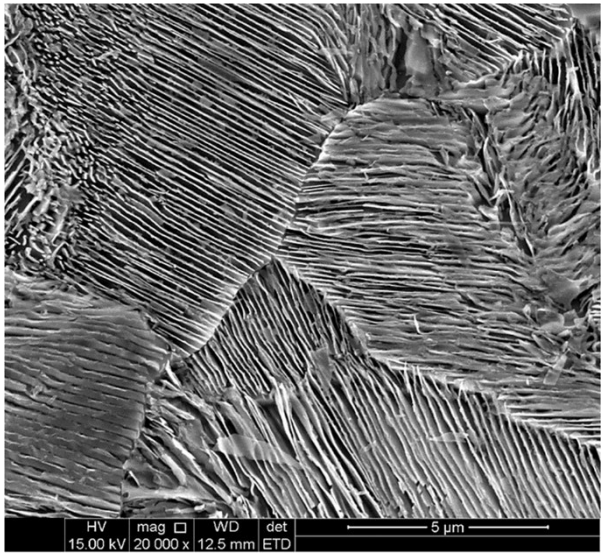
Origin and evolution of life: origin of complexity?

Complexity ("patterns") in inorganic world



Stripe domains in ferromagnetic thin films

Stripes on a beach in tide zone



Pearlitic structure in rail steel (Sci Rep 9, 7454 (2019))

Microstructures in metals and alloys

Do we understand this? No, or, at least, not completely

What is complexity?

- Something that we immediately recognize when we see it, but very hard to define quantitatively

- S. Lloyd, “Measures of complexity: a non-exhaustive list” – 40 different definitions

- Can be roughly divided into two categories:

- computational/descriptive complexities (“ultraviolet”)

- effective/physical complexities (“infrared” or inter-scale)

Computational and descriptive complexities

- Prototype – the Kolmogorov complexity:

the length of the shortest description (in a given language) of the object of interest

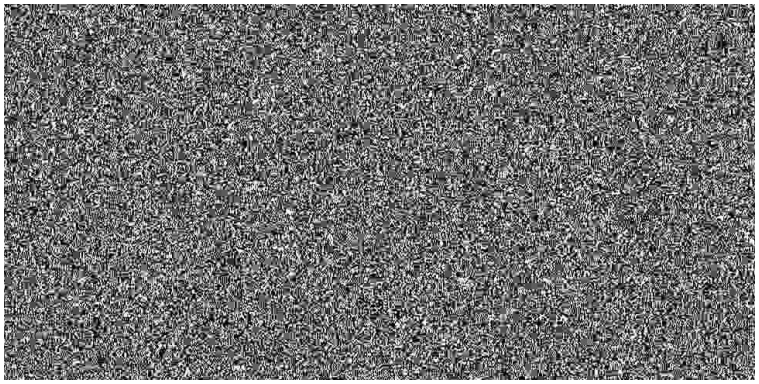
- Examples:

- Number of gates (in a predetermined basis) needed to create a given state from a reference one

- Length of an instruction required by file compressing program to restore image

Descriptive complexity

- The more random – the more complex:



White noise
970 x 485 pixels, gray scale, 253 Kb

>

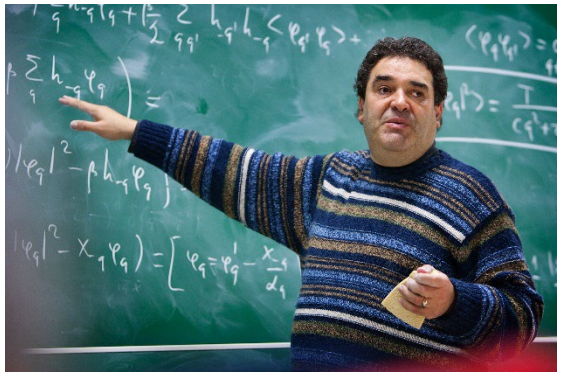
Vermeer "View of Delft"
750 x 624 pixels, colored, 234 Kb

Descriptive complexity II

The longer instruction – the more complex?

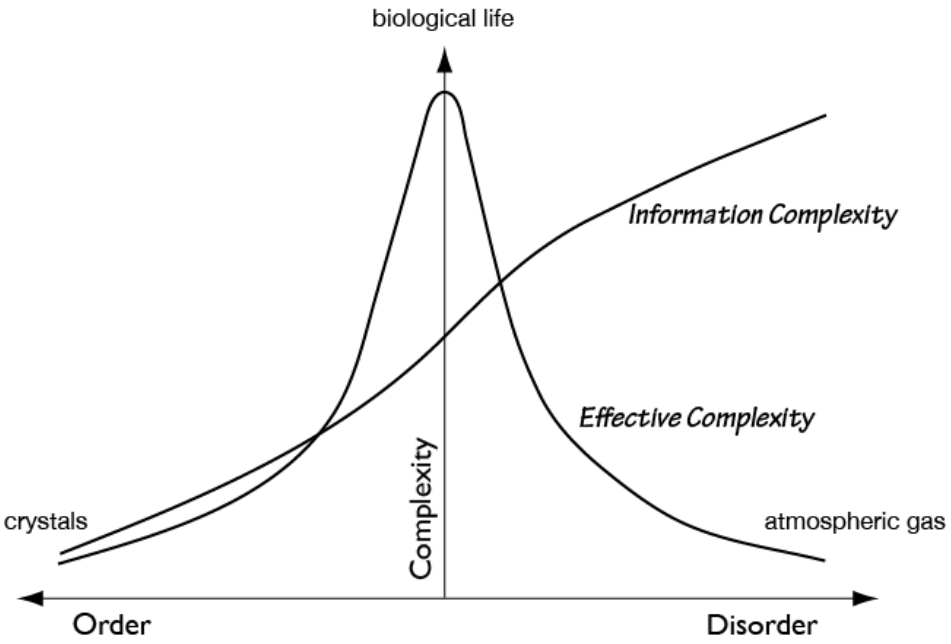
Paris japonica - 150 billion base pairs in DNA

$\gg$



Homo sapiens - 3.1 billion base pairs in DNA

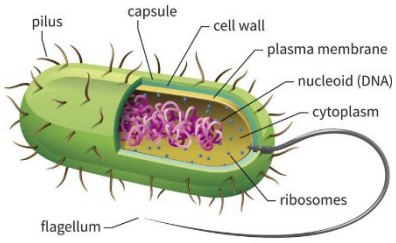
Effective complexity



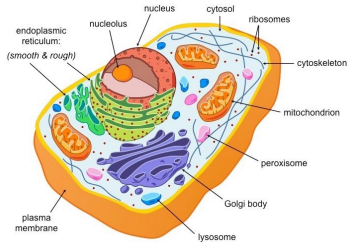
Can we come up with a quantitative measure?..

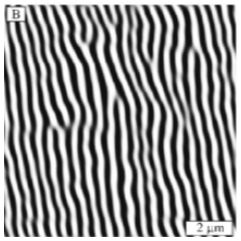
Not a mere philosophical question

- What happens at the major evolutionary transitions?
- Why are simple neural algorithms capable of solving complex many-body problems?
- Why do many natural patterns appear to be universal?

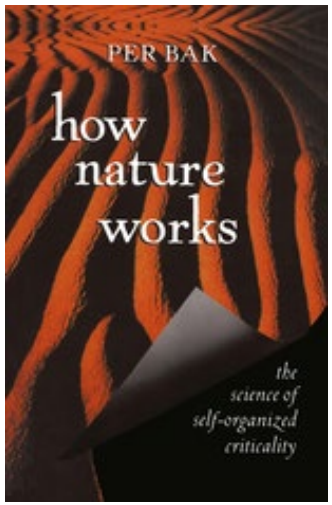


→





Attempts: Self-Organized Criticality?



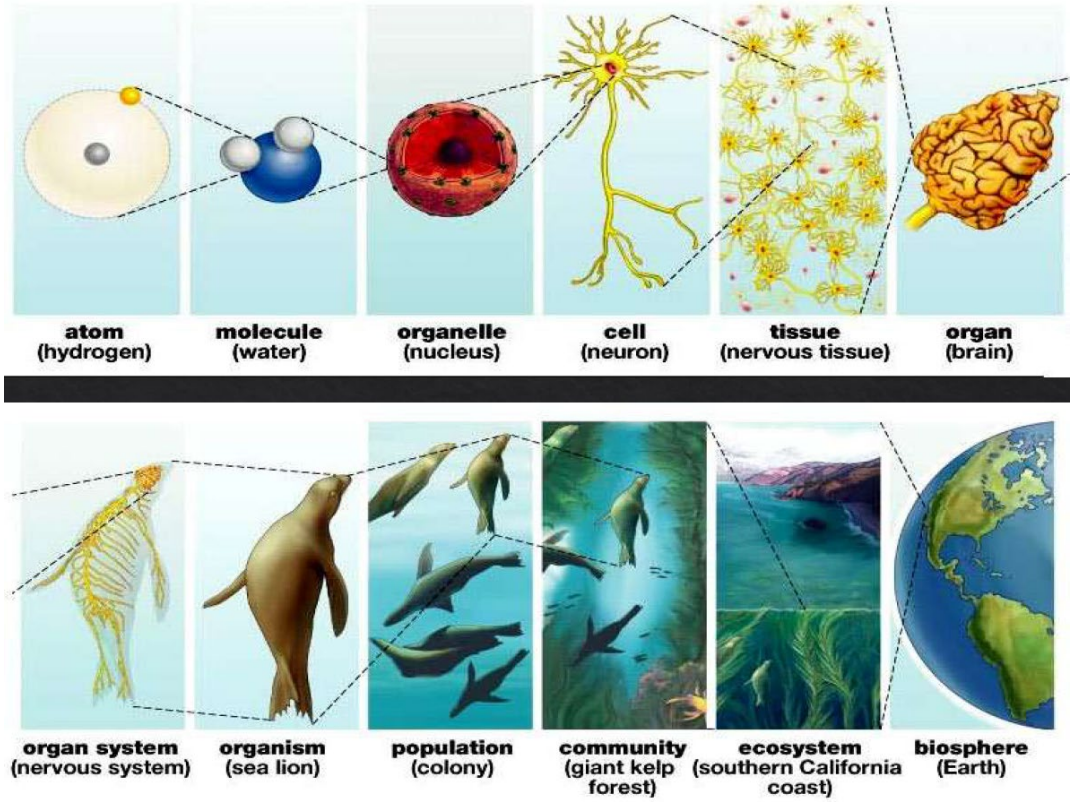
Per Bak: Complexity *is* criticality

Some complicated (marginally stable) systems demonstrate self-similarity and "fractal" structure

This is intuitively more complex behavior than just white noise but can we call it "complexity"?

But: complexity is hierarchical!

Our idea is dissimilarity at different scales



Multiscale structural complexity

Multi-scale structural complexity of natural patterns

PNAS 117, 30241 (2020)

Andrey A. Bagrov[a,b,1,2], Ilia A. Iakovlev[b,1], Askar A. Iliasov[c], Mikhail I. Katsnelson[c,b], and Vladimir V. Mazurenko[b]

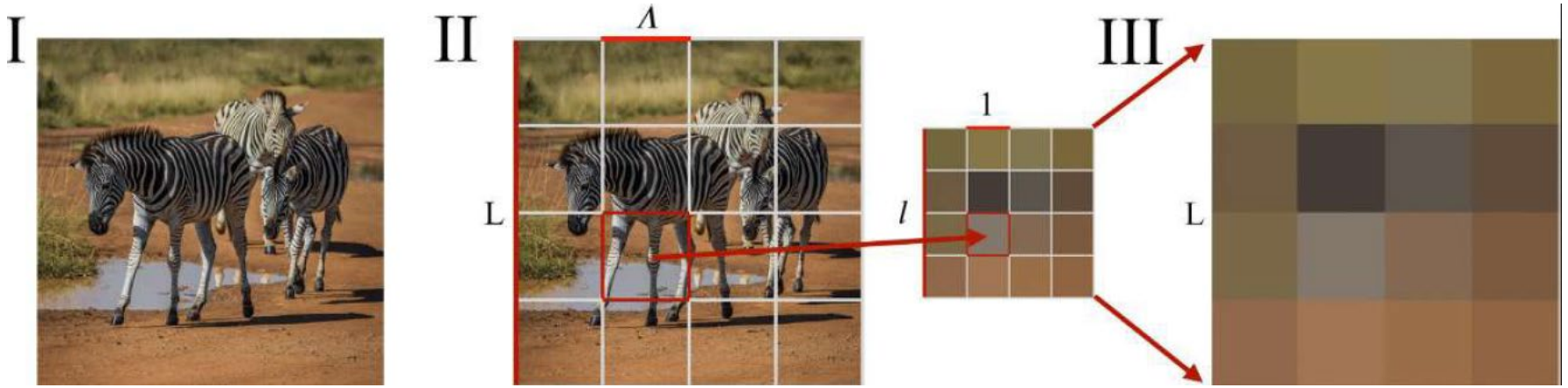
The idea (from holographic complexity and common sense): Complexity is dissimilarity at various scales

Let $f(x)$ be a multidimensional pattern

$f_\Lambda(x)$ its coarse-grained version (Kadanoff decimation, convolution with Gaussian window functions,…)

Complexity is related to distances between $f_\Lambda(x)$ and $f_{\Lambda+d\Lambda}(x)$

Structural complexity II



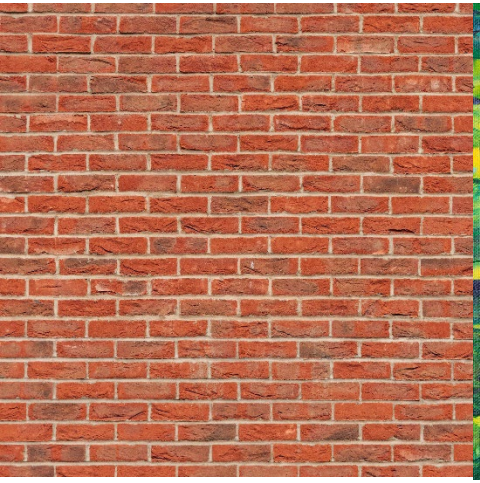
$$\Delta_\Lambda = |\langle f_\Lambda(x)|f_{\Lambda+d\Lambda}(x)\rangle - \frac{1}{2}\left(\langle f_\Lambda(x)|f_\Lambda(x)\rangle + \langle f_{\Lambda+d\Lambda}(x)|f_{\Lambda+d\Lambda}(x)\rangle\right)| = \frac{1}{2}|\langle f_{\Lambda+d\Lambda}(x) - f_\Lambda(x)|f_{\Lambda+d\Lambda}(x) - f_\Lambda(x)\rangle|,$$

$$\langle f(x)|g(x)\rangle = \int_D dx f(x)g(x)$$

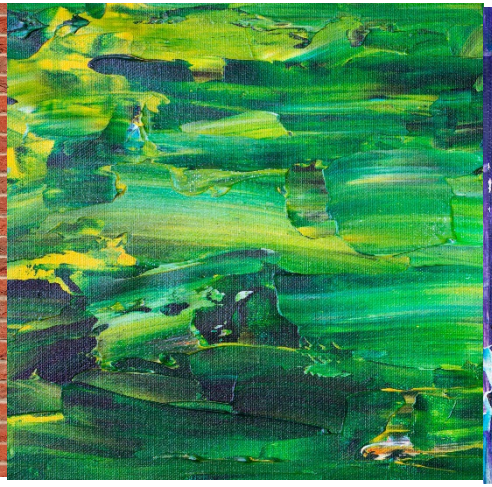
$$C = \sum_\Lambda \frac{1}{d\Lambda}\Delta_\Lambda \to \int |\langle \frac{\partial f}{d\Lambda}|\frac{\partial f}{d\Lambda}\rangle| d\Lambda, \text{ as } d\Lambda \to 0$$

Different ways of coarse-graining: average, "winner takes all" (Kadanoff decimation), cut-off in reciprocal space for Fourier image (Wilson RG…)

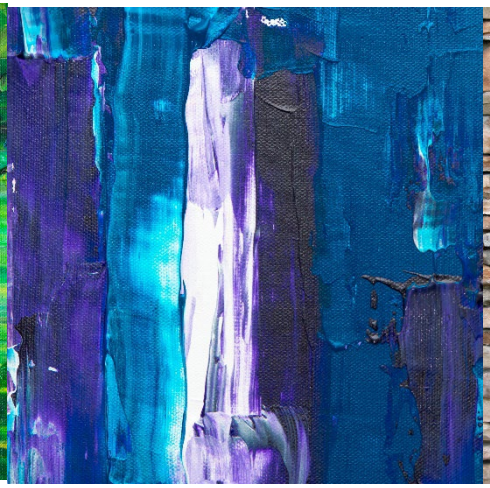
Art objects (and walls)



$C = 0.1076$



$C = 0.2010$

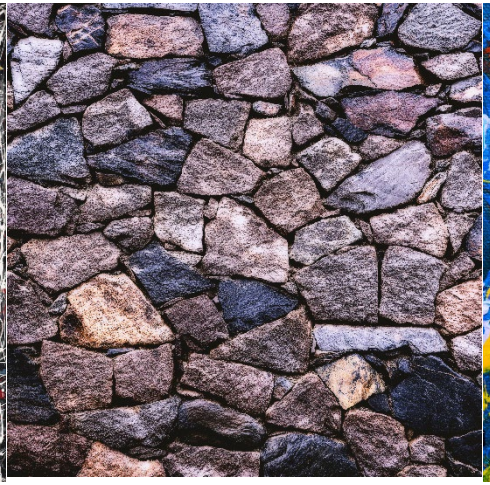


$C = 0.2147$

$C = 0.2765$

$C = 0.4557$

$C = 0.4581$



$C = 0.4975$

$C = 0.5552$

Other objects

$C = 0.353$

$C = 0.152$

$C = 0.204$

$C = 0.260$

$C = 0.167$

$C = 0.316$

$C = 0.209$

Photos by V. V. Mazurenko

Solution of an ink drop in water

Entropy should grow, but complexity is not! And indeed…

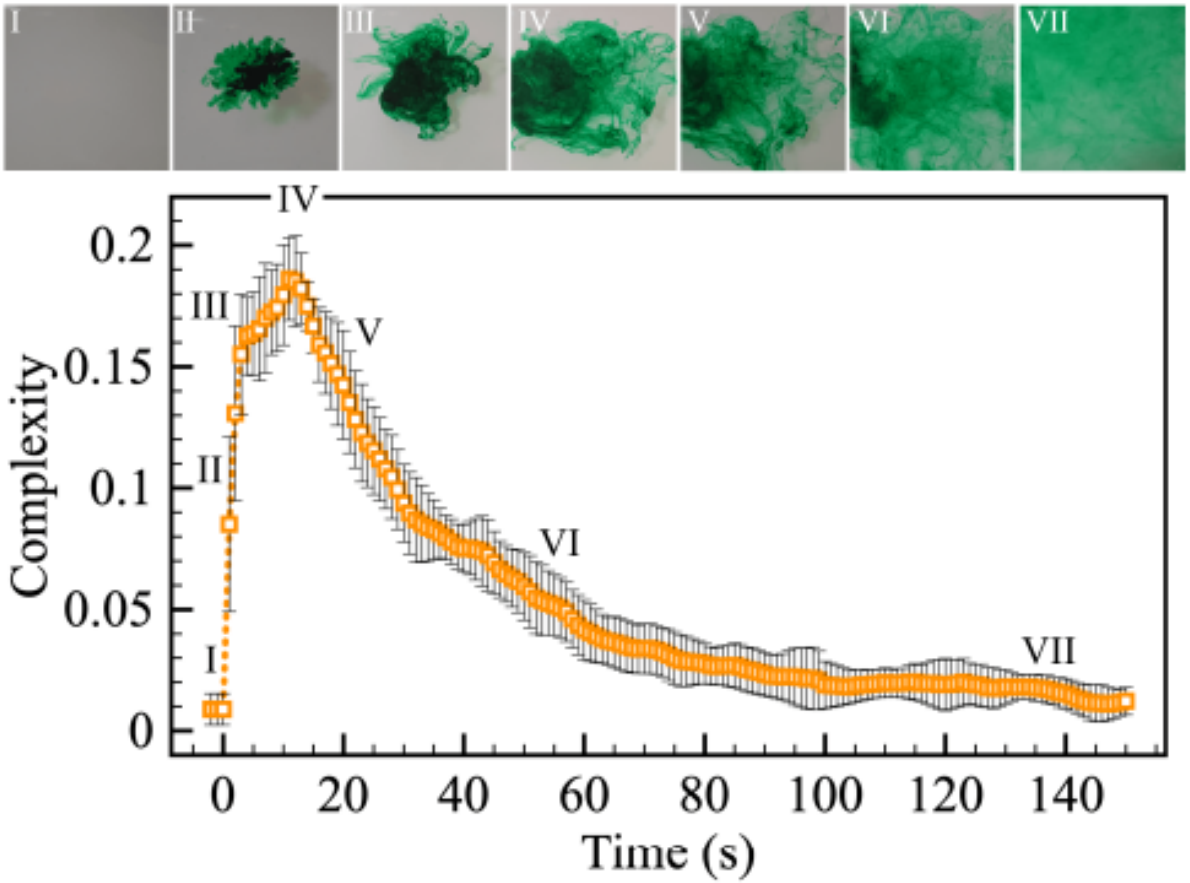
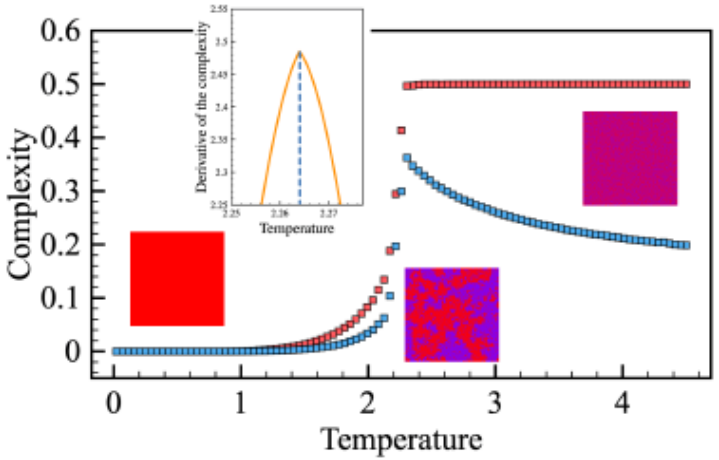


FIG. 7. The evolution of the complexity during the process of dissolving a food dye drop of 0.3 ml in water at 31°C.

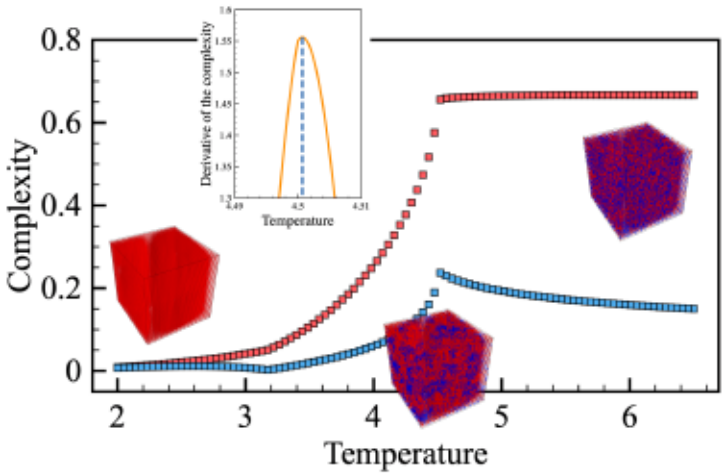
Structural complexity: Ising model

Can be used as a numerical tool to find T_C from finite-size simulations

2D



3D



Different ways of coarse-graining give different pictures but T_c is always a cusp

Structural complexity: Magnetic patterns II

Simulations of magnetic systems $H = -J\sum_{nn'} \mathbf{S}_n\mathbf{S}_{n'} - \mathbf{D}\sum_{nn'}[\mathbf{S}_n \times \mathbf{S}_{n'}] - \sum_n BS_n^z$

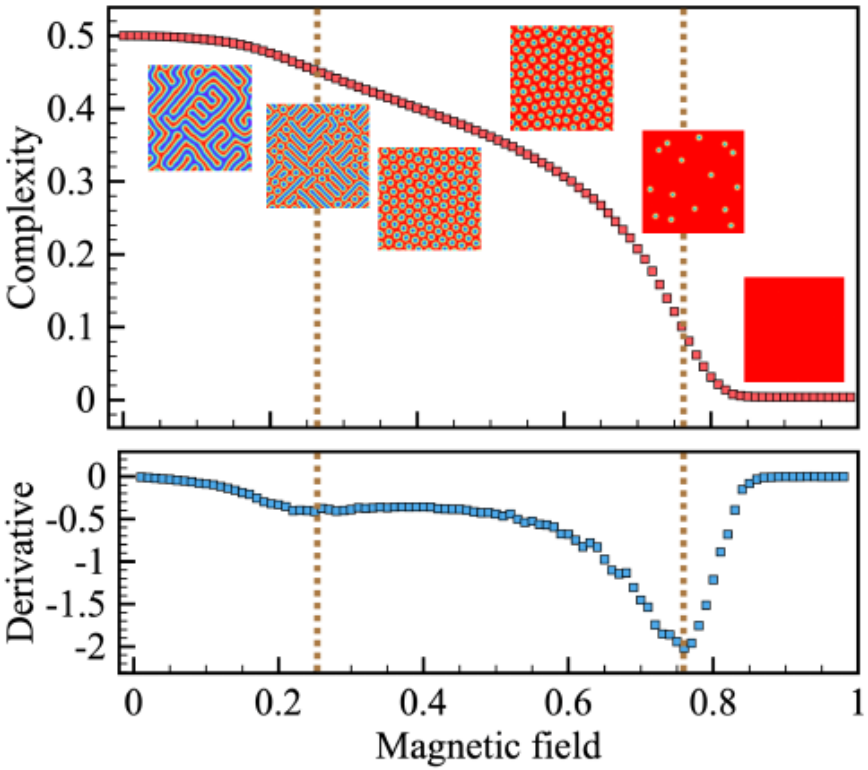


FIG. 4. (a) Magnetic field dependence of the complexity obtained from the simulations with spin Hamiltonian containing DM interaction with $J = 1$, $|\mathbf{D}| = 1$, $T = 0.02$. The error bars are smaller than the symbol size. (b) Complexity derivative we used for accurate detection of the phases boundaries.

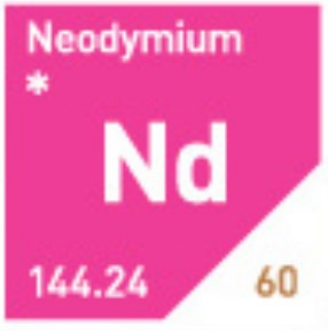
Experimental observation of self-induced spin glass state: elemental Nd

Self-induced spin glass state in elemental and crystalline neodymium

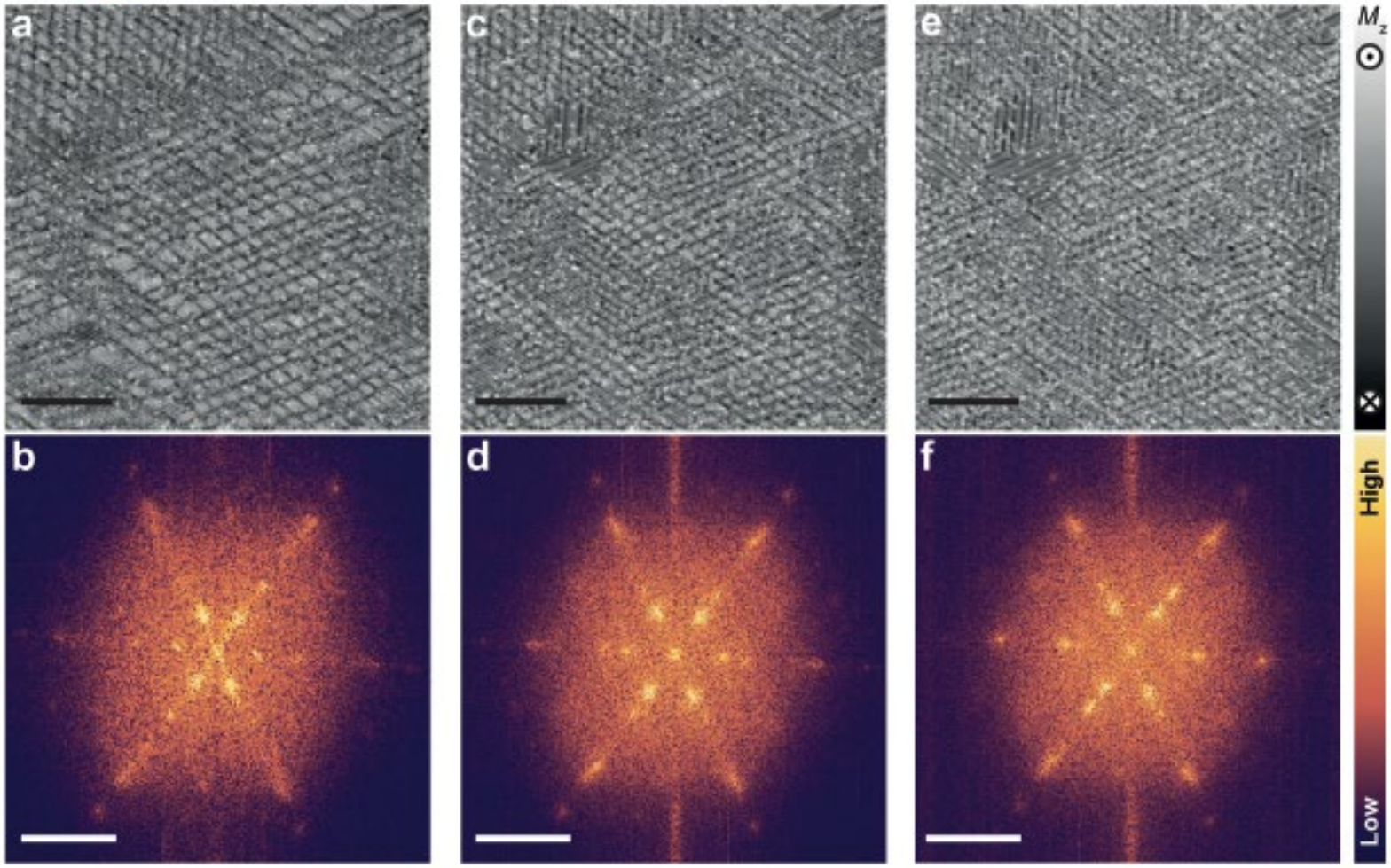
Science **368**, 966 (2020)

Umut Kamber, Anders Bergman, Andreas Eich, Diana Iuşan, Manuel Steinbrecher, Nadine Hauptmann, Lars Nordström, Mikhail I. Katsnelson, Daniel Wegner*, Olle Eriksson, Alexander A. Khajetoorians*

Spin-polarized STM experiment, Radboud University



Magnetic structure: local correlations



The most important observation: aging. At thermocycling (or cyling magnetic field) the magnetic state is not exactly reproduced

Order from disorder

Thermally induced magnetic order from glassiness in elemental neodymium

NATURE PHYSICS | VOL 18 | AUGUST 2022 | 905–911

Benjamin Verlhac[1], Lorena Niggli[1], Anders Bergman[2], Umut Kamber[1], Andrey Bagrov[1,2], Diana Iușan[2], Lars Nordström[2], Mikhail I. Katsnelson[1], Daniel Wegner[1], Olle Eriksson[2,3] and Alexander A. Khajetoorians[1]✉

Glassy state at low T and long-range order at T increase

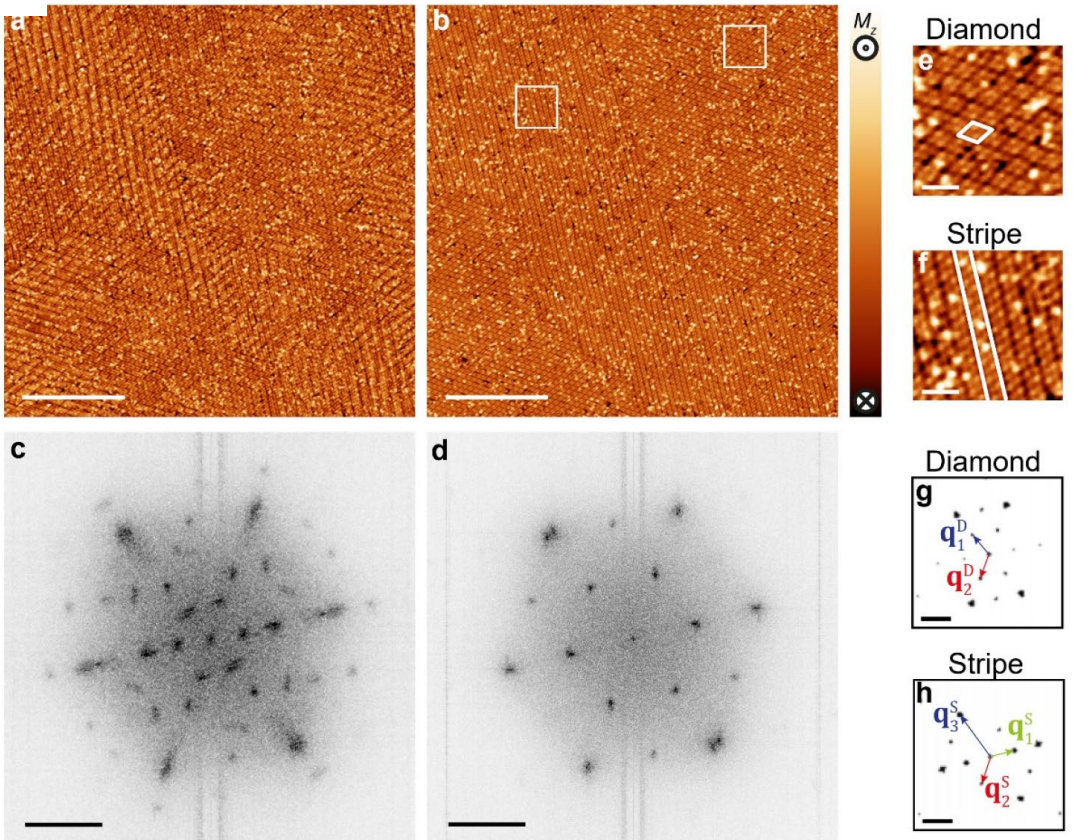
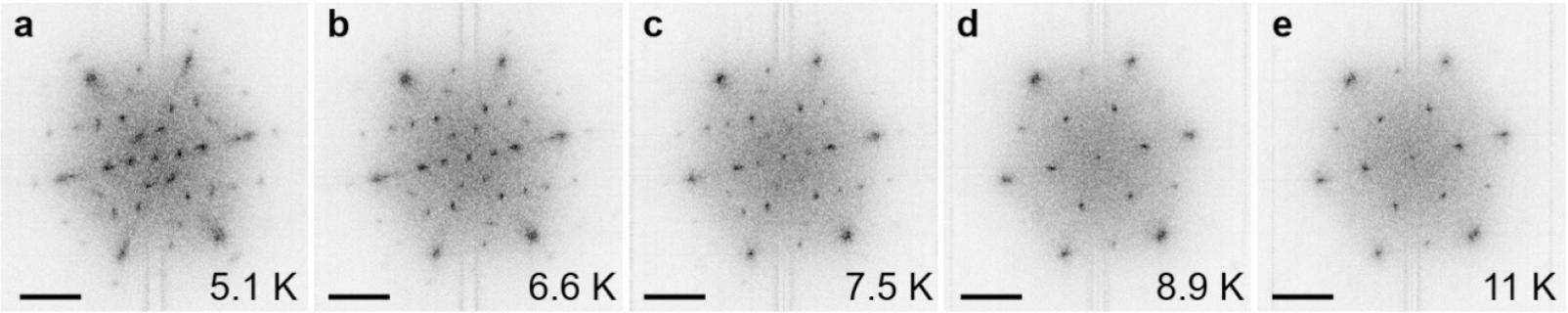


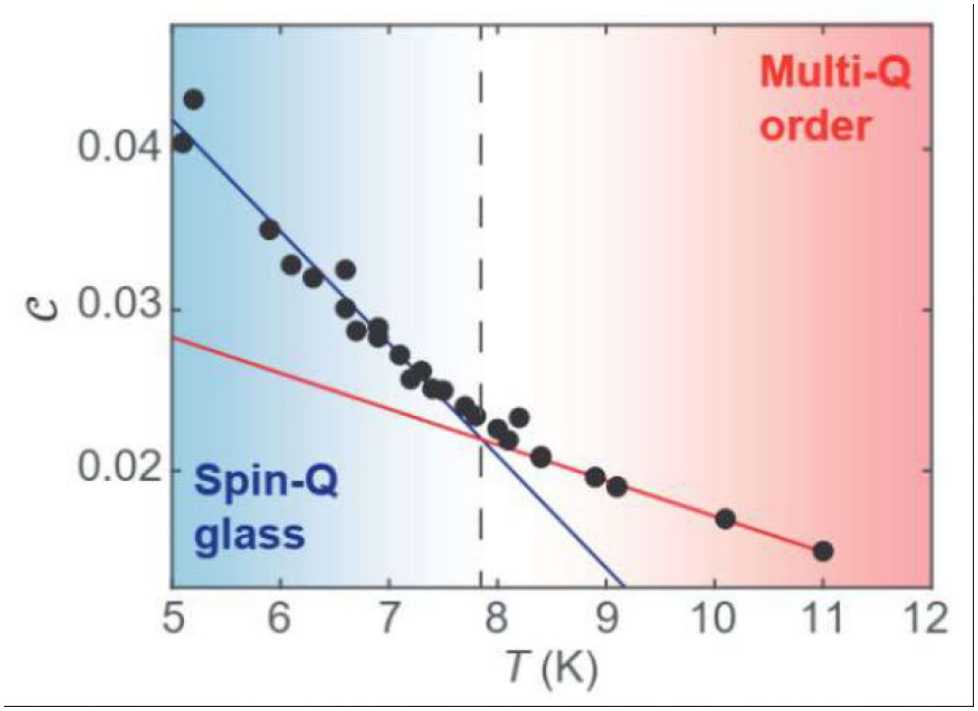
Figure 2: Emergence of long-range multi-Q order from the spin-Q glass state at elevated temperature. a,b. Magnetization images of the same region at T = 5.1 K and 11 K, respectively (I_t = 100 pA, a-b, scale bar: 50 nm). c,d. Corresponding Q-space images (scale bars: 3 nm^{-1}), illustrating the changes from strong local (i.e. lack of long-range) Q order toward multiple large-scale domains with well-defined long-range multi-Q order. e,f. Zoom-in images of the diamond-like (e) and stripe-like (f) patterns (scale bar: 5 nm). The locations of these images is shown by the white squares in b. g,h. Display of multi-Q state maps of the two apparent domains in the multi-Q ordered phase, where (g)

T=5K (a,c): spin glass

T=11K(b,d): (noncollinear) AFM

Order from disorder II





Phase transition at approx. 8K (seen via our complexity measure)

Certification of quantum states

Certification of quantum states with hidden structure of their bitstrings

npj Quantum Information (2022)8:41

O. M. Sotnikov[1], I. A. Iakovlev[1], A. A. Iliasov[2], M. I. Katsnelson[2], A. A. Bagrov[2,3] and V. V. Mazurenko[1]✉

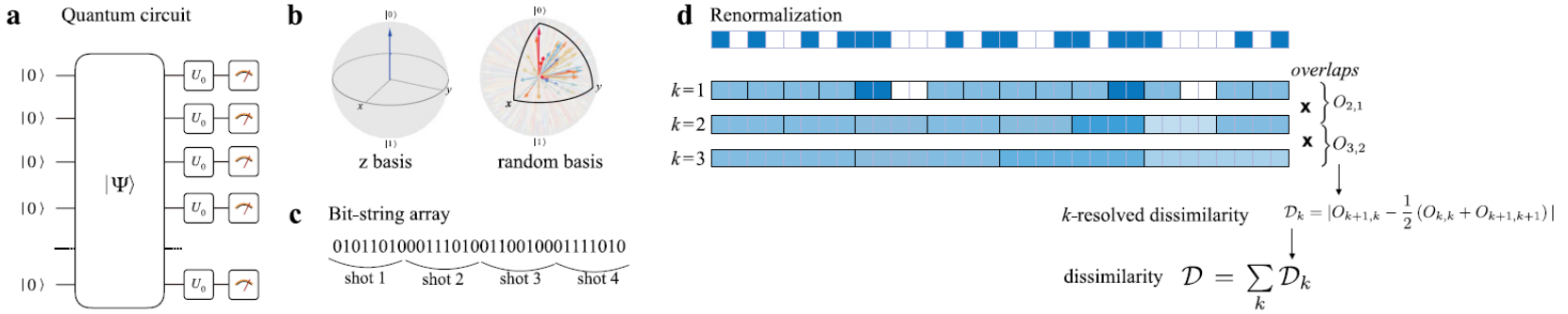


Fig. 1 Protocol for computing dissimilarity of a quantum state. a First, one prepares a state on a quantum device and chooses the measurement basis by applying rotational gates U_0 to individual qubits. **b** In this paper, we work with σ^z and random bases whose Bloch sphere representations are shown in the picture. We say that the set of measurements is performed in a random basis if, for each shot of measurement, a random vector belonging to the highlighted sector of the Bloch sphere is uniformly sampled and the corresponding parameters of gate U_0 are applied. **c** A number of measurements is performed and their outcomes — bitstrings of length N — are then stacked together in a one-dimensional binary array of length $N \times N_{\text{shots}}$ that serves as a classical representation of the quantum state. **d** The array is coarse-grained in several steps (indexed with k). Different schemes can be employed, but here we use plain averaging with fixed filter size Λ. In the picture, blue and white squares in the top line correspond to '0' and '1' bits in the array shown in (**c**), and black rectangles depict the blocks where averaging occurs at every step of coarse-graining. Overlap-based dissimilarities $\mathcal{D}_k$ between subsequent arrays are computed and summed up to the overall dissimilarity $\mathcal{D}$. See Methods section for more details.

First make at least two complementary measurements, then analyze the measurement results (relation to Bohr's complementarity principle)

Certification of quantum states II

Two-dimensional map can be used to characterize the type of quantum states

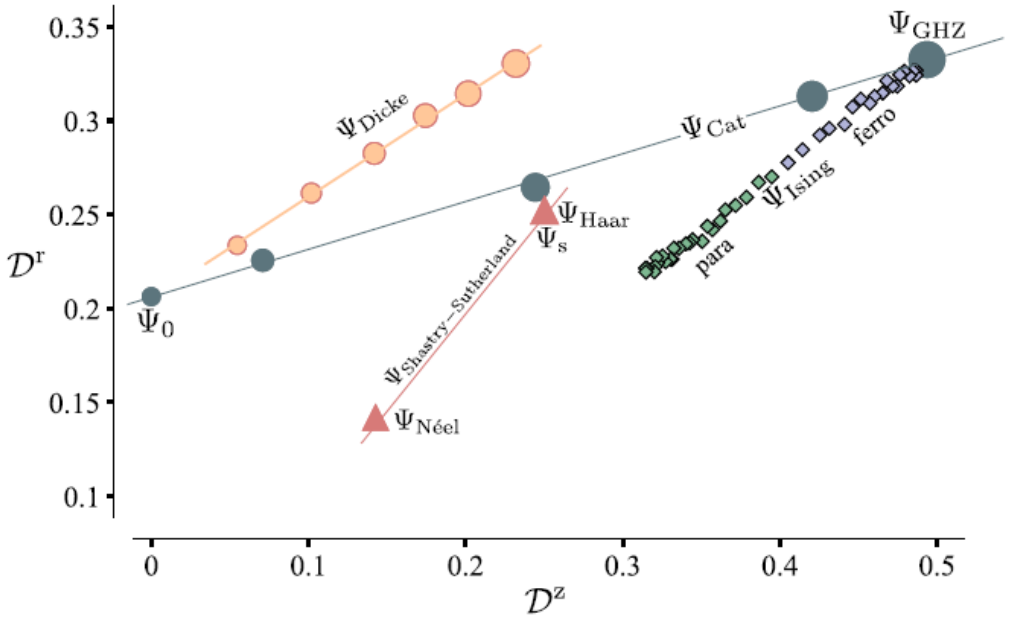


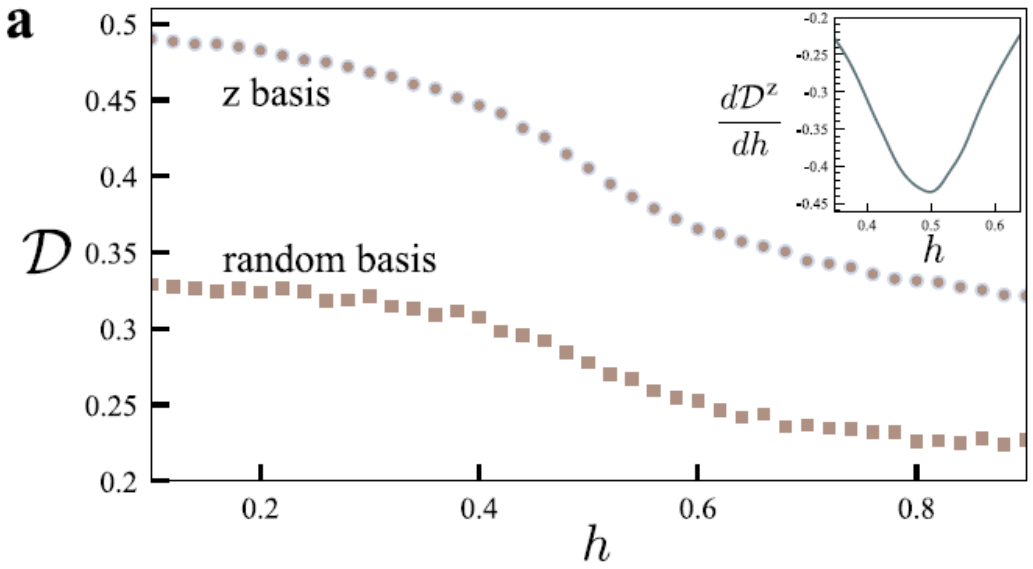
Fig. 10 Dissimilarity map. Low-dimensional representation of the 16-qubit quantum states studied in this work with respect to their dissimilarity calculated in σ^z and random bases. Ψ_0, Ψ_s, Ψ_{Haar} denote the trivial $|0\rangle^{\otimes N}$, the singlet and the random quantum states, respectively.

Certification of quantum states III

One can characterize a type of quantum states and, again, find (quantum) critical point

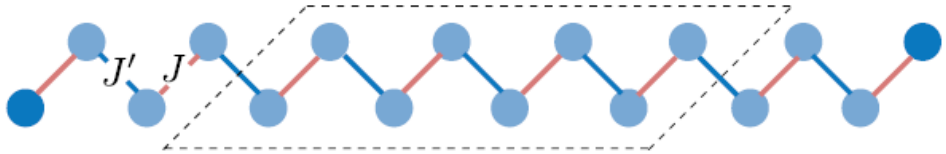
Ising in transverse field $H = J\sum_{ij} \hat{S}_i^z \hat{S}_j^z + h\sum_i \hat{S}_i^x$

1D chain; quantum critical point at $h_c = 0.5|J|$



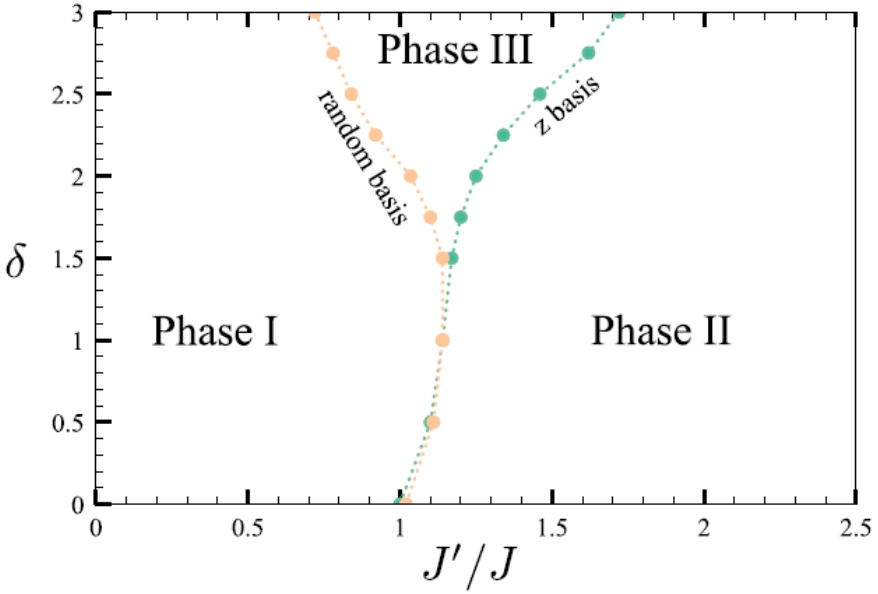
Certification of quantum states IV

The way to detect topological phases

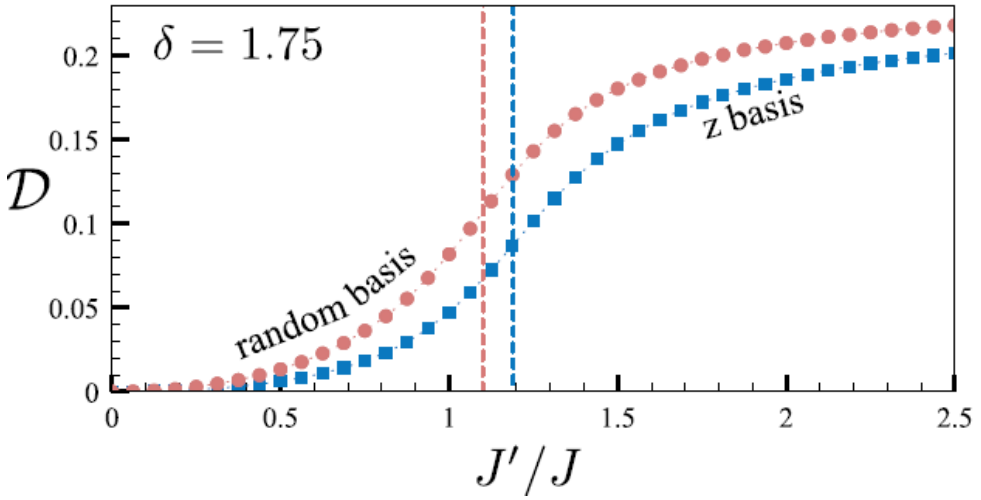


$$H_{XXZ} = J \sum_{ij \in \mathbf{O}} (\hat{S}_i^x \hat{S}_j^x + \hat{S}_i^y \hat{S}_j^y + \delta \hat{S}_i^z \hat{S}_j^z)$$
$$+ J' \sum_{ij \in \mathbf{E}} (\hat{S}_i^x \hat{S}_j^x + \hat{S}_i^y \hat{S}_j^y + \delta \hat{S}_i^z \hat{S}_j^z)$$

Three different phases: trivial (I), topological (II), antiferromagnetic (III)



Phase diagram constructing from dissimilarity (MSC), from maxima of its derivatives



Dissimilarities with different basis show different phase boundaries

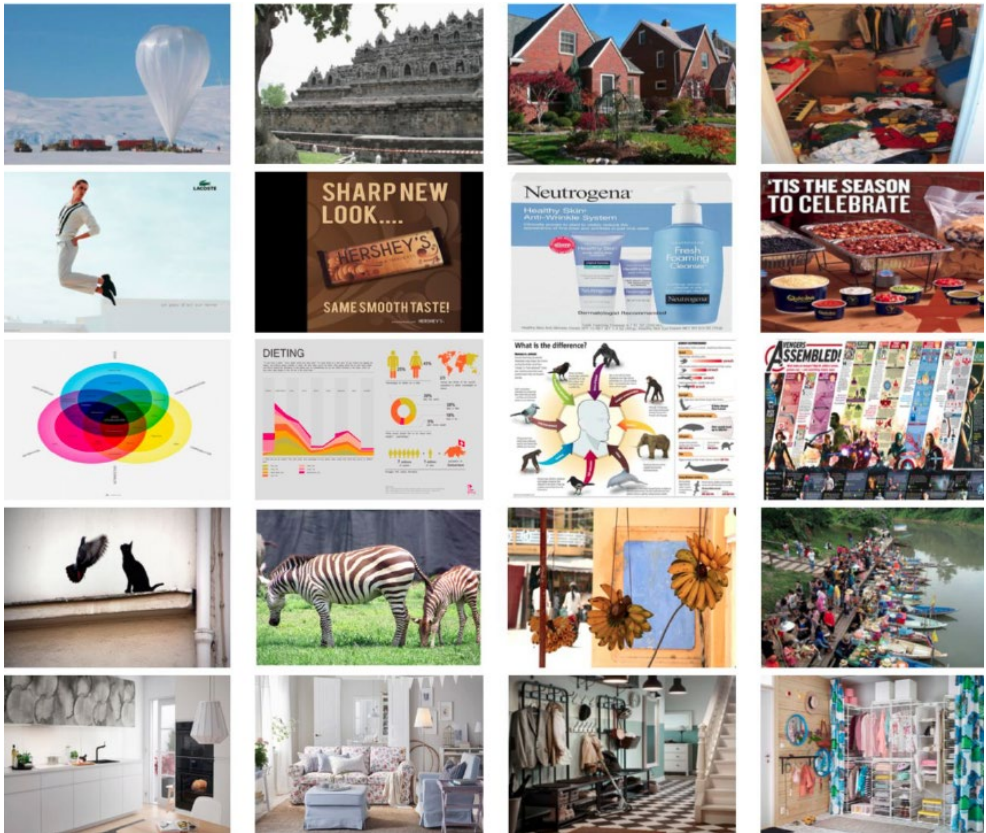
Psychology of human visual perception

We wanted definition of complexity in agreement with our intuitive understanding of complexity – did we succeed?

Multiscale structural dissimilarity in human perception of visual complexity
Anna Kravchenko, Andrey Bagrov, MIK, Veronica Dudarev (in preparation)

To analyze: **SAVOIAS: A Diverse, Multi-Category Visual Complexity Dataset**
Elham Saraee, Mona Jalal, Margrit Betke; arXiv:1810.01771

- Multiple domains: Scenes, Advertisements, Infographics, Objects, Interior design, Art, and Suprematism
- Well-studied for other existing complexity measures
- Obtained by crowdsourcing more than 37,000 pairwise comparisons of images, rankings converted into 1-100 scale



Human visual perception II

Choosing coarse-graining method

There are many ways to do coarse graining

Evidence suggests processing on early layers of visual cortex can be approximated by Fourier Transform (Campbell and Robson 1968, Ochs 1979, Kulikowski and Bishop 1981, Olshausen 2003, Kesserwani 2002)

DFT:

$$F(k,l) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} f(i,j)\, e^{-\iota 2\pi \left(\frac{ki}{N} + \frac{lj}{N}\right)}$$

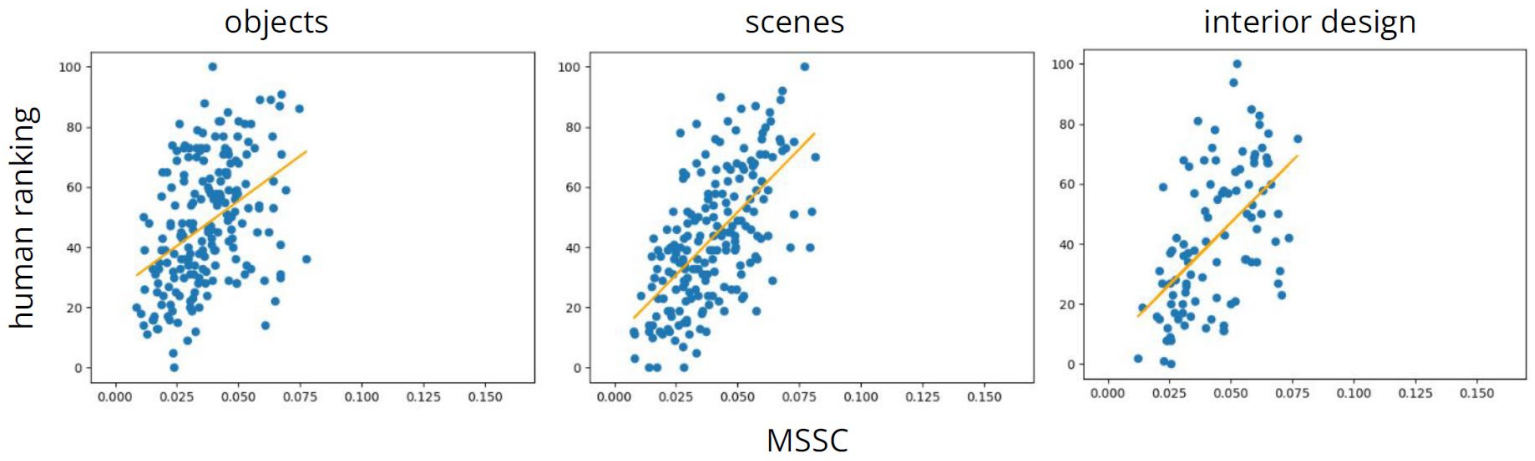
Inverse Fourier transform:

$$f(a,b) = \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} F(k,l)\, e^{\iota 2\pi \left(\frac{ka}{N} + \frac{lb}{N}\right)}$$

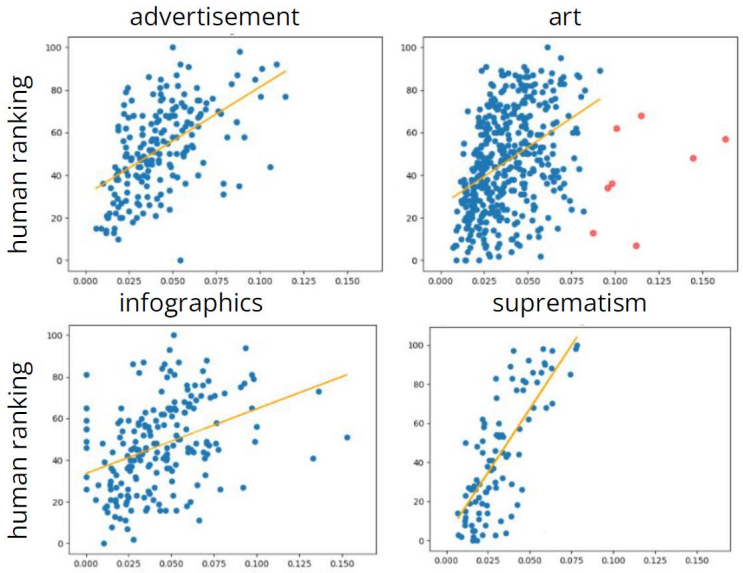
The best correlation is reached when making coarse-graining of Fourier images which can be interesting by itself

Human visual perception III

Natural scenes



Man-made images



- Worse correlation
- Obvious outliers

What does this imply?

Human visual perception IV

Comparison with existing measures

		edge density	compression ratio	number of regions	feature congestion	subband entropy	MSSC	max r
natural	scenes	0.16	0.3	0.57	0.42	0.16	0.62	0.57
	objects	0.28	0.16	0.29	0.3	0.1	0.46	0.3
Human -made	suprematism	0.18	0.6	0.84	0.48	0.39	0.76	0.84
	interior design	0.61	0.68	0.67	0.58	0.31	0.60	0.68
	advertisements	0.54	0.56	0.41	0.56	0.54	0.52	0.56
symbolic	art	0.48	0.51	0.65	0.22	0.33	0.42	0.65
	infographics	0.57	0.55	0.38	0.52	0.61	0.38	0.61

From (Saree et al. 2018)

Surpasses state-of-the-art on natural scenes, falls below state-of-the-art methods for images conveying information

What is missing in our definition? Contextual complexity (that is, cultural references)

Human visual perception V

Perceptual vs. conceptual complexity: art as a language

Outliers:

Visually straining but conceptually boring

Paintings identified as outliers, having been excluded from prediction by a threshold greater than two standard deviations, exhibited the same distinctive visual characteristic: broad lines featuring contrasting colors (9). This artistic style can be visually straining, however, it's relatively simplistic in terms of information conveyed through it.

Other applications –biology

Nucleic Acids Research, 2024, 52, 11045–1105

Long range segmentation of prokaryotic genomes by gene age and functionality

Yuri I. Wolf[1], Ilya V. Schurov[2], Kira S. Makarova[1], Mikhail I. Katsnelson[2], Eugene V. Koonin[1]

Multilevel structural complexity was used to analyze observed patterns in prokaryotic genomes vs predictions of various models

To summarize the first part: computationally simple but useful concept

Statistical mechanics of machine learning

Based on the formalism

Vanchurin, V. The world as a neural network. Entropy **22**, 1210 (2020)

Vanchurin, V. Towards a theory of machine learning. Mach. Learn.: Sci. Technol. **2**, 035012 (2021)

Applications to be discussed here:

M. I. Katsnelson and V. Vanchurin, Emergent quantumness in neural networks
FOUND PHYS 51, 94 (2021)

From Boltzmann's time: problem of derivation of irreversible equations of statistical thermodynamics from reversible microscopic equations of classical (or quantum) physics

We invert it: initially, equations of machine learning are irreversible but "thermodynamic" growth of entropy can be compensate due tp entropy decrease via learning

Entropy is a measure of our ignorance of the state of the system: if we decrease our ignorance we decrease entropy

Statistical mechanics of machine learning II

Consider a learning system described by a coupled dynamics of *trainable* variables, $\mathbf{q}$, and non-trainable or *hidden* variables, $\mathbf{x}$. In "epistomological" kind of approaches [6, 14–16] one can identify the trainable variables with characteristics of a human mind whereas the hidden variables represent an external world, but this identification is not needed for our formal consideration which we will try to keep as general as possible. In context of artificial neural networks the trainable variables determine the weight matrix and bias vector, and the hidden variables represent the state vector of neurons [13]. It is assumed that on the shortest time-scales the dynamics of the trainable variables undergoes diffusion

$$\begin{aligned}\frac{\partial p(t,\mathbf{q})}{\partial t} &= \sum_k \frac{\partial}{\partial q_k}\left(D\frac{\partial p(t,\mathbf{q})}{\partial q_k} - \frac{dq_k}{dt}p(t,\mathbf{q})\right) \\ &= \sum_k \frac{\partial}{\partial q_k}\left(D\frac{\partial p(t,\mathbf{q})}{\partial q_k} - \gamma\frac{\partial F(t,\mathbf{q})}{\partial q_k}p(t,\mathbf{q})\right)\end{aligned} \tag{1}$$

and the dynamics of hidden variables is only described through its free energy

$$\begin{aligned}\frac{d}{dt}F(t,\mathbf{q}) &= \frac{\partial F(t,\mathbf{q})}{\partial t} + \sum_k \frac{dq_k}{dt}\frac{\partial F(t,\mathbf{q})}{\partial q_k} \\ &= \frac{\partial F(t,\mathbf{q})}{\partial t} + \gamma\sum_k\left(\frac{\partial F(t,\mathbf{q})}{\partial q_k}\right)^2\end{aligned}$$

Learning dynamics: towards optimization of F

$$\frac{dq_k}{dt} = \gamma\frac{\partial F(t,\mathbf{q})}{\partial q_k}.$$

Entropy production

Shannon entropy $$S_q(t) = -\int d^K q \;\; p(t,\mathbf{q}) \log\left(p(t,\mathbf{q})\right)$$

$$\begin{aligned}
\frac{dS_q(t)}{dt} &= -\int d^K q\, p \frac{\partial \log(p)}{\partial t} - \int d^K q \;\log(p)\frac{\partial p}{\partial t} \\
&= -\frac{d}{dt}\int d^K q\, p - \int d^K q \;\log(p)\frac{\partial p}{\partial t} \\
&= -\int d^K q \;\log(p) \sum_k \frac{\partial}{\partial q_k}\left(D\frac{\partial p}{\partial q_k} - \gamma \frac{\partial F}{\partial q_k} p\right) \\
&= D\int d^K q \;\sum_k \frac{1}{p}\left(\frac{\partial p}{\partial q_k}\right)^2 - \gamma \int d^K q \;\sum_k \frac{\partial p}{\partial q_k}\frac{\partial F}{\partial q_k}.
\end{aligned}$$

The first term describes conventional entropy growth, the second is its decrease due to learning.

Near the learning equilibrium (that is, in well-trained systems)

$$\frac{dS_q(t)}{dt} \approx \int d^K q \;\sqrt{p}\left(-4D\sum_k \frac{\partial^2}{\partial q_k^2}\right)\sqrt{p}$$

Entropy production II

$$\mathcal{S}[p,F,\lambda] = \int_0^T dt \frac{dS_q}{dt} + \lambda \int_0^T dt d^K q\, p\left(\frac{\partial F}{\partial t} + \gamma \sum_k \left(\frac{\partial F}{\partial q_k}\right)^2 + \frac{V}{\epsilon}\right),$$

$$= \int_0^T dt\, d^K q \sqrt{p}\left(-4D \sum_k \frac{\partial^2}{\partial q_k^2} + \lambda \frac{\partial F}{\partial t} + \lambda\gamma \sum_k \left(\frac{\partial F}{\partial q_k}\right)^2 + \lambda \frac{V}{\epsilon}\right)\sqrt{p}, \tag{7}$$

where the total time-averaged free energy production pre unit time step ϵ is

$$V(\mathbf{q}) \equiv -\left\langle \epsilon \frac{d}{dt} F(t,\mathbf{q}) \right\rangle_t. \tag{8}$$

where λ is a Lagrange multiplier

This can be rewritten as

$$\mathcal{S}[p,F,\hbar] = \frac{\lambda}{\epsilon} \int_0^T dt\, d^K q \sqrt{p}\left(-\frac{\hbar^2}{2m} \sum_k \frac{\partial^2}{\partial q_k^2} + \frac{\partial(\epsilon F)}{\partial t} + \frac{1}{2m} \sum_k \left(\frac{\partial(\epsilon F)}{\partial q_k}\right)^2 + V\right)\sqrt{p} \tag{9}$$

where

$$m \equiv \frac{\epsilon}{2\gamma}, \tag{10}$$

and

$$\hbar \equiv \epsilon \sqrt{\frac{4D}{\gamma\lambda}}. \tag{11}$$

Further derivation is similar to that in “logical inference” approach

De Raedt, H., Katsnelson, M.I., Michielsen, K.: Quantum theory as the most robust description of reproducible experiments. Ann. Phys. **347**, 45–73 (2014)

Schrödinger equation

The main difference between (7) and (9) is that instead of solving the equations for p, F and λ, we are now solving them for p, F and $\hbar$. The optimal solutions are obtained by setting all possible variations of (9) to zero

$$\frac{\partial}{\partial \hbar}\mathcal{S}[p,F,\hbar] = -\int_0^T dt\, d^K q \frac{\hbar}{m}\sqrt{p}\sum_k \frac{\partial^2}{\partial q_k^2}\sqrt{p} = 0 \tag{12}$$

$$\frac{\delta}{\delta F}\mathcal{S}[p,F,\hbar] = -\frac{\partial}{\partial t}p - \frac{1}{m}\sum_k \frac{\partial}{\partial q_k}\left(\frac{\partial(\epsilon F)}{\partial q_k}p\right) = 0 \tag{13}$$

$$\frac{\delta}{\delta p}\mathcal{S}[p,F,\hbar] = -\frac{\hbar^2}{2m}\frac{1}{\sqrt{p}}\sum_k \frac{\partial^2\sqrt{p}}{\partial q_k^2} + \frac{\partial(\epsilon F)}{\partial t} + \frac{1}{2m}\sum_k\left(\frac{\partial(\epsilon F)}{\partial q_k}\right)^2 + V = 0. \tag{14}$$

$$\left\langle \frac{d}{dt}S_q \right\rangle_t = 0, \tag{15}$$

and by equations (13) and (14), which are the Madelung hydrodynamic equations [20]

$$\frac{\partial}{\partial t}p = -\sum_k \frac{\partial}{\partial q_k}(u_k p) \tag{16}$$

$$\frac{\partial}{\partial t}u_j = -\sum_k u_k \frac{\partial}{\partial q_k}u_j - \frac{1}{m}\frac{\partial}{\partial q_j}\left(V - \frac{\hbar^2}{2m}\sum_k \frac{\partial^2\sqrt{p}}{\partial q_k^2}\right) \tag{17}$$

with velocity of the fluid

$$u_k \equiv \frac{1}{m}\frac{\partial}{\partial q_k}(\epsilon F). \tag{18}$$

Schrödinger equation II

It is well known that the Madelung equations can be derived from the Schrödinger equation

$$-i\hbar\frac{\partial}{\partial t}\Psi = \left(\frac{\hbar^2}{2m}\sum_k \frac{\partial^2}{\partial q_k^2} - V\right)\Psi \tag{19}$$

where the wave function is defined as

$$\Psi \equiv \sqrt{p}\exp\left(\frac{iF\epsilon}{\hbar}\right). \tag{20}$$

The difference is that the phase lives on a circle, not on an infinite straight line.

Last step: neural networks with changeable number of neurons, grand canonical ensemble

$$F \cong F + \mu n \qquad \forall n \in \mathbb{Z}$$

$$\hbar = \pm\frac{\mu\epsilon}{2\pi}. \tag{26}$$

By imposing the condition (26) on $\hbar$ in (22) we arrive at the Schrödinger action

$$\mathcal{S}[\Psi] = \frac{\lambda}{\epsilon}\int_0^T dt\, d^K q \left(\frac{\hbar^2}{2m}\sum_k \frac{\partial\Psi^*}{\partial q_k}\frac{\partial\Psi}{\partial q_k} - i\hbar\Psi^*\frac{\partial\Psi}{\partial t} + V\Psi^*\Psi\right) \tag{27}$$

Schrödinger equation III

Learning process in well-trained network (near equilibrium) with changeable number of neurons is described by something which looks like quantum equation

Emergent quantumness – contrary to ideas of "quatum biology", "quantum consciousness" etc.

"Planck constant" is determined by characteristics of the network

Emergent criticality (scaling)

M. I. Katsnelson, V. Vanchurin, and T. Westerhout, Emergent scale invariance in neural networks PHYSICA A 610, 128401 (2023)

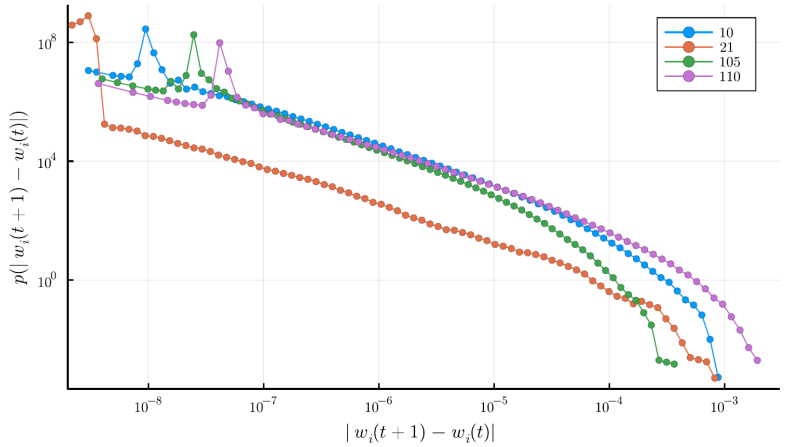
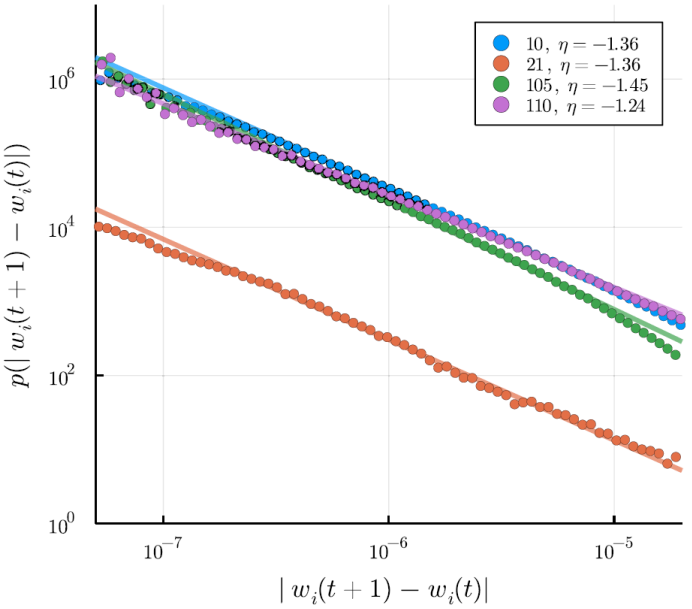


Fig. 2. Distribution of local fluctuations $w_i(t+1)-w_i(t)$ for a few randomly selected weights. We have checked that the distribution is symmetric and for clarity show only $|w_i(t+1)-w_i(t)|$. Labels 10, 21, 105, and 110 indicate different values of i, i.e. different weights. Weights 10 (blue) and 21 (orange) come from the first layer, and 105 (green) and 110 (purple) — from the second. For each i, more than 10^7 data points were used to calculate the distribution.



Analytical consideration connects power-law scaling to existence of "soft" directions: neurons which state can be changed quite essentially without the effect on loss function (cf "neutral evolution" in biology, see below)

Frustrations and complexity: Quantum case

NATURE COMMUNICATIONS | (2020)11:1593

Generalization properties of neural network approximations to frustrated magnet ground states

Tom Westerhout[1✉], Nikita Astrakhantsev[2,3,4✉], Konstantin S. Tikhonov[5,6,7✉], Mikhail I. Katsnelson[1,8] & Andrey A. Bagrov[1,8,9✉]

How to find true ground state of the quantum system?

In general, a very complicated problem (difficult to solve even for quantum computer!)

Idea: use of variational approach and train neural network to find "the best" trial function (G. Carleo and M. Troyer, Science 355, 602 (2017))

$$|\Psi_{GS}\rangle = \sum_{i=1}^{K} \psi_i |\mathcal{S}_i\rangle = \sum_{i=1}^{K} s_i |\psi_i| |\mathcal{S}_i\rangle$$

Generalization problem: to train NN for relatively small basis (*K* much smaller than total dim. of quantum space) and find good approximation to the true ground state

Frustrations and complexity: Quantum case II

Quantum S=1/2 Hamiltonian
NN and NNN interactions

$$\hat{H} = J_1 \sum_{\langle a,b \rangle} \hat{\boldsymbol{\sigma}}_a \otimes \hat{\boldsymbol{\sigma}}_b + J_2 \sum_{\langle\langle a,b \rangle\rangle} \hat{\boldsymbol{\sigma}}_a \otimes \hat{\boldsymbol{\sigma}}_b$$

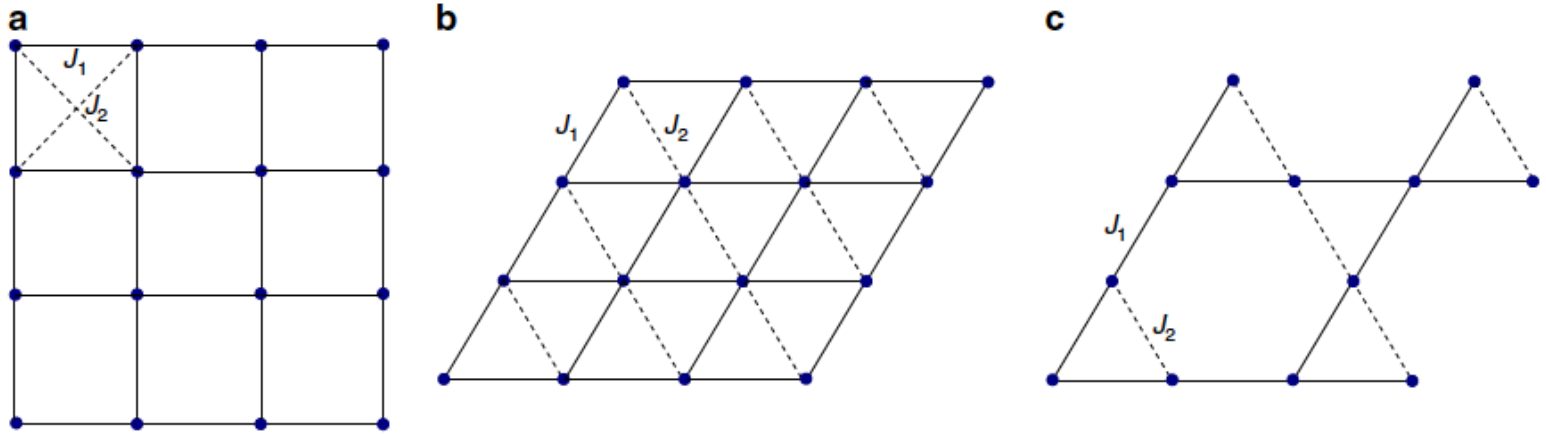


Fig. 1 Lattices considered in this work. We studied three frustrated antiferromagnetic Heisenberg models: **a** next-nearest neighbor J_1–J_2 model on square lattice; **b** anisotropic nearest-neighbor model on triangular lattice; **c** spatially anisotropic Kagome lattice. In all cases $J_2 = 0$ corresponds to the absence of frustration.

24 spins, dimensionality of Hilbert space $d = C_{12}^{24} \simeq 2.7 \cdot 10^6$

Still possible to calculate ground state exactly
Training for $K = 0.01\ d$ (small trial set)

Frustrations and complexity: Quantum case III

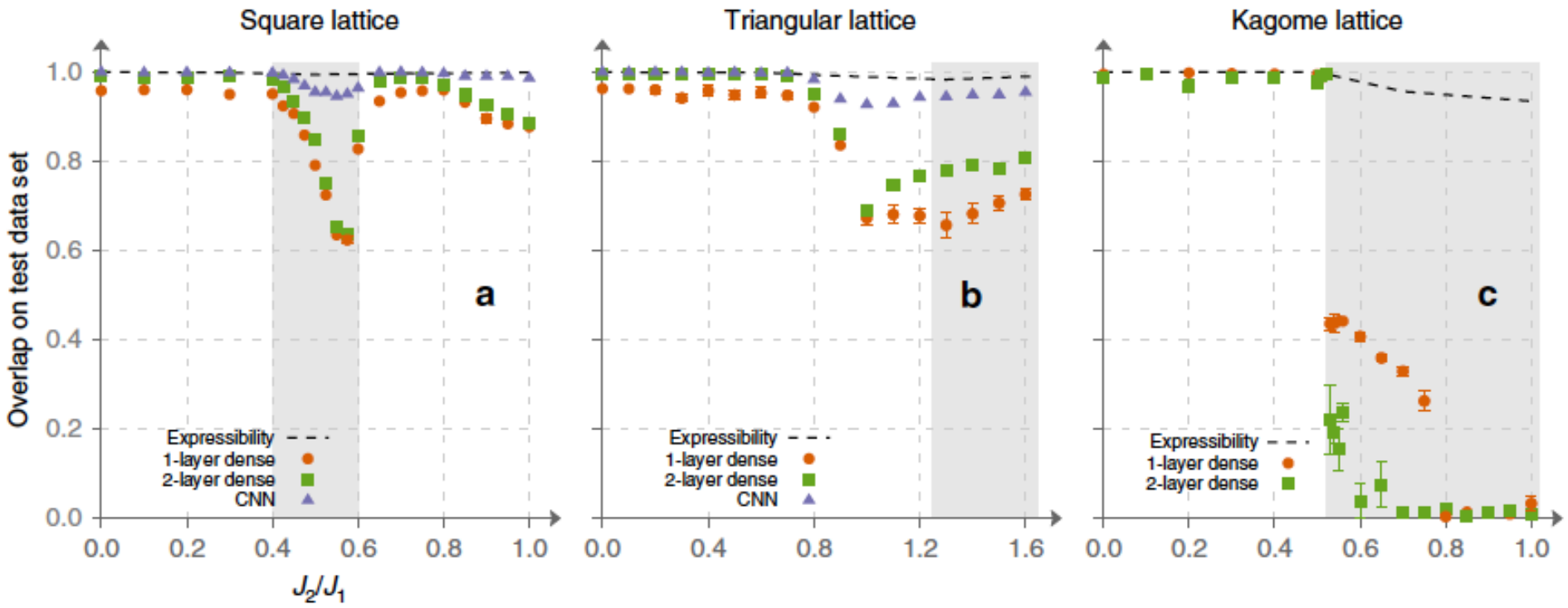


Fig. 2 Optimization results for 24-site clusters obtained with supervised learning and stochastic reconfiguration. Subfigures **a**–**c** were obtained using supervised learning of the sign structure. Overlap of the variational wave function with the exact ground state is shown as function of J_2/J_1 for square **a**, triangular **b**, and Kagome **c** lattices. Overlap was computed on the test dataset (not included into training and validation datasets). Note that generalization is poor in the frustrated regions (which are shaded on the plots). 1-layer dense, 2-layer dense, and convolutional neural network (CNN) architectures are described in Supplementary Note 1. Subfigures **d**–**f** show overlap between the variational wave function optimized using Stochastic Reconfiguration and the exact ground state for square, triangular, and Kagome lattices, respectively. Variational wave function was represented by two two-layer dense networks. A correlation between generalization quality and accuracy of the SR method is evident. On this figure, as well as on all the subsequent ones (both in the main text and Supplementary Notes 1 and 2), error bars represent standard error (SE) obtained by repeating simulations multiple times.

Frustrations and complexity: Quantum case IV

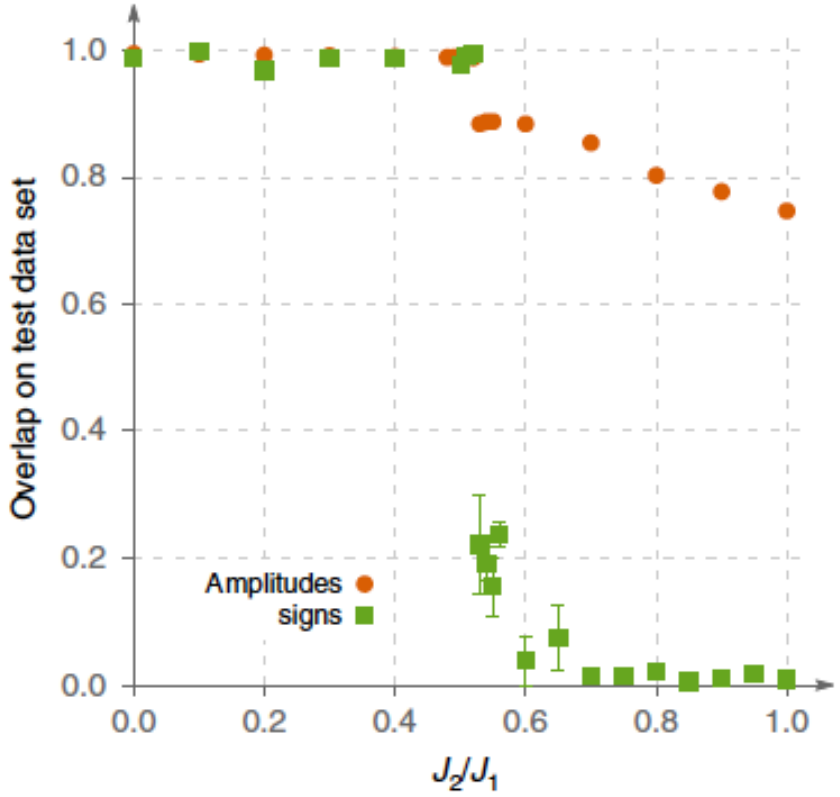


Fig. 4 Generalization of signs and amplitudes. We compare generalization quality as measured by overlap for learning the sign structure (red circles) and amplitude structure (green squares) for 24-site Kagome lattice for two-layer dense architecture. Note that both curves decrease in the frustrated region, but the sign structure is much harder to learn.

It is *sign* structure which is difficult to learn in frustrated case!!!

Relation to sign problem in QMC?!

"Somehow it seems to fill my head with ideas –only I don't exactly know what they are!" (Through the Looking-Glass, and What Alice Found There)

Further development

Many-body quantum sign structures as non-glassy Ising models

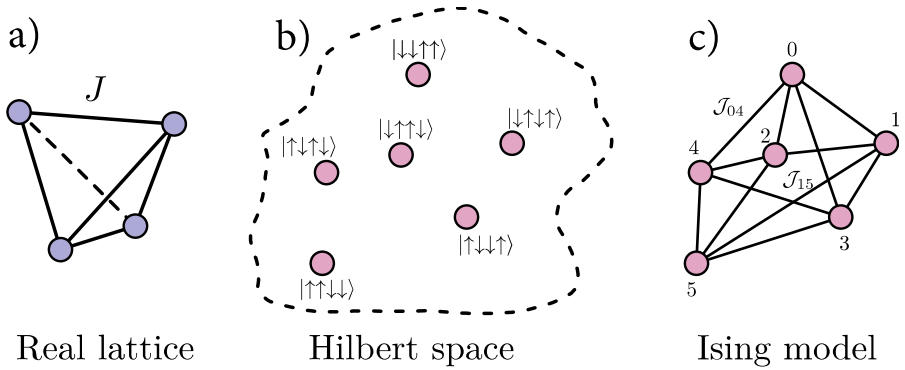
Tom Westerhout, Mikhail I. Katsnelson, Andrey A. Bagrov

Communications Physics **volume 6**, Article number: 275 (2023)

The idea: use machine learning to find amplitudes and then map onto efficient Ising model

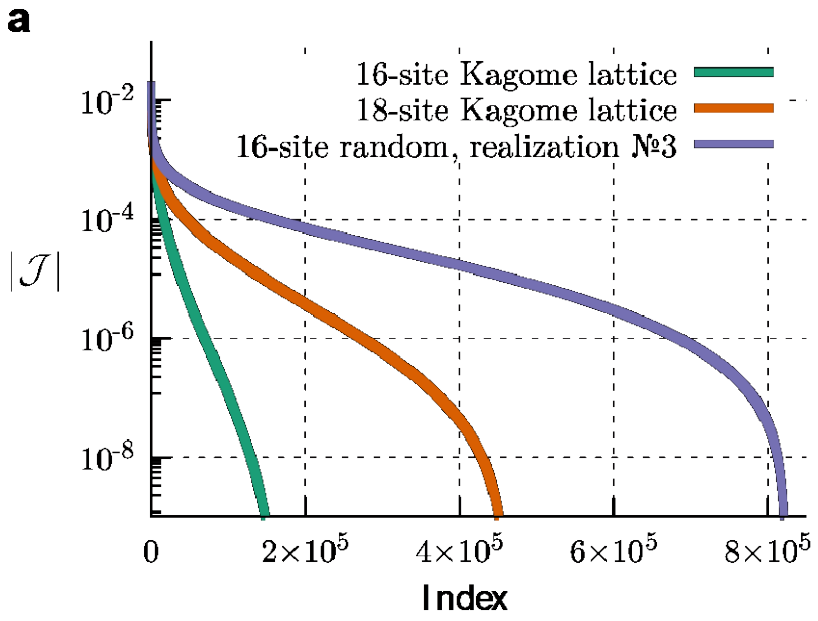
$$|\Psi_{GS}\rangle = \sum_{i=1}^{K} \psi_i |\mathcal{S}_i\rangle = \sum_{i=1}^{K} s_i |\psi_i| |\mathcal{S}_i\rangle$$

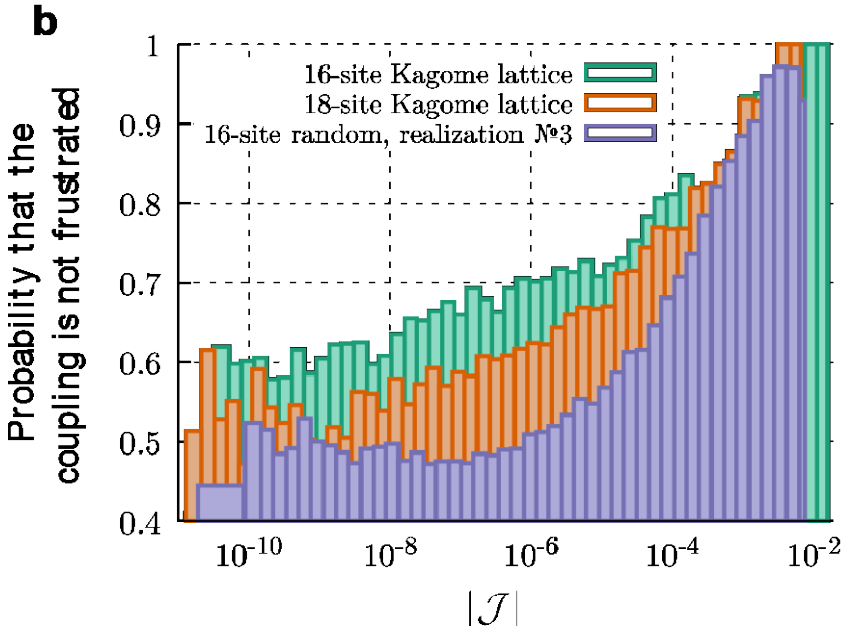
When amplitudes are known the trial ground state energy $< \Psi|H|\Psi >$ is a bilinear function of signs s_i, and we have Ising optimization problem in K-dimensional space; K is very big but it turns out that the model is not glassy and can be optimized without too serious problems



Real lattice Hilbert space Ising model

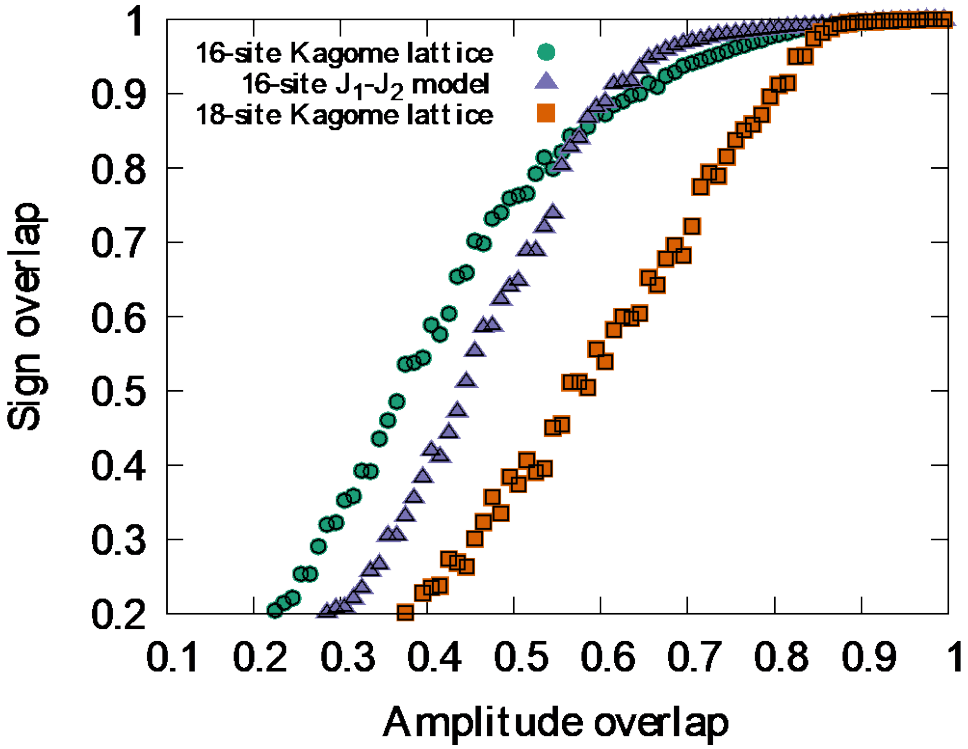
Further development II





It turns out that even for initially frustrated quantum spin models the effective Ising model is not frustrated, both couplings are small and optimization is quite efficient

Further development III



The quality of optimization is quite robust with respect to uncertainties in amplitudes (overlap with the exact ground state)

Thermodynamics/Learning/Evolution

Toward a theory of evolution as multilevel learning

Vitaly Vanchurin[a,b,1], Yuri I. Wolf[a], Mikhail I. Katsnelson[c], and Eugene V. Koonin[a,1]

PNAS 2022 Vol. 119 No. 6 e2120037119

Thermodynamics of evolution and the origin of life

Vitaly Vanchurin[a,b,1], Yuri I. Wolf[a], Eugene V. Koonin[a,1], and Mikhail I. Katsnelson[c,1]

PNAS 2022 Vol. 119 No. 6 e2120042119

Table 1. Corresponding quantities in thermodynamics, machine learning, and evolutionary biology

	Thermodynamics	Machine learning	Evolutionary biology
$\mathbf{x}$	Microscopic physical degrees of freedom	Variables describing training dataset (nontrainable variables)	Variables describing environment
$\mathbf{q}$	Generalized coordinates (e.g., volume)	Weight matrix and bias vector (trainable variables)	Trainable variables (genotype, phenotype)
$H(\mathbf{x},\mathbf{q})$	Energy	Loss function	Additive fitness, $H(x,q)=-T\log f(\mathrm{q})$
$S(\mathbf{q})$	Entropy of physical system	Entropy of nontrainable variables	Entropy of biological system
$U(\mathbf{q})$	Internal energy	Average loss function	Average additive fitness
$Z(T,\mathbf{q})$	Partition function	Partition function	Macroscopic fitness
$F(T,\mathbf{q})$	Helmholtz free energy	Free energy	Adaptive potential (macroscopic additive fitness)
$\Omega(T,\mu)$	Grand potential, $\Omega_p(\mathscr{T},\mathscr{M})$	Grand potential	Grand potential, $\Omega_b(T,\mu)$
T or $\mathscr{T}$	Physical temperature, $\mathscr{T}$	Temperature	Evolutionary temperature, T
μ or $\mathscr{M}$	Chemical potential, $\mathscr{M}$	Absent in conventional machine learning	Evolutionary potential, μ
N_e or N	Number of molecules, N	Number of neurons, N	Effective population size, N_e
K	Absent in conventional physics	Number of trainable variables	Number of adaptable variables

Energy landscape in physics is similar to fitness landscape in biology

Thermodynamics of learning

Consider an arbitrary learning system described by trainable variables $\mathbf{q}$ and nontrainable variables $\mathbf{x}$, such that nontrainable variables undergo stochastic dynamics and trainable variables undergo learning dynamics. In the limit when the nontrainable variables $\mathbf{x}$ have already equilibrated, but the trainable variables $\mathbf{q}$ are still in the process of learning, the conditional probability distribution $p(\mathbf{x}|\mathbf{q})$ over nontrainable variables $\mathbf{x}$ can be obtained from the maximum entropy principle whereby Shannon (or Boltzmann) entropy

$$S = -\int d^N x\, p(\mathbf{x}|\mathbf{q}) \log p(\mathbf{x}|\mathbf{q}) \qquad \textbf{[2.1]}$$

is maximized subject to the appropriate constraints on the system, such as average loss

$$\int d^N x\, H(\mathbf{x},\mathbf{q}) p(\mathbf{x}|\mathbf{q}) = U(\mathbf{q}) \qquad \textbf{[2.2]}$$

and normalization condition

$$\int d^N x\, p(\mathbf{x}|\mathbf{q}) = 1. \qquad \textbf{[2.3]}$$

Its simplicity notwithstanding, the condition **[2.2]** is crucial. This condition means, first, that learning can be mathematically described as minimization of some function $U(\mathbf{q})$ of trainable variables only, and second that this function can be represented as the average of some function $H(\mathbf{x},\mathbf{q})$ of both trainable, $\mathbf{q}$, and nontrainable, $\mathbf{x}$, variables over the space of the latter. Eq.

We postulate that a system under consideration obeys the maximum entropy principle but is also learning or evolving by minimizing the average loss function $U(\mathbf{q})$ **[2.2]**. The corresponding maximum entropy distribution can be calculated using the method of Lagrange multipliers, that is, by solving the following variational problem:

$$\frac{\delta\Big(S - \beta\Big(\int d^N y\, H(\mathbf{y},\mathbf{q}) p(\mathbf{y}|\mathbf{q}) - U\Big) - \nu\Big(\int d^N y\, p(\mathbf{y}|\mathbf{q}) - 1\Big)\Big)}{\delta p(\mathbf{x}|\mathbf{q})} = 0, \qquad \textbf{[2.4]}$$

where β and ν are the Lagrange multipliers which impose, respectively, the constraints **[2.2]** and **[2.3]**. The solution of **[2.4]** is the Boltzmann (or Gibbs) distribution

$$\begin{aligned} -\log p(\mathbf{x}|\mathbf{q}) - 1 - \beta H(\mathbf{x},\mathbf{q}) - \nu = 0 \\ p(\mathbf{x}|\mathbf{q}) = \exp\left(-\beta H(\mathbf{x}) - 1 - \nu\right) = \frac{\exp\left(-\beta H(\mathbf{x},\mathbf{q})\right)}{Z(\beta,\mathbf{q})}, \end{aligned} \qquad \textbf{[2.5]}$$

where

$$Z(\beta,\mathbf{q}) = \exp(1+\nu) = \int d^N x\, \exp\left(-\beta H(\mathbf{x},\mathbf{q})\right) = \int d^N x\, \varphi(\mathbf{x},\mathbf{q}) \qquad \textbf{[2.6]}$$

is the partition function (Z stands for German *Zustandssumme*, sum over states).

Formally, the partition function $Z(\beta,\mathbf{q})$ is simply a normalization constant in Eq. **2.5**, but its dependence on β and $\mathbf{q}$ contains a wealth of information about the learning system and its environment. For example, if the partition function is known, then the average loss can be easily calculated by simple differentiation

$$\begin{aligned} U(\mathbf{q}) &= \frac{\int d^N x\, H(\mathbf{x},\mathbf{q}) \exp\left(-\beta H(\mathbf{x},\mathbf{q})\right)}{\int d^N x\, \exp\left(-\beta H(\mathbf{x},\mathbf{q})\right)} = -\frac{\partial}{\partial\beta}\log Z(\beta,\mathbf{q}) \\ &= \frac{\partial}{\partial\beta}(\beta F(\beta,\mathbf{q})), \end{aligned} \qquad \textbf{[2.7]}$$

where the biological equivalent of free energy is defined as

$$F \equiv -T\log Z = -\beta^{-1}\log Z = U - TS \qquad \textbf{[2.8]}$$

Analogies with biological evolution

Can the change of e.g. biological temperature switch fitness landscape from a few well-defined peaks to a glassy-like with many directions of possible evolution?

Explaining the Cambrian "Explosion" of Animals

Charles R. Marshall

Annu. Rev. Earth Planet. Sci.
2006. 34:355–84

Australian Journal of Zoology
http://dx.doi.org/10.1071/ZO13052

The evolution of morphogenetic fitness landscapes: conceptualising the interplay between the developmental and ecological drivers of morphological innovation

Charles R. Marshall

Cambrian Exposion as an analog of magnetic phase transitions in neodymium?!

Well… for me (as a physicist) it is a good place to stop

Fundamental principles of evolution

What are the requirements for a universe to be observable? The possibility to make meaningful observations implies a degree of order and complexity in the observed universe emerging from evolutionary processes, and such evolvability itself seems to be predicated on several fundamental principles. It has to be emphasized that "observation" and "learning" here by no means imply "mind" or "consciousness" but a far more basic requirement. To learn and survive in an environment, a system (or observer) must predict, with some minimal but sufficient degree of accuracy, the response of that environment to various actions and to be able to choose such actions that are compatible with the observer's continued existence in that environment. In this sense, any life-form is an observer, and so are even inanimate systems endowed with the ability of feedback reaction. In this most general sense, observation is a prerequisite for evolution. We first formulate the basic principles underlying observability and evolvability and then, give the pertinent comments and explanations.

P1. Loss function. In any evolving system, there exists a loss function of time-dependent variables that is minimized during evolution.

P2. Hierarchy of scales. Evolving systems encompass multiple dynamical variables that change on different temporal scales (with different characteristic frequencies).

P3. Frequency gaps. Dynamical variables are split among distinct levels of organization separated by sufficiently wide frequency gaps.

P4. Renormalizability. Across the entire range of organization of evolving systems, a statistical description of faster-changing (higher-frequency) variables is feasible through the slower-changing (lower-frequency) variables.

P5. Extension. Evolving systems have the capacity to recruit additional variables that can be utilized to sustain the system and the ability to exclude variables that could destabilize the system.

P6. Replication. In evolving systems, replication and elimination of the corresponding information-processing units (IPUs) can take place on every level of organization.

P7. Information flow. In evolving systems, slower-changing levels absorb information from faster-changing levels during learning and pass information down to the faster levels for prediction of the state of the environment and the system itself.

Fundamental evolutionary phenomena

E1. IPUs. Discrete IPUs (that is, self- vs. nonself-differentiation and discrimination) exist at all levels of organization. All biological systems at all levels of organization, such as genes, cells, organisms, populations, and so on up to the level of the entire biosphere, possess some degree of self-coherence that separates them, first and foremost, from the environment at large and from other similar-level IPUs.

E2. Frustration. All complex, dynamical systems face multidimensional and multiscale optimization problems, which generically lead to frustration resulting from conflicting objectives at different scales. This is a key, intrinsic feature of all such systems and a major force driving the advent of increasing multilevel complexity (12). Frustration is an extremely general physical phenomenon that is by no account limited to biology but rather, occurs already in much simpler physical systems, such as spin and structural glasses, the behavior of which is determined by competing interactions so that a degree of complexity is attained (31, 32).

E3. Multilevel Hierarchy. The hierarchy of multiple levels of organization is an intrinsic, essential feature of evolving biological systems in terms of both the structure of these systems (genes, genomes, cells, organisms, kin groups, populations, species, communities, and more) and the substrate the evolutionary forces act upon.

E4. Near Optimality. Stochastic optimization or the use of stochastic optimization algorithms is the only feasible approach to complex optimization, but it guarantees neither finding the globally optimal solution nor retention of the optimal configuration when and if it is found. Rather, stochastic optimization tends to rapidly find local optima and keeps the system in their vicinity, sustaining the value of the loss function at a near-optimal level.

E5. Diversity of Near-Optimal Solutions. Solutions on the loss function landscapes that arise in complex optimization problems span numerous local peaks of comparable heights.

E6. Separation of Phenotype from Genotype. This quintessential feature of life embodies two distinct (albeit inseparable in known organisms) symmetry-breaking phenomena: 1) separation between dedicated digital information storage media (stable, rarely updatable, tending to distributions with discrete values) and mostly analog processing devices and 2) asymmetry of the information flow within the IPUs whereby the genotype provides "instructions" for the phenotype, whereas the phenotype largely loses the ability to update the genotype directly. The separation between the information storage and processing subsystems is a prerequisite for efficient evolution that probably emerged early on the path from prebiotic entities to the emergence of life.

E7. Replication. Emergence of long-term digital storage devices, that is genomes consisting of RNA or DNA (E6) provides for long-term information preservation, facilitates adaptive reactions to changes in the environment, and promotes the stability of IPUs to the point where (at least in chemical systems) it is limited by the energy of the chemical bonds rather than the energy of thermal fluctuations. Obviously, however, as long as this information is confined to a single IPU, it will disappear with the inevitable eventual destruction of that IPU. Should this be the case, other IPUs of similar architecture would need to accumulate a comparable amount of information from scratch to reach the same level of stability. Thus, copying and sharing information are essential for long-term (effectively, indefinite) persistence of IPUs.

E8. Natural Selection. Evolution by natural selection (Darwinian evolution) arises from the combination of all the principles and phenomena described above.

E9. Parasitism. Parasites and host–parasite coevolution are ubiquitous across biological systems at multiple levels of organization and are both intrinsic to and indispensable for the evolution of life.

E10. Programmed Death. Programmed (to various degrees) death is an intrinsic feature of life.

To summarize

Whether you can observe a thing or not
depends on the theory which you use.
It is theory which decides what can be observed
(A. Einstein)

MANY THANKS FOR YOUR ATTENTION