Title: Replacing neural networks by optimal predictive models for the detection of phase transitions

Speakers: Julian Arnold

Series: Machine Learning Initiative

Date: February 24, 2023 - 11:00 AM

URL: https://pirsa.org/23020061

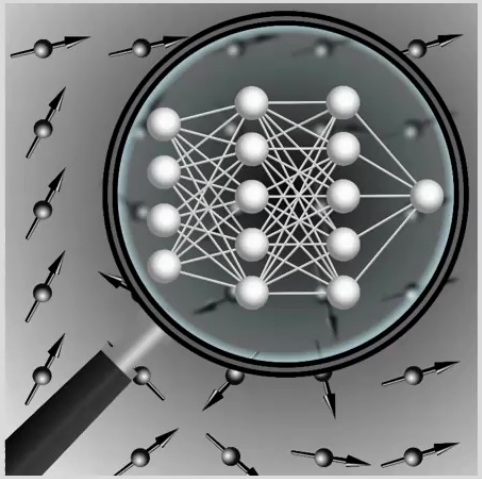
Abstract:

In recent years, machine learning has been successfully used to identify phase transitions and classify phases of matter in a data-driven manner. Neural network (NN)-based approaches are particularly appealing due to the ability of NNs to learn arbitrary functions. However, the larger an NN, the more computational resources are needed to train it, and the more difficult it is to understand its decision making. Thus, we still understand little about the working principle of such machine learning approaches, when they fail or succeed, and how they differ from traditional approaches. In this talk, I will present analytical expressions for the optimal predictions of three popular NN-based methods for detecting phase transitions that rely on solving classification and regression tasks using supervised learning at their core. These predictions are optimal in the sense that they minimize the target loss function. Therefore, in practice, optimal predictive models are well approximated by high-capacity predictive models, such as large NNs after ideal training. I will show that the analytical expressions we have derived provide a deeper understanding of a variety of previous NN-based studies and enable a more efficient numerical routine for detecting phase transitions from data.

Zoom Link: https://pitp.zoom.us/j/91642481966?pwd=alkrWEFFcFBvRlJEbDRBZWV3MFFDUT09

Replacing neural networks by optimal predictive models for the detection of phase transitions

Julian Arnold
Department of Physics, University of Basel
Bruder group

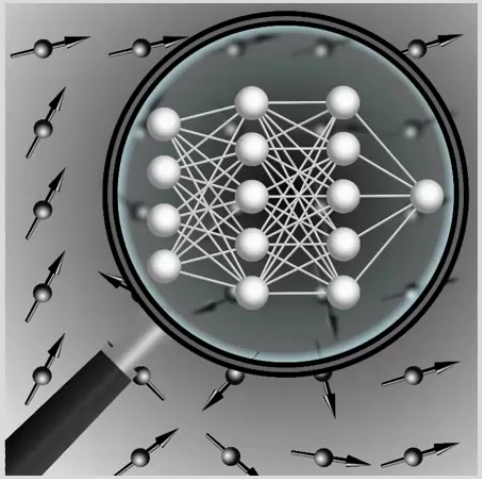


JA and Schäfer, Phys. Rev. X **12**, 031044 (2022)

FNSNF Swiss National Science Foundation
QSIT Quantum Science and Technology National Centre of Competence in Research
University of Basel

Replacing neural networks by optimal predictive models for the detection of phase transitions

Julian Arnold
Department of Physics, University of Basel
Bruder group

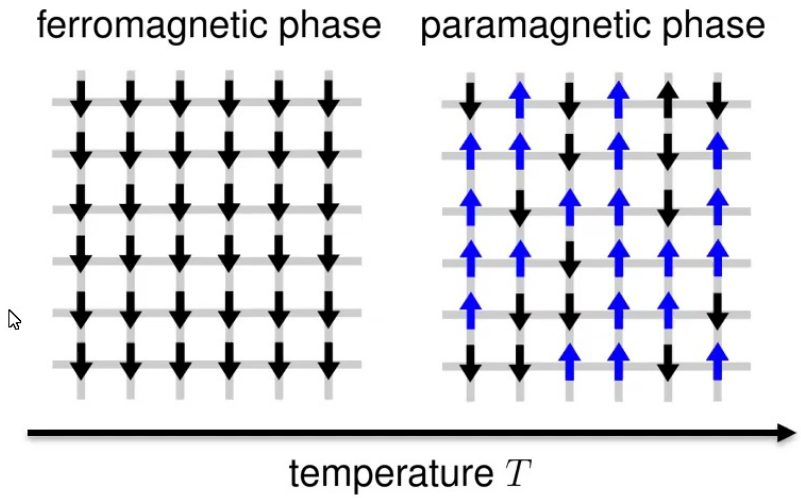


JA and Schäfer, Phys. Rev. X **12**, 031044 (2022)

FNSNF Swiss National Science Foundation
QSIT Quantum Science and Technology National Centre of Competence in Research
University of Basel

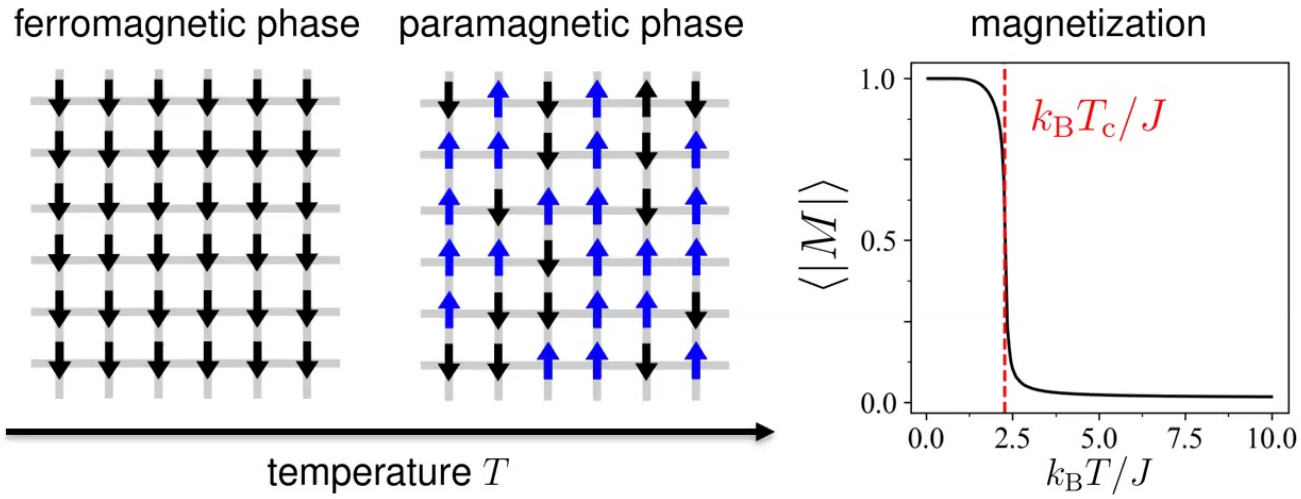
Symmetry-breaking phase transition in Ising model

- Hamiltonian $H(\boldsymbol{\sigma}) = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j$
- spin configuration $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_{L\times L})$, with $\sigma_i \in \{+1, -1\}$
- Boltzmann distribution $\mathrm{P}(\boldsymbol{\sigma}) = \dfrac{e^{-H(\boldsymbol{\sigma})/k_\mathrm{B}T}}{Z}$



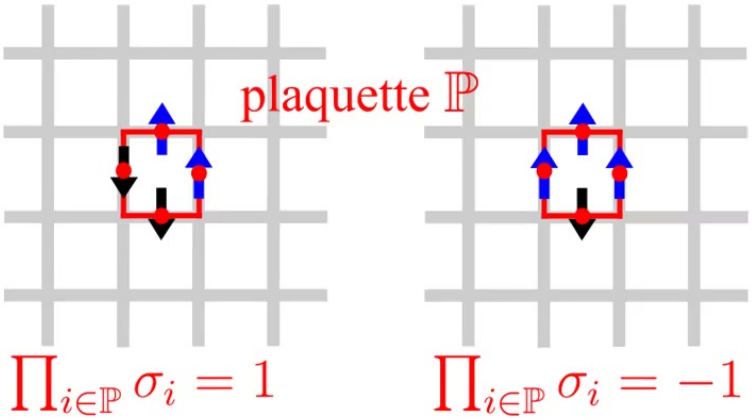
Symmetry-breaking phase transition in Ising model

- magnetization serves as order parameter $M(\boldsymbol{\sigma}) = \frac{1}{L^2}\sum_i \sigma_i$
- Onsager's solution $k_{\mathrm{B}}T_{\mathrm{c}}/J = \frac{2}{\ln(1+\sqrt{2})}$



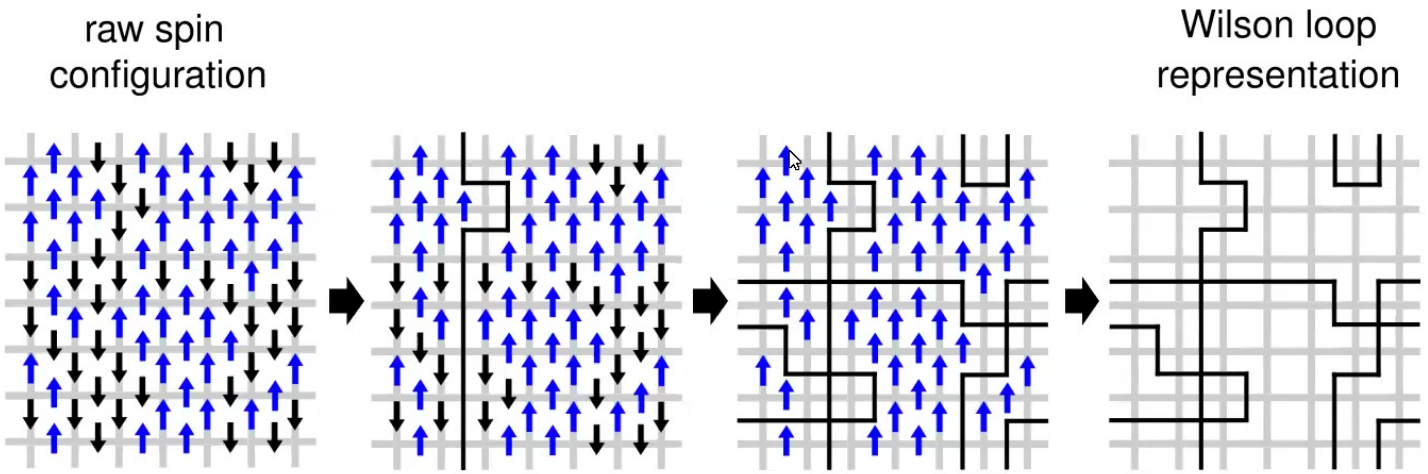
Topological crossover in Ising gauge theory

- Hamiltonian $H(\boldsymbol{\sigma}) = -J \sum_{\mathbb{P}} \prod_{i \in \mathbb{P}} \sigma_i$

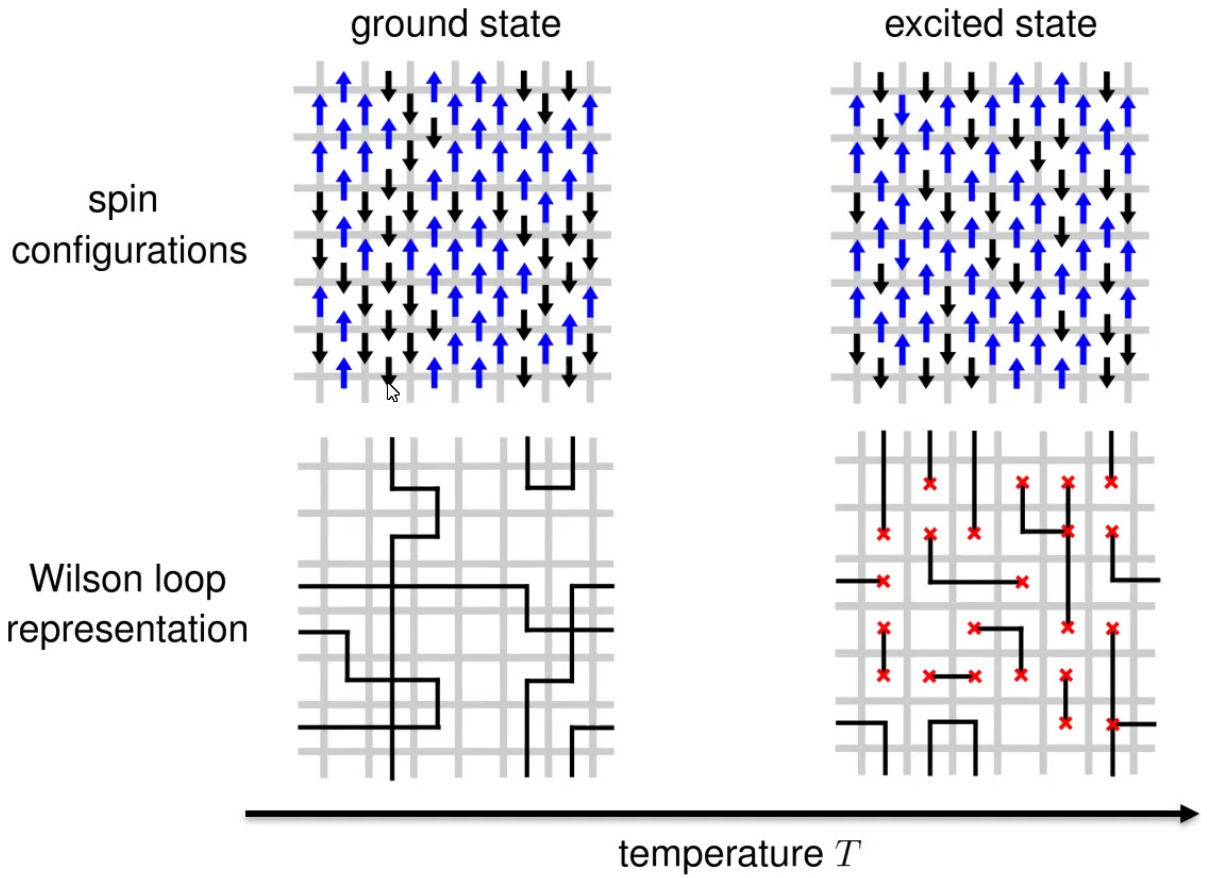


- spin configuration $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_{L \times L})$, with $\sigma_i \in \{+1, -1\}$
- Boltzmann distribution $\mathrm{P}(\boldsymbol{\sigma}) = \dfrac{e^{-H(\boldsymbol{\sigma})/k_{\mathrm{B}}T}}{Z}$

Topological crossover in Ising gauge theory



Topological crossover in Ising gauge theory

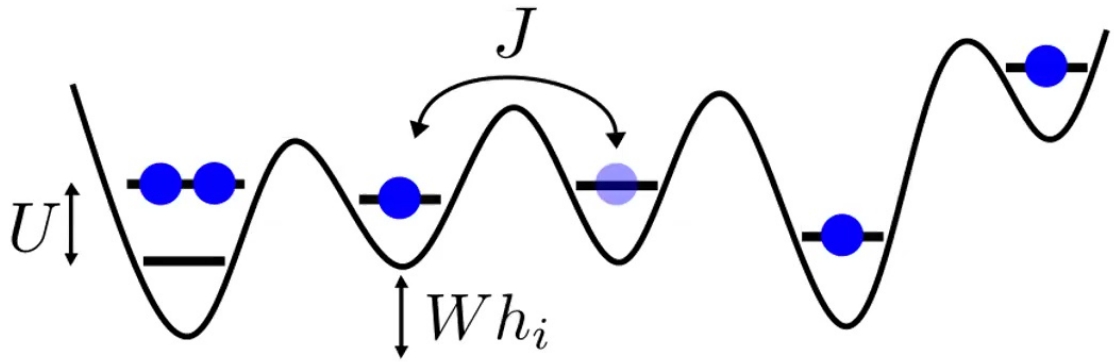


Many-body localization phase transition in Bose-Hubbard chain

- Hamiltonian $\hat{H} = -J \sum_{i=1}^{L-1} (\hat{b}_{i+1}^{\dagger} \hat{b}_i + \text{H.c.}) + \sum_{i=1}^{L} \frac{U}{2} \hat{n}_i (\hat{n}_i - 1) + W h_i \hat{n}_i$

  $\Rightarrow$ fixed interaction strength $U/J = 2.9$

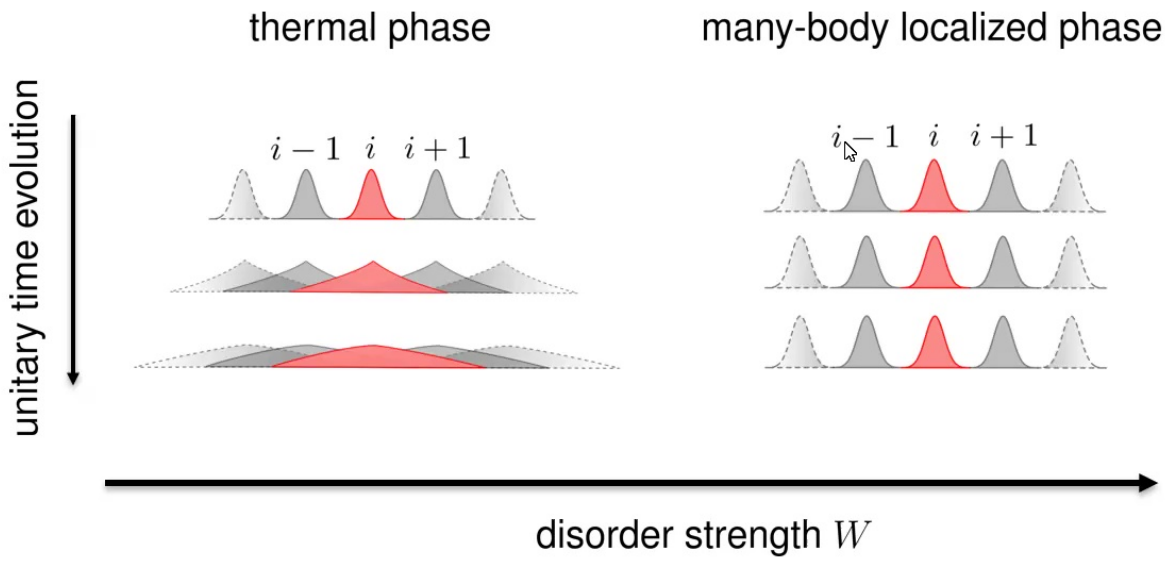
  $\Rightarrow$ quasi-periodic potential $h_i = \cos(2\pi\alpha i + \phi)$, where α is fixed and $\phi \in [0, 2\pi)$ mimicking on-site disorder with strength W



Many-body localization phase transition in Bose-Hubbard chain

- Hamiltonian $\hat{H} = -J \sum_{i=1}^{L-1} (\hat{b}_{i+1}^{\dagger} \hat{b}_i + \text{H.c.}) + \sum_{i=1}^{L} \frac{U}{2} \hat{n}_i (\hat{n}_i - 1) + W h_i \hat{n}_i$

  $\Rightarrow$ let system evolve starting from Mott-insulating state $|\Psi(0)\rangle = |11 \ldots 1\rangle$



Motivation

- detecting phase transitions autonomously from readily accessible data

  $\Rightarrow$ does not require prior theoretical knowledge

  $\Rightarrow$ could enable discovery of new phases of matter

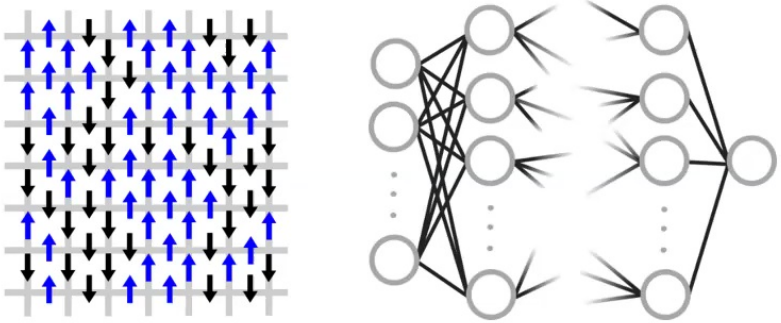
Motivation

- detecting phase transitions autonomously from readily accessible data

  $\Rightarrow$ does not require prior theoretical knowledge

  $\Rightarrow$ could enable discovery of new phases of matter

use (deep) neural networks which proved to be successful in traditional image classification tasks



Detecting phase transitions using neural networks

- supervised learning (SL)

nature physics
LETTERS
PUBLISHED ONLINE: 13 FEBRUARY 2017 | DOI: 10.1038/NPHYS4035

Machine learning phases of matter

Juan Carrasquilla[1]* and Roger G. Melko[1,2]

- learning by confusion (LBC)

nature physics
LETTERS
PUBLISHED ONLINE: 13 FEBRUARY 2017 | DOI: 10.1038/NPHYS4037

Learning phase transitions by confusion

Evert P. L. van Nieuwenburg*, Ye-Hua Liu and Sebastian D. Huber

- prediction-based method (PBM)

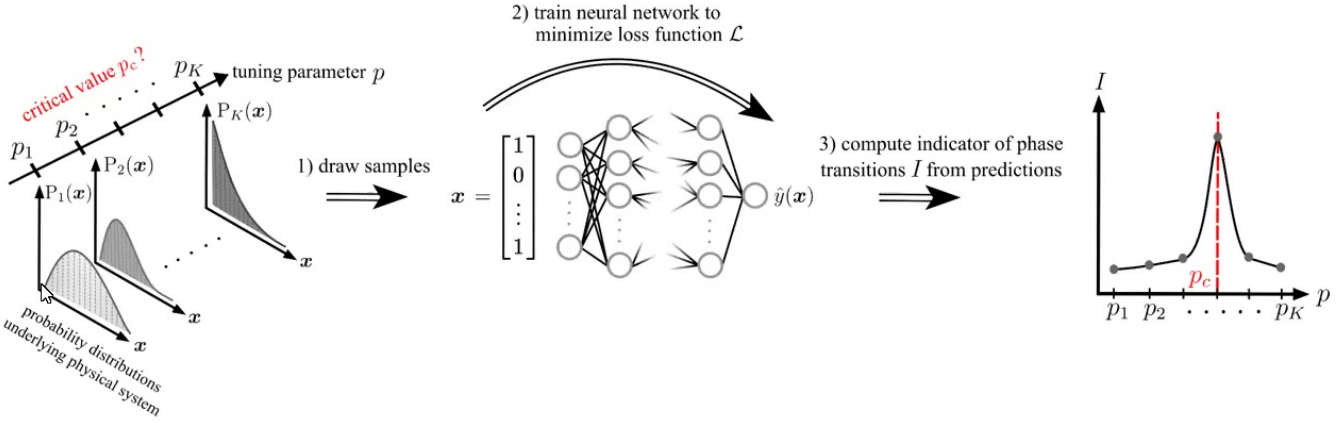
PHYSICAL REVIEW E **99**, 062107 (2019)

Vector field divergence of predictive model output as indication of phase transitions

Frank Schäfer and Niels Lörch*
Department of Physics, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, Switzerland

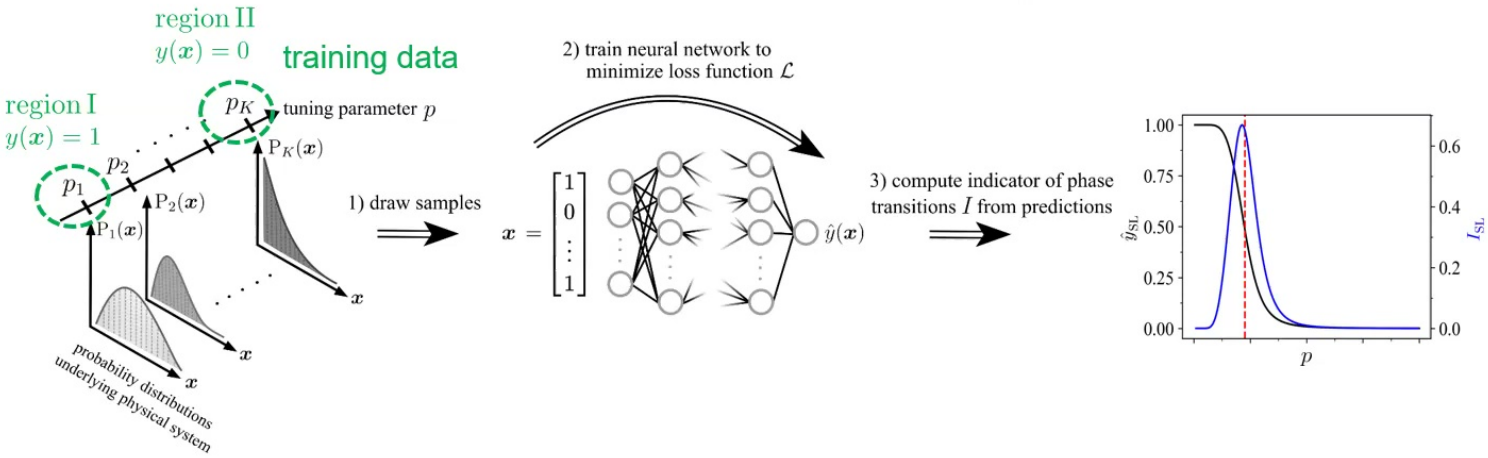
(Received 3 December 2018; revised manuscript received 16 May 2019; published 5 June 2019)

Detecting phase transitions using neural networks



Supervised learning

M number of drawn samples at each parameter value
K number of sampled parameter values
$\mathcal{X}_k$ data set at parameter value p_k
$y(\boldsymbol{x})$ label of sample $\boldsymbol{x}$
$\hat{y}(\boldsymbol{x})$ prediction for sample $\boldsymbol{x}$



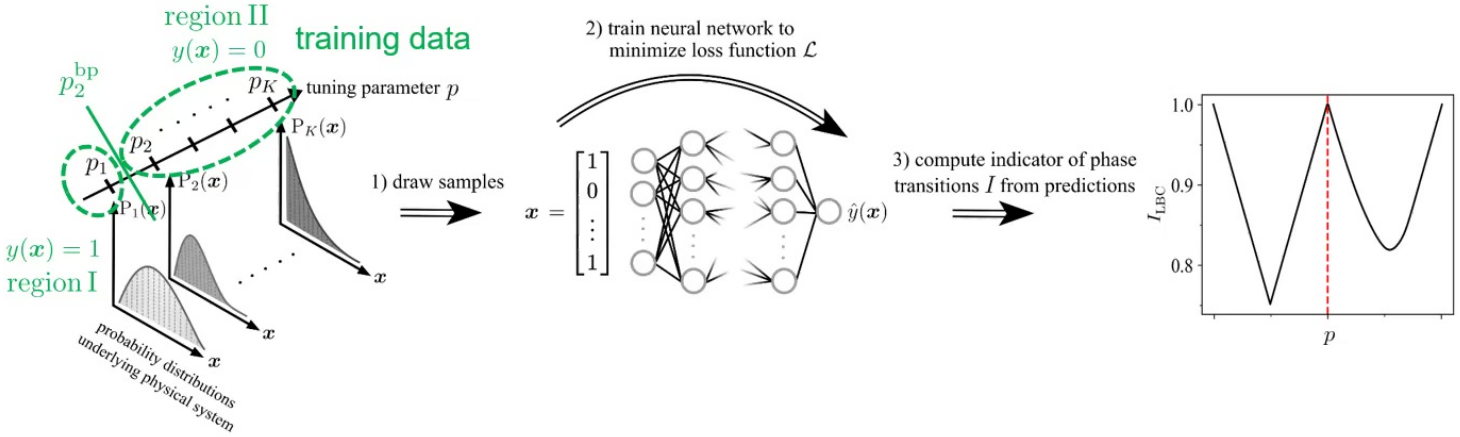
- train neural network to minimize binary cross-entropy loss $\mathcal{L}_{\mathrm{SL}}$
- compute mean prediction at each sampled value of tuning parameter

$$\hat{y}_{\mathrm{SL}}(p_k) = \frac{1}{M} \sum_{\boldsymbol{x} \in \mathcal{X}_k} \hat{y}(\boldsymbol{x})$$

- compute indicator of phase transitions $I_{\mathrm{SL}}(p_k) = -\left.\frac{\partial \hat{y}_{\mathrm{SL}}(p)}{\partial p}\right|_{p_k}$

Learning by confusion

M number of drawn samples at each parameter value
K number of sampled parameter values
$\mathcal{X}_k$ data set at parameter value p_k
$y(\boldsymbol{x})$ label of sample $\boldsymbol{x}$
$\hat{y}(\boldsymbol{x})$ prediction for sample $\boldsymbol{x}$

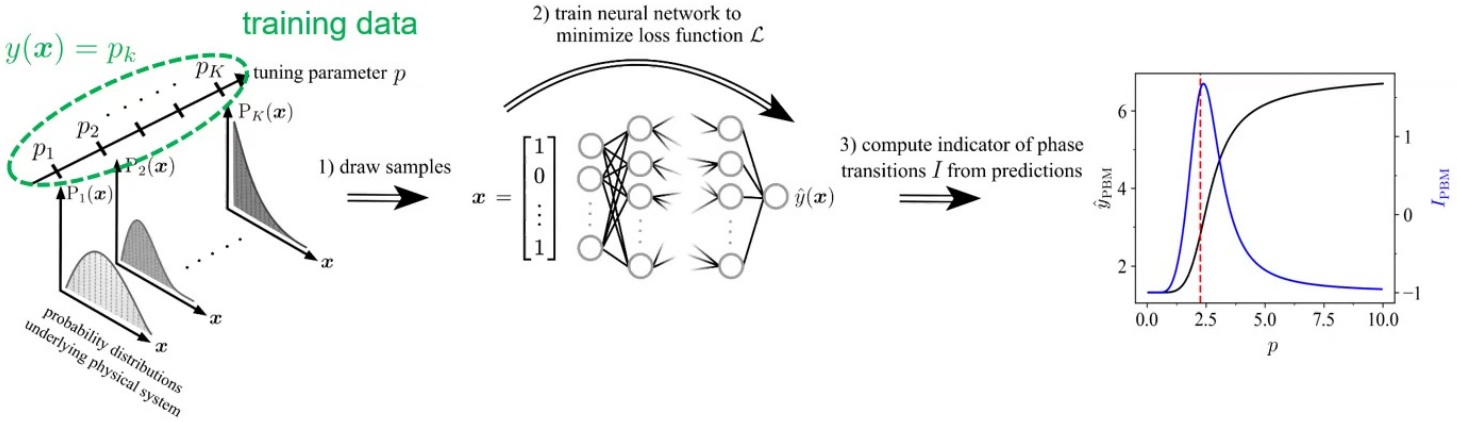


for each bipartition point p_k^{bp}

- train neural network to minimize binary cross-entropy loss $\mathcal{L}_{\mathrm{LBC}}$
- compute mean classification accuracy which serves as an indicator of phase transitions $I_{\mathrm{LBC}}(p_k^{\mathrm{bp}})$

Prediction-based method

M number of drawn samples at each parameter value
K number of sampled parameter values
$\mathcal{X}_k$ data set at parameter value p_k
$y(\boldsymbol{x})$ label of sample $\boldsymbol{x}$
$\hat{y}(\boldsymbol{x})$ prediction for sample $\boldsymbol{x}$



- train neural network to minimize mean-squared error loss $\mathcal{L}_{\mathrm{PBM}}$
- compute mean prediction at each sampled value of tuning parameter

  $$\hat{y}_{\mathrm{PBM}}(p_k) = \frac{1}{M} \sum_{\boldsymbol{x} \in \mathcal{X}_k} \hat{y}(\boldsymbol{x})$$

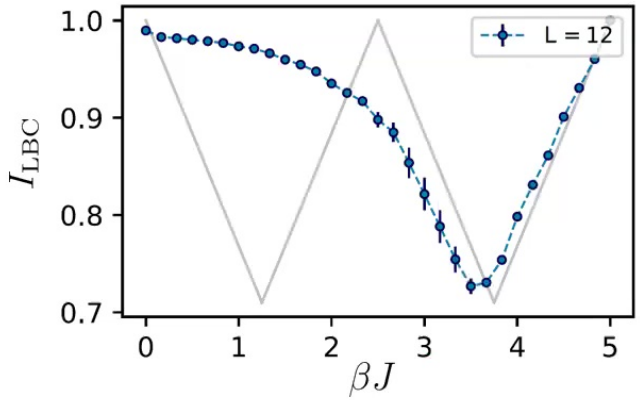
- compute indicator of phase transitions $I_{\mathrm{PBM}}(p_k) = \left. \frac{\partial \delta y_{\mathrm{PBM}}(p)}{\partial p} \right|_{p_k} = \left. \frac{\partial \hat{y}_{\mathrm{PBM}}(p)}{\partial p} \right|_{p_k} - 1$
  with $\delta y_{\mathrm{PBM}}(p_k) = \hat{y}_{\mathrm{PBM}}(p_k) - p_k$

What's the problem?

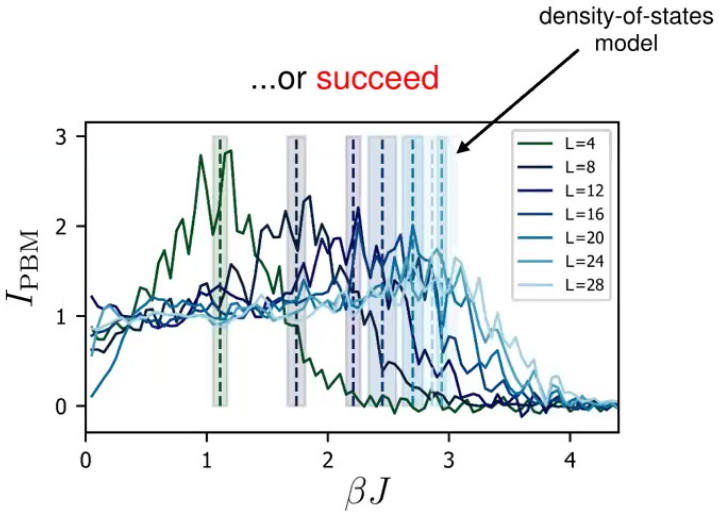
- methods were motivated in a heuristic fashion
- (deep) neural networks are difficult to interpret

  ⇒ have limited understanding of their working principle

unclear when and why they fail...



...or succeed



Greplova *et al.*, New J. Phys. **22** 045003 (2020)

What's the problem?

- methods were motivated in a heuristic fashion
- (deep) neural networks are difficult to interpret

  $\Rightarrow$ have limited understanding of their working principle

high model capacity

$\Updownarrow$

low interpretability and high computational cost

Regaining interpretability through reduced capacity

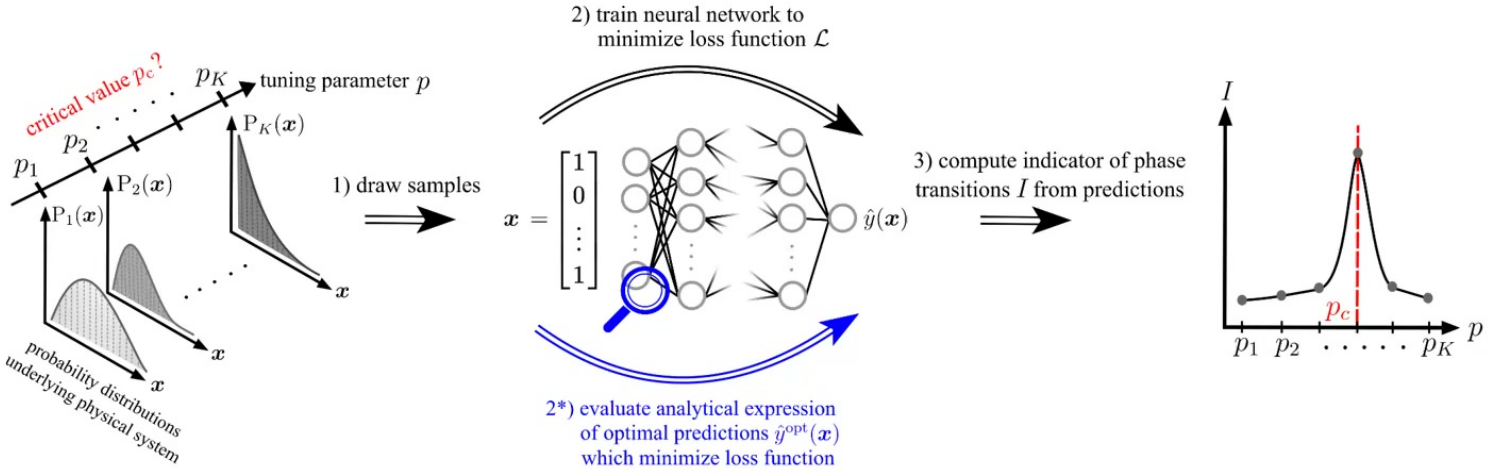
low model capacity
$\Downarrow$
high interpretability

many works up to now...

- ...use linear methods, such as (kernel) support vector machines
- ...engineer input features or reduce neural network size systematically

  $\Rightarrow$ neural network can be well approximated by linear function
- ...use standard ML interpretability tools: analyze dependence of output on input data via truncated Taylor expansion

Replacing neural networks by optimal predictive models



$$\arg\min_{\hat{y}(\boldsymbol{x})} \mathcal{L} = \hat{y}^{\mathrm{opt}}(\boldsymbol{x})$$

$$\lim_{\text{model capacity} \to \infty} \hat{y}(\boldsymbol{x}) = \hat{y}^{\mathrm{opt}}(\boldsymbol{x})$$

$\mathrm{P}_k(\boldsymbol{x})$ probability to draw sample $\boldsymbol{x}$ at parameter value p_k
$\mathrm{I/II}$ denotes the two training regions

Optimal predictions

- supervised learning: $\hat{y}_{\mathrm{SL}}^{\mathrm{opt}}(\boldsymbol{x}) = \dfrac{\mathrm{P_I}(\boldsymbol{x})}{\mathrm{P_I}(\boldsymbol{x}) + \mathrm{P_{II}}(\boldsymbol{x})}$, where $\mathrm{P_{I/II}}(\boldsymbol{x}) = \sum_{k \in \mathrm{I/II}} \mathrm{P}_k(\boldsymbol{x})$

- learning by confusion: $\hat{y}_{\mathrm{LBC}}^{\mathrm{opt}}(\boldsymbol{x}) = \dfrac{\mathrm{P_I}(\boldsymbol{x})}{\mathrm{P_I}(\boldsymbol{x}) + \mathrm{P_{II}}(\boldsymbol{x})}$

- prediction-based method: $\hat{y}_{\mathrm{PBM}}^{\mathrm{opt}}(\boldsymbol{x}) = \dfrac{\sum_{k=1}^{K} \mathrm{P}_k(\boldsymbol{x})\, p_k}{\sum_{k=1}^{K} \mathrm{P}_k(\boldsymbol{x})}$

$\Rightarrow$ calculate optimal indicator I^{opt} given optimal predictions $\hat{y}^{\mathrm{opt}}$

$\mathrm{P}_k(\boldsymbol{x})$ probability to draw sample x at parameter value p_k
$\mathrm{I/II}$ denotes the two training regions

Optimal predictions

- supervised learning: $\hat{y}_{\mathrm{SL}}^{\mathrm{opt}}(\boldsymbol{x}) = \dfrac{\mathrm{P_I}(\boldsymbol{x})}{\mathrm{P_I}(\boldsymbol{x}) + \mathrm{P_{II}}(\boldsymbol{x})}$, where $\mathrm{P}_{\mathrm{I/II}}(\boldsymbol{x}) = \sum_{k \in \mathrm{I/II}} \mathrm{P}_k(\boldsymbol{x})$
- learning by confusion: $\hat{y}_{\mathrm{LBC}}^{\mathrm{opt}}(\boldsymbol{x}) = \dfrac{\mathrm{P_I}(\boldsymbol{x})}{\mathrm{P_I}(\boldsymbol{x}) + \mathrm{P_{II}}(\boldsymbol{x})}$
- prediction-based method: $\hat{y}_{\mathrm{PBM}}^{\mathrm{opt}}(\boldsymbol{x}) = \dfrac{\sum_{k=1}^{K} \mathrm{P}_k(\boldsymbol{x})\, p_k}{\sum_{k=1}^{K} \mathrm{P}_k(\boldsymbol{x})}$

expressions reveal dependence of output on input data

$\Rightarrow$ phase transitions are detected as changes in underlying probability distributions

Optimal predictions

$\mathrm{P}_k(\boldsymbol{x})$ probability to draw sample x at parameter value p_k
I/II denotes the two training regions
M number of drawn samples at each parameter value
$M_k(\boldsymbol{x})$ number times sample x is drawn at p_k

- supervised learning: $\hat{y}^{\mathrm{opt}}_{\mathrm{SL}}(\boldsymbol{x}) = \dfrac{\mathrm{P}_{\mathrm{I}}(\boldsymbol{x})}{\mathrm{P}_{\mathrm{I}}(\boldsymbol{x}) + \mathrm{P}_{\mathrm{II}}(\boldsymbol{x})}$, where $\mathrm{P}_{\mathrm{I/II}}(\boldsymbol{x}) = \sum_{k \in \mathrm{I/II}} \mathrm{P}_k(\boldsymbol{x})$

- learning by confusion: $\hat{y}^{\mathrm{opt}}_{\mathrm{LBC}}(\boldsymbol{x}) = \dfrac{\mathrm{P}_{\mathrm{I}}(\boldsymbol{x})}{\mathrm{P}_{\mathrm{I}}(\boldsymbol{x}) + \mathrm{P}_{\mathrm{II}}(\boldsymbol{x})}$

- prediction-based method: $\hat{y}^{\mathrm{opt}}_{\mathrm{PBM}}(\boldsymbol{x}) = \dfrac{\sum_{k=1}^{K} \mathrm{P}_k(\boldsymbol{x})\, p_k}{\sum_{k=1}^{K} \mathrm{P}_k(\boldsymbol{x})}$

  $\Rightarrow$ can be evaluated given

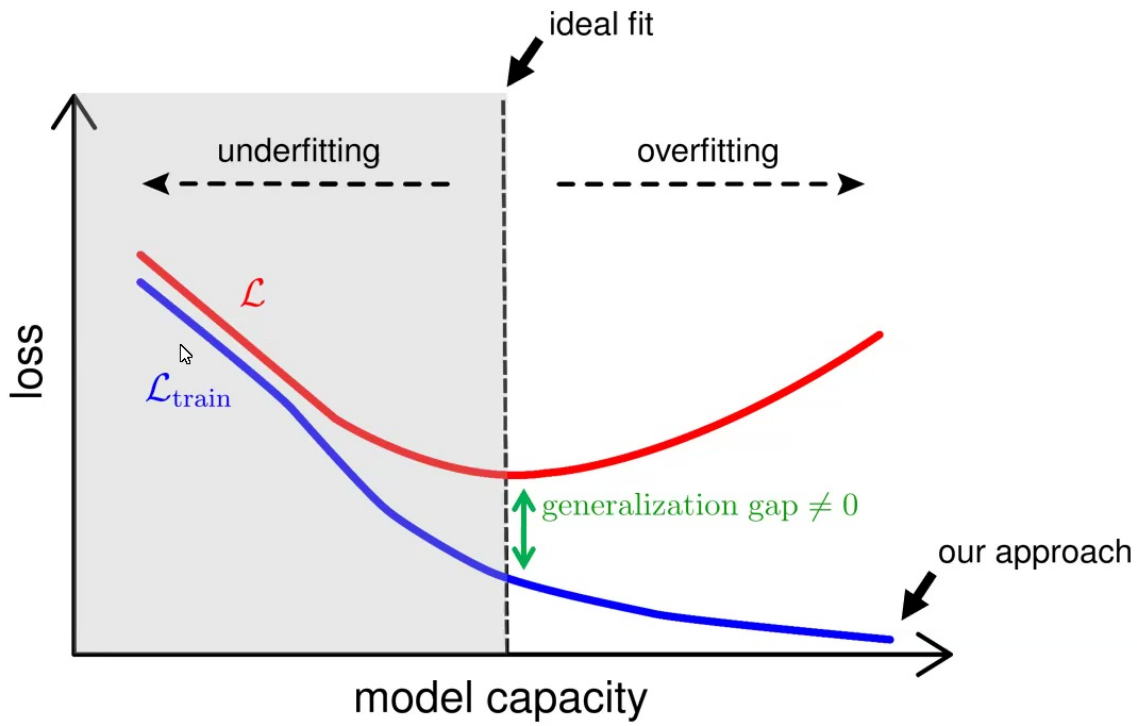
  a) analytical expression for $\mathrm{P}_k(\boldsymbol{x})$

  b) exact numerical values for $\mathrm{P}_k(\boldsymbol{x})$

  c) numerical estimate for $\mathrm{P}_k(\boldsymbol{x})$, e.g., from drawn samples $\mathrm{P}_k(\boldsymbol{x}) \approx M_k(\boldsymbol{x})/M$

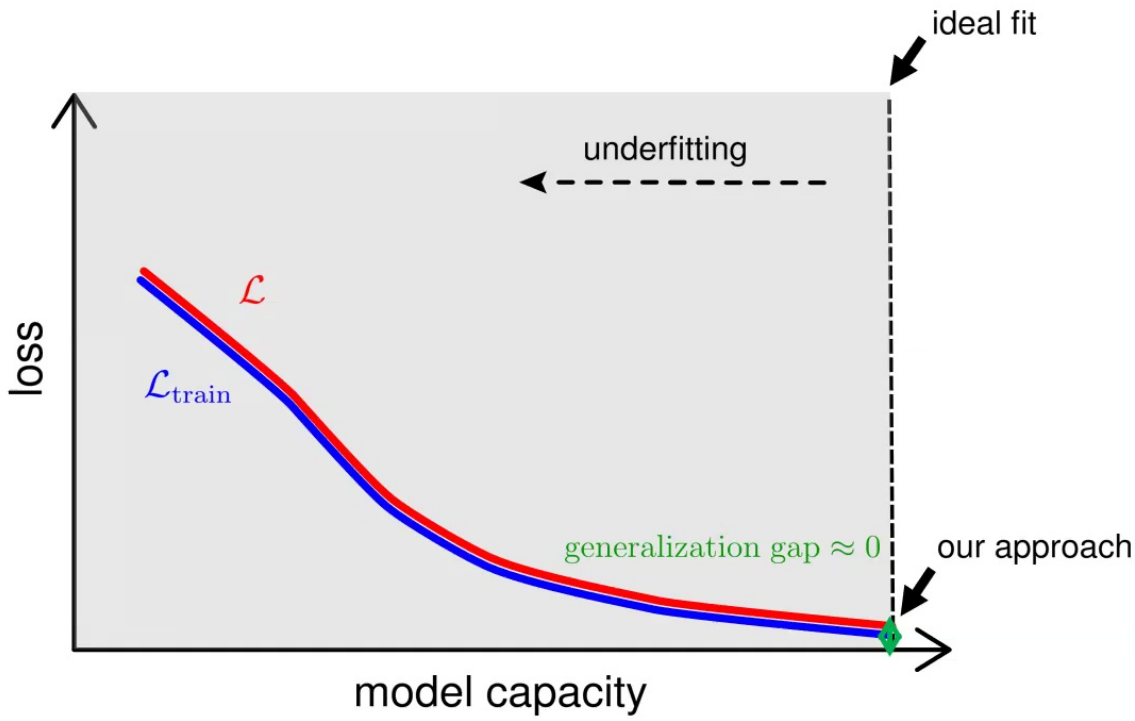
  $\Rightarrow$ empirically optimal predictions

Empirically optimal predictions: Small data set



Vapnik, *The Nature of Statistical Learning Theory*, Springer (2000)

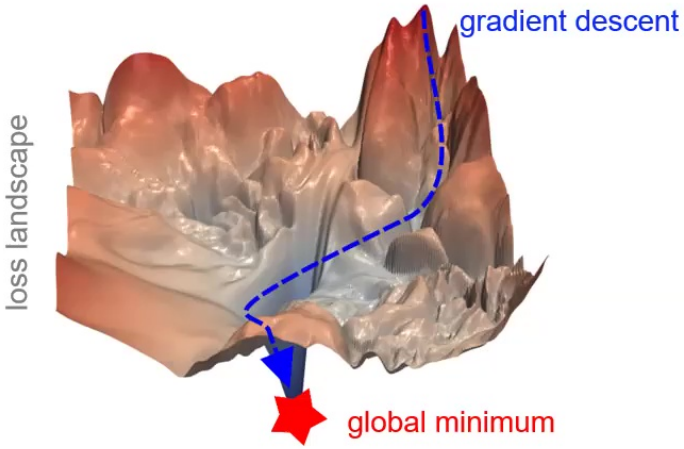
Empirically optimal predictions: Large data set



Vapnik, *The Nature of Statistical Learning Theory*, Springer (2000)

Reaching high model capacity with neural networks

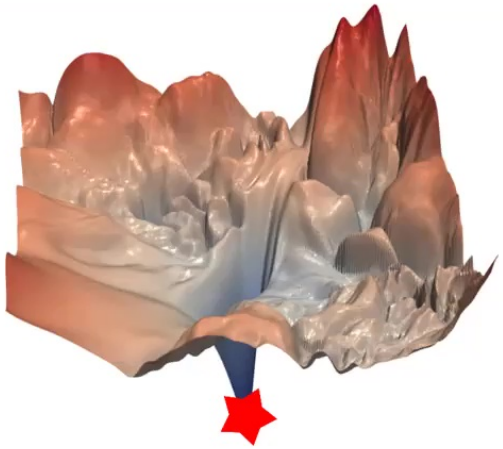
training neural networks



⇒ requires
- large neural networks
- many training epochs
- hyperparameter tuning

⇒ no guarantee of convergence

evaluating optimal predictions



⇒ requires

≈ same time as single training step of neural network with minimal size

⇒ convergence guarantee *by construction*

Li *et al.*, Adv. Neural Inf. Process. Syst. 31 (2018)

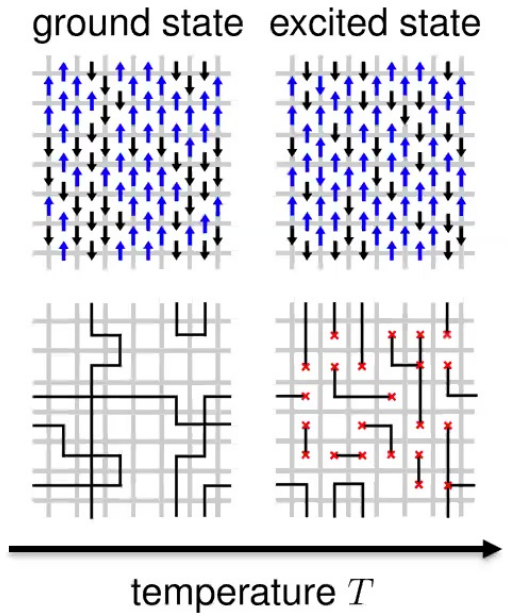
Application to classical equilibrium systems

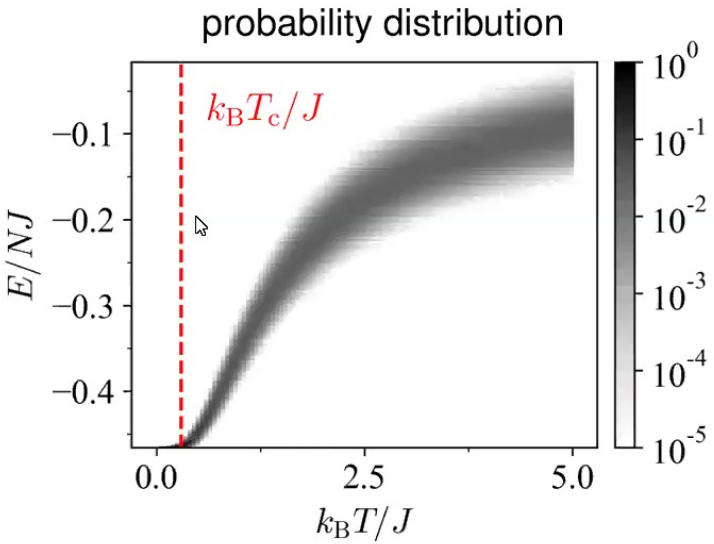
- for systems governed by Boltzmann distribution $\mathrm{P}_k(\boldsymbol{\sigma}) = \dfrac{e^{-H(\boldsymbol{\sigma})/k_\mathrm{B}T_k}}{Z_k}$
  the energy $E = H(\boldsymbol{\sigma})$ is a sufficient statistic

  $\Rightarrow$ switch from raw configurations to energy as input
  yields optimal lossless compression of state space

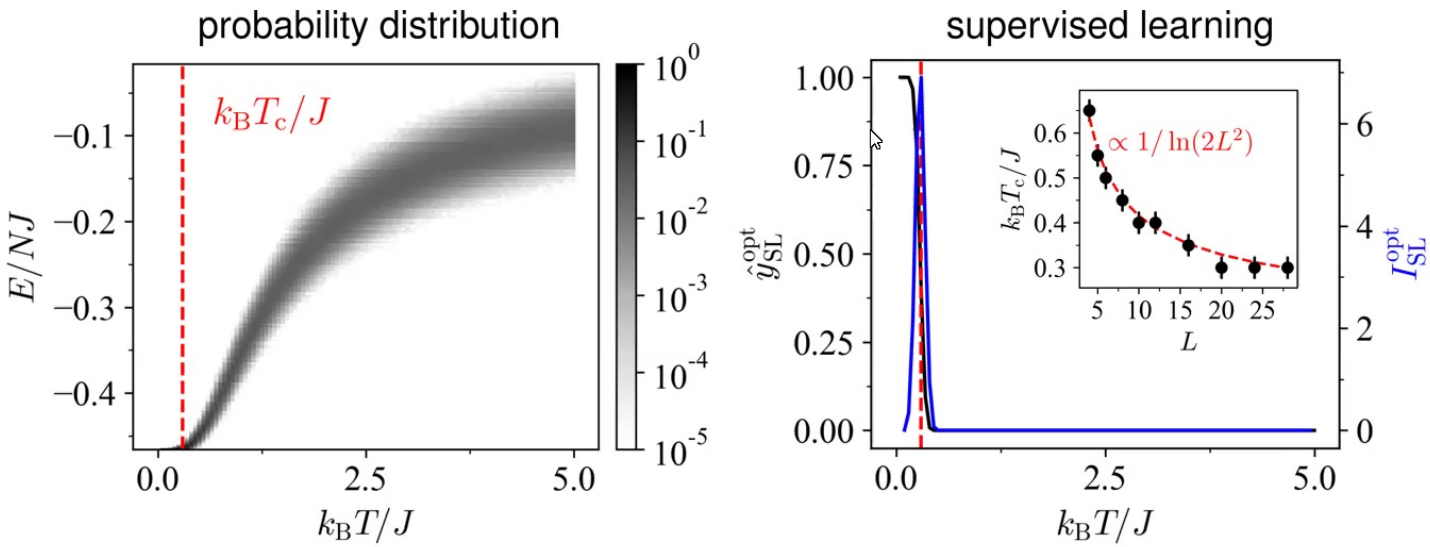
  $\Rightarrow$ for any predictive model: energy is single relevant feature

Example: Ising gauge theory (L=28)





Example: Ising gauge theory (L=28)



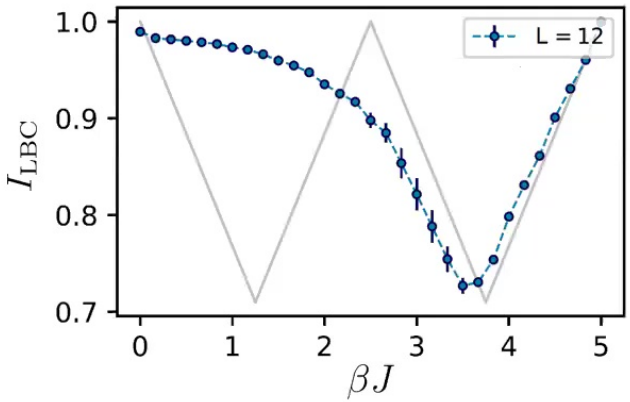
- can derive: $\hat{y}_{\mathrm{SL}}^{\mathrm{opt}}(p_k) \propto \mathrm{P}_k(E_{\mathrm{gs}})$

  $\Rightarrow$ supervised learning tracks the relevant physical quantity

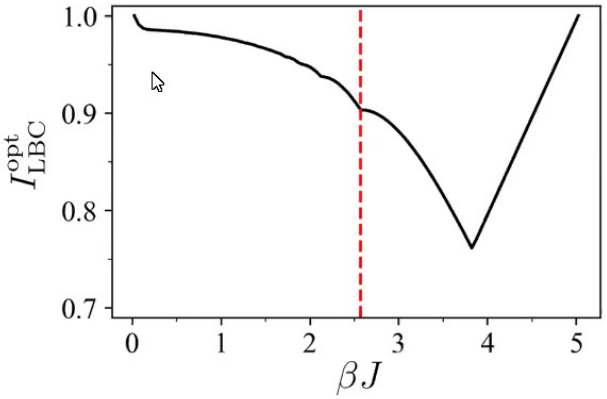
Carrasquilla and Melko, Nat. Phys. **13** 431-434 (2017)

Example: Ising gauge theory (L=12)

training neural networks



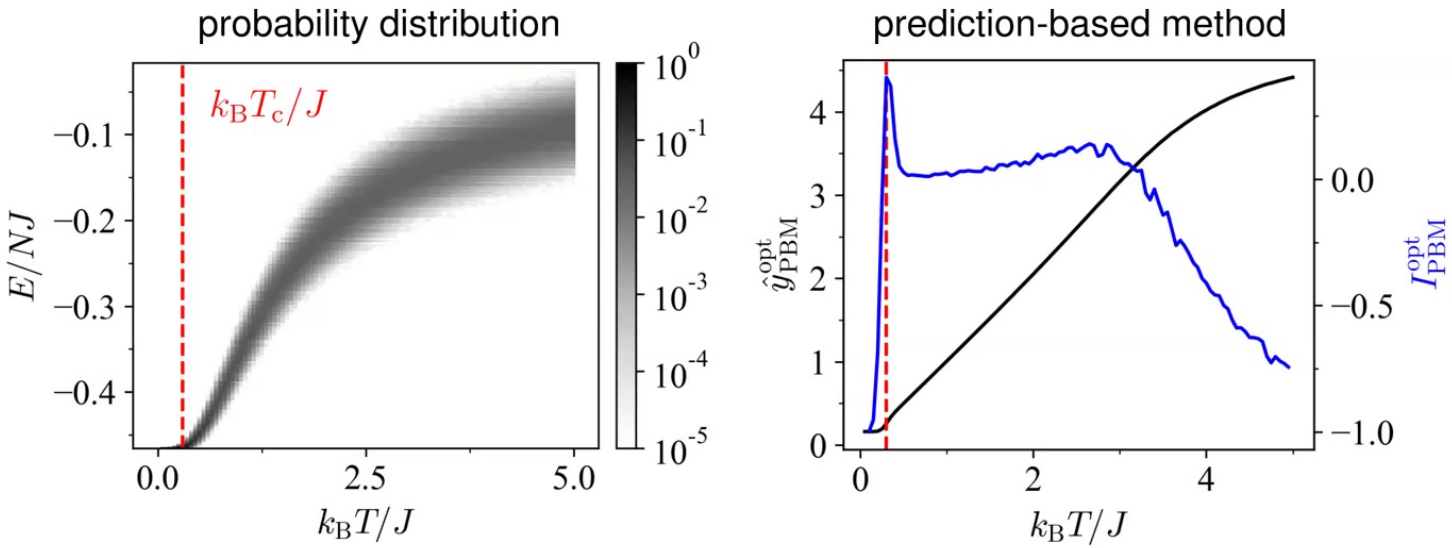
evaluating optimal predictions



$\Rightarrow$ learning by confusion fails in this setting

Greplova *et al.*, New J. Phys. **22** 045003 (2020)

Example: Ising gauge theory (L=28)

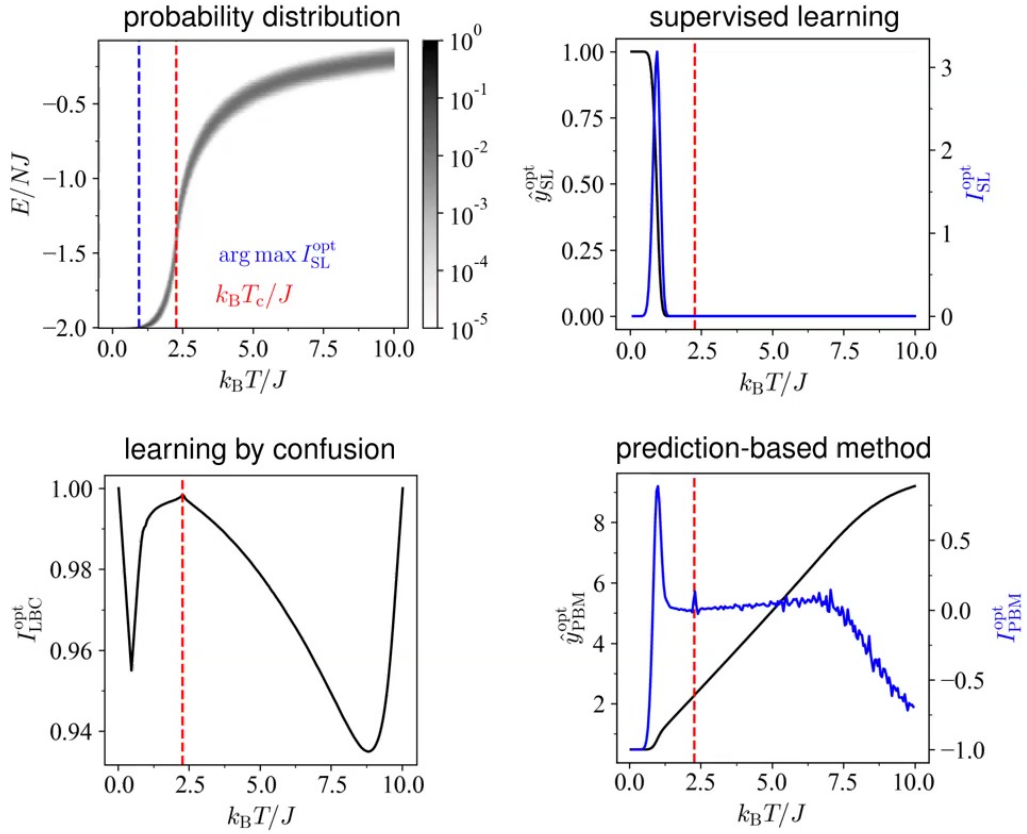


- for Boltzmann distribution: $\hat{y}^{\mathrm{opt}}_{\mathrm{PBM}} \equiv \hat{y}_{\mathrm{DOS}}$

  $\Rightarrow$ optimal predictive model is equivalent to density-of-states (DOS) model

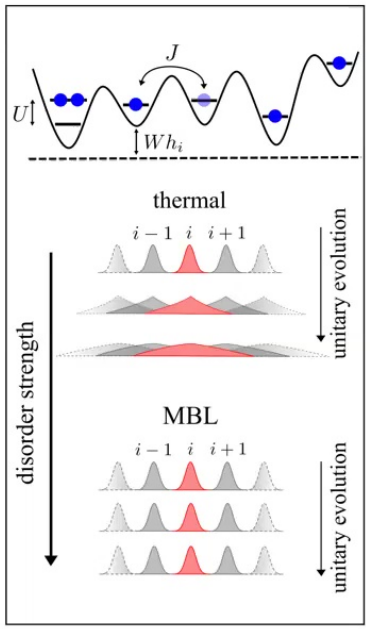
Greplova *et al.*, New J. Phys. **22** 045003 (2020)

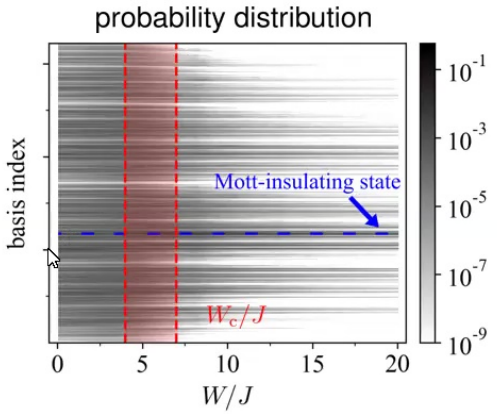
Example: Ising model (L=60)

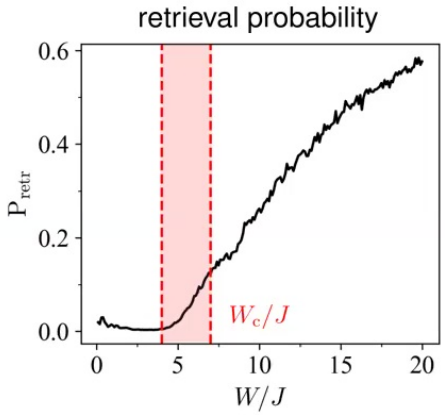


Schäfer and Lörch, Phys. Rev. E **99** 062107 (2019)

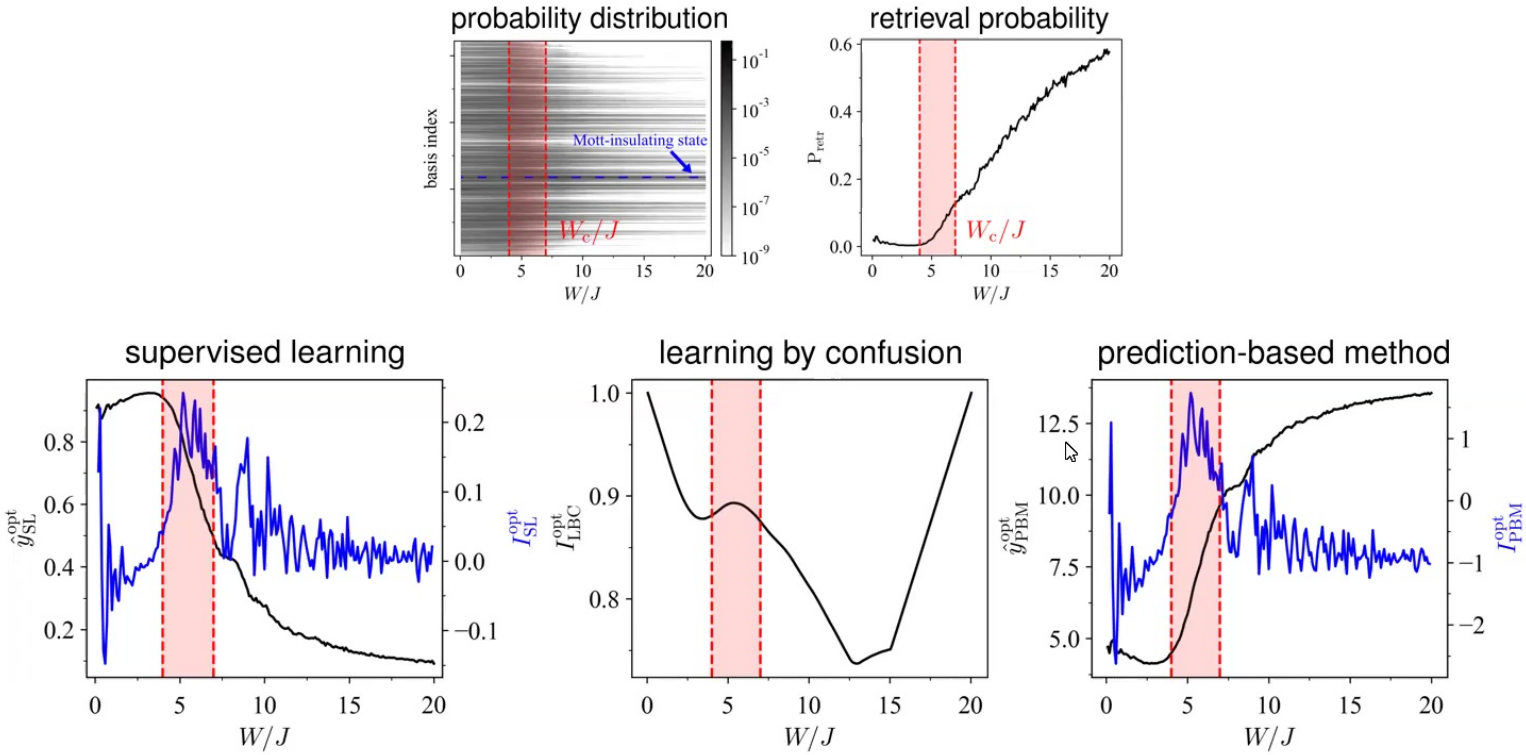
Example: Bose-Hubbard chain (L=8)







Example: Bose-Hubbard chain (L=8)



$\Rightarrow$ approach is *in principle* applicable to experimental measurement data

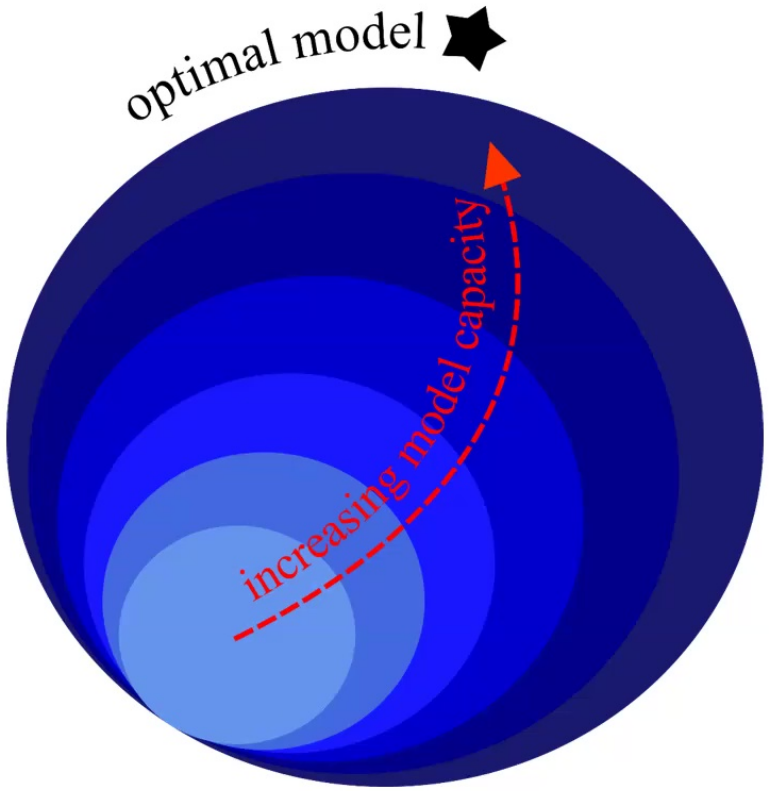
---

Bohrdt *et al.*, Phys. Rev. Lett. **127** 150504 (2021)

...and many more

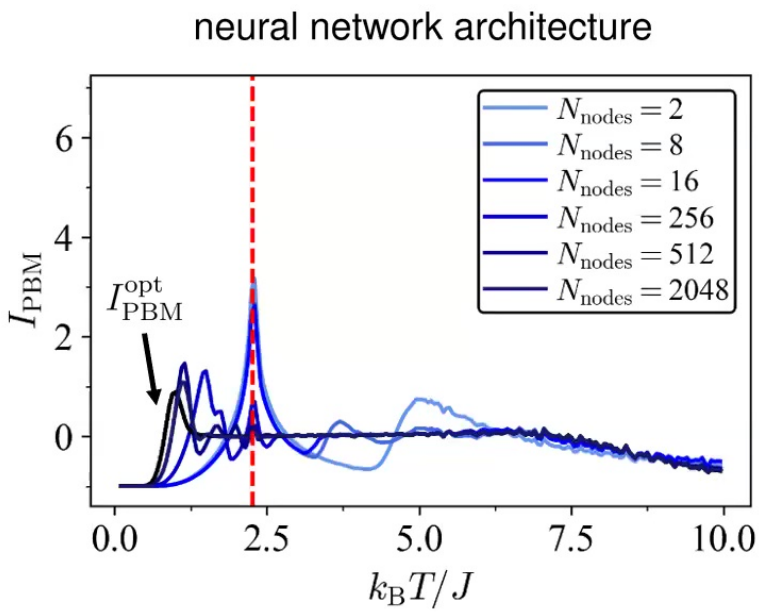
- Berezinskii-Kosterlitz-Thouless transition in XY model
- first-order phase transition in XXZ chain
- Mott-insulator to superfluid transition in Bose-Hubbard model
- topological phase transition in Kitaev chain

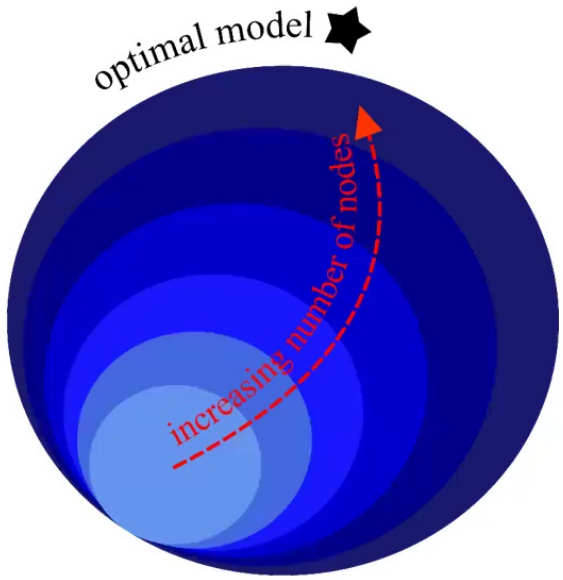
Controlling model capacity



Controlling model capacity

Example: Prediction-based method for Ising model ($L = 60$)

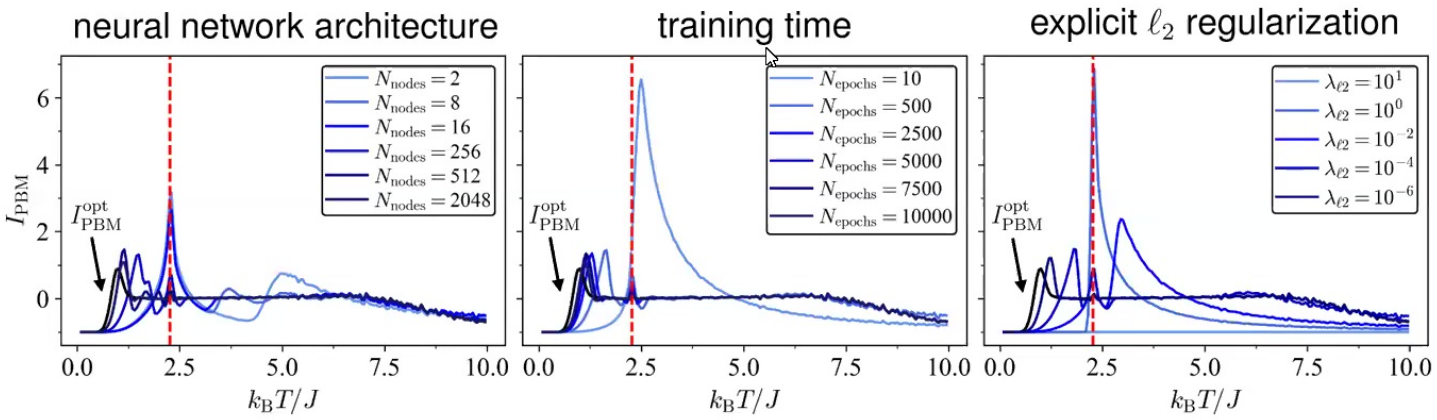




Controlling model capacity

N_{nodes} number of nodes
N_{epochs} number of epochs
ℓ_2 regularization: $\mathcal{L} \to \mathcal{L} + \lambda_{\ell 2} \sum_i \theta_i^2$

Example: Prediction-based method for Ising model ($L = 60$)



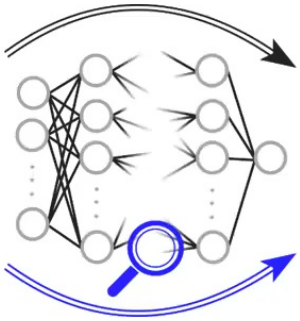
$\Rightarrow$ optimal predictions can be recovered in practice in high-capacity limit

$\Rightarrow$ failure of optimal model signals fundamental mismatch between task and goal

Summary

high model capacity
with
low interpretability and high computational cost

neural networks



non-parametric predictive models

high model capacity
with
high interpretability at low computational cost

Outlook

- apply similar analyses to other machine learning methods and classification tasks in condensed matter physics

- improve numerical routine using methods for density estimation

  $\Rightarrow$ compatible with autoregressive networks and matrix product states

- compare indicators to other choices of statistical distances

  $\Rightarrow$ establish relation to Fisher information