Interpreting Neural Networks in the Context of Physical Phase Diagrams

Sebastian J. Wetzel

Perimeter Institute Quantum Intelligence Lab (PIQUIL), Waterloo
Institute for Theoretical Physics, University of Heidelberg

PERIMETER INSTITUTE

STRUCTURES CLUSTER OF EXCELLENCE

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Overview

Physics:

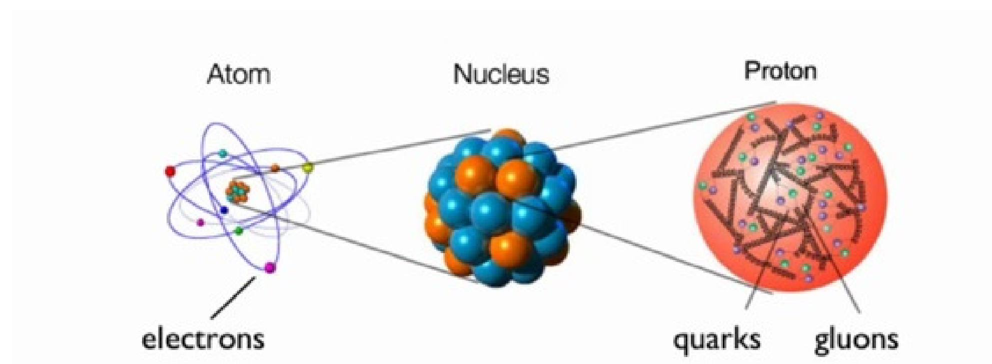
- Phase Diagrams
- Standard Model

Machine Learning

- Applying Neural Networks to Discover Phase Transitions
- Interpretation of Neural Networks

Phase Diagrams

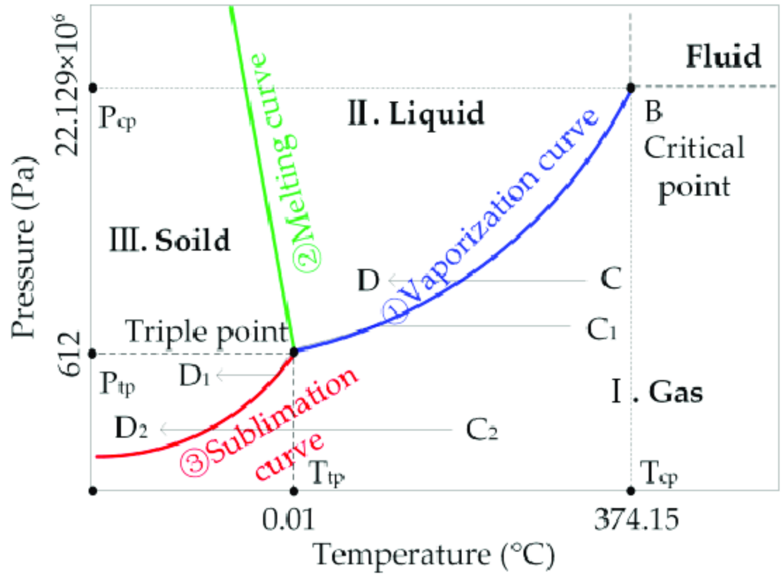
Water



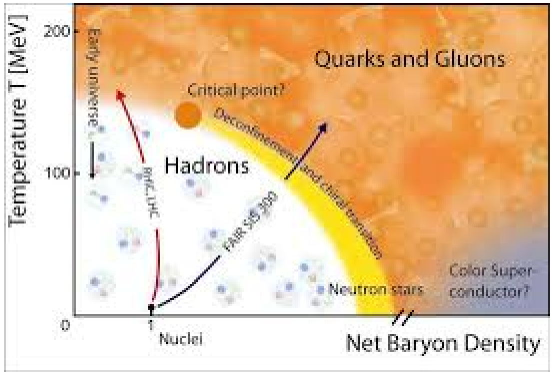
Atomic core

Magnet

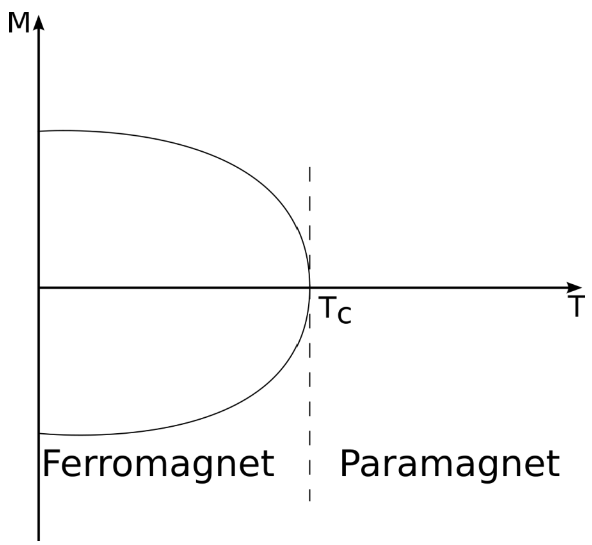
Phase Diagrams



Water



Atomic core



Magnet

Motivation

Theoretical Physics Goal:

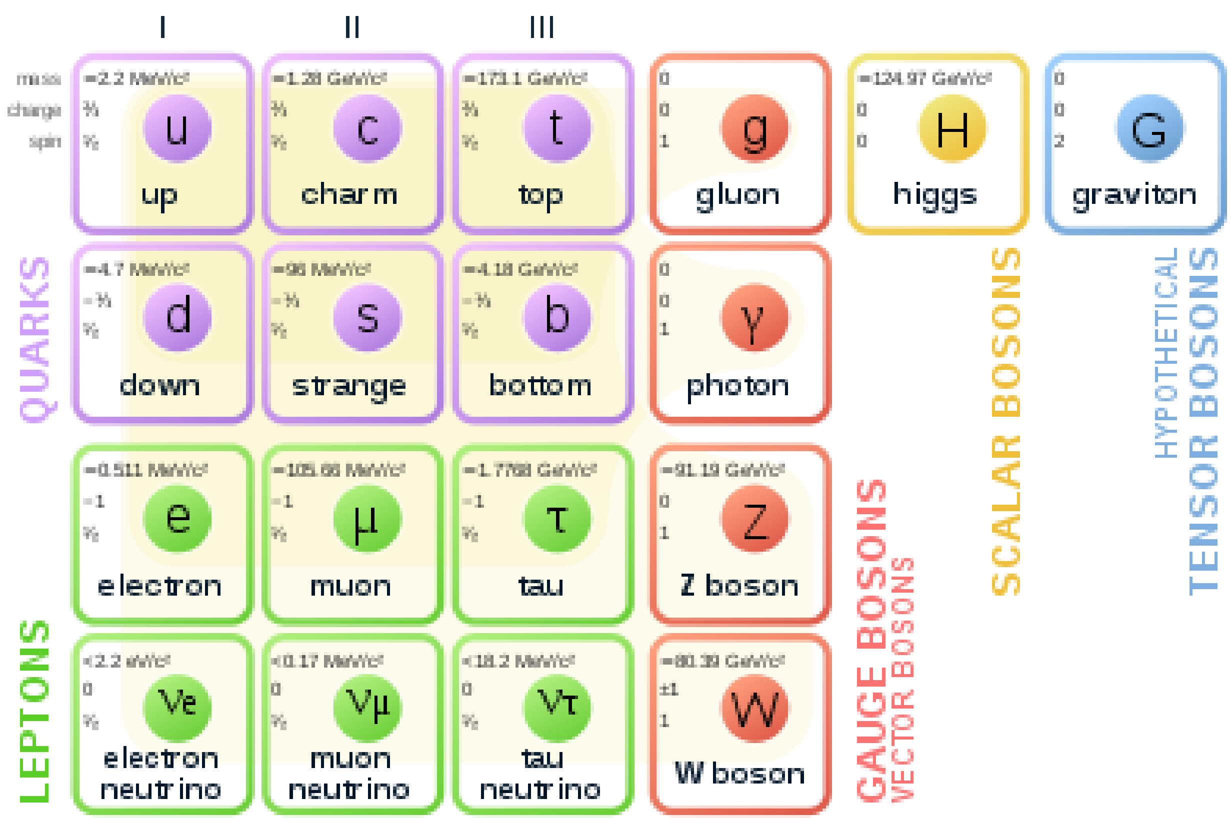
Determine macroscopic phase diagrams from a microscopic description

- Determine the existence of phases
- Pin down the phase transition
- Find the dominant characteristics of phases

Standard Model of Particle Physics + Gravity

- Experimentally verified paricle content + hypothetical graviton

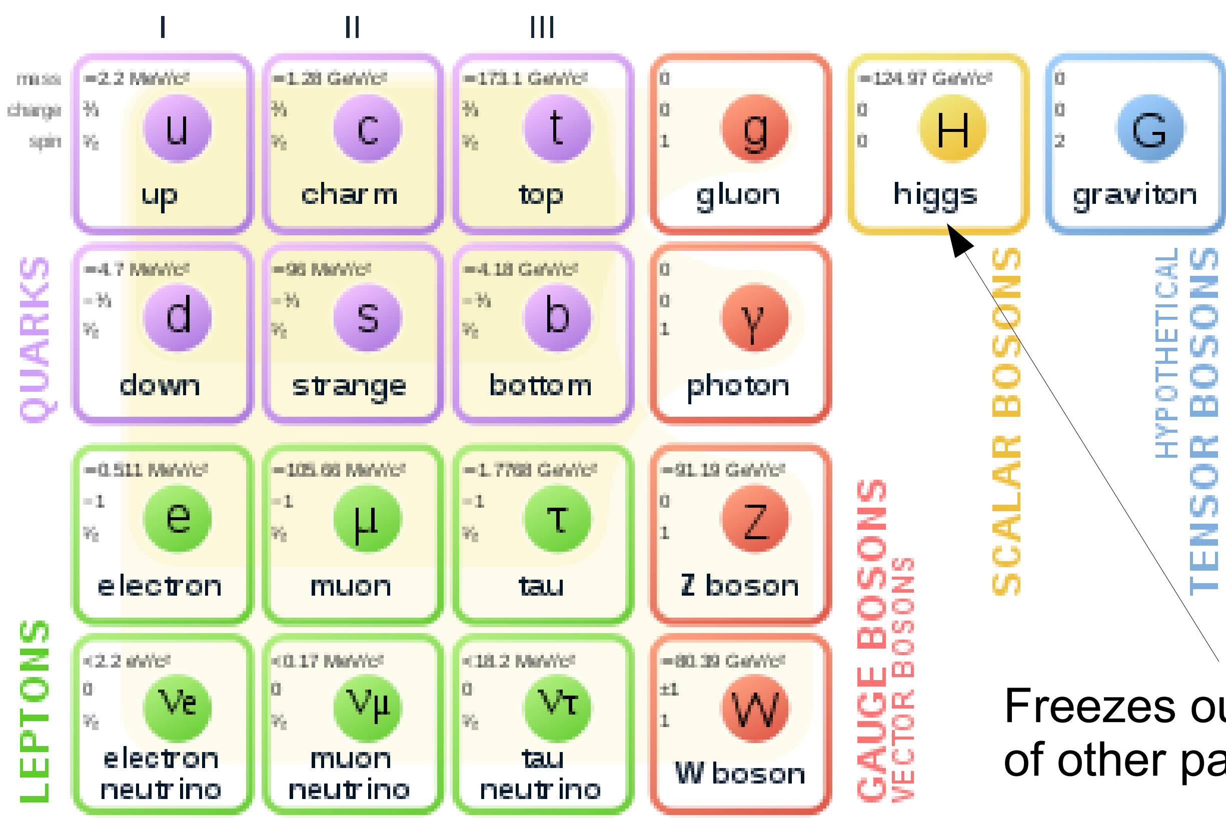
Standard Model of Elementary Particles + Gravity



Standard Model of Particle Physics + Gravity

- Experimentally verified paricle content + hypothetical graviton

Standard Model of Elementary Particles + Gravity

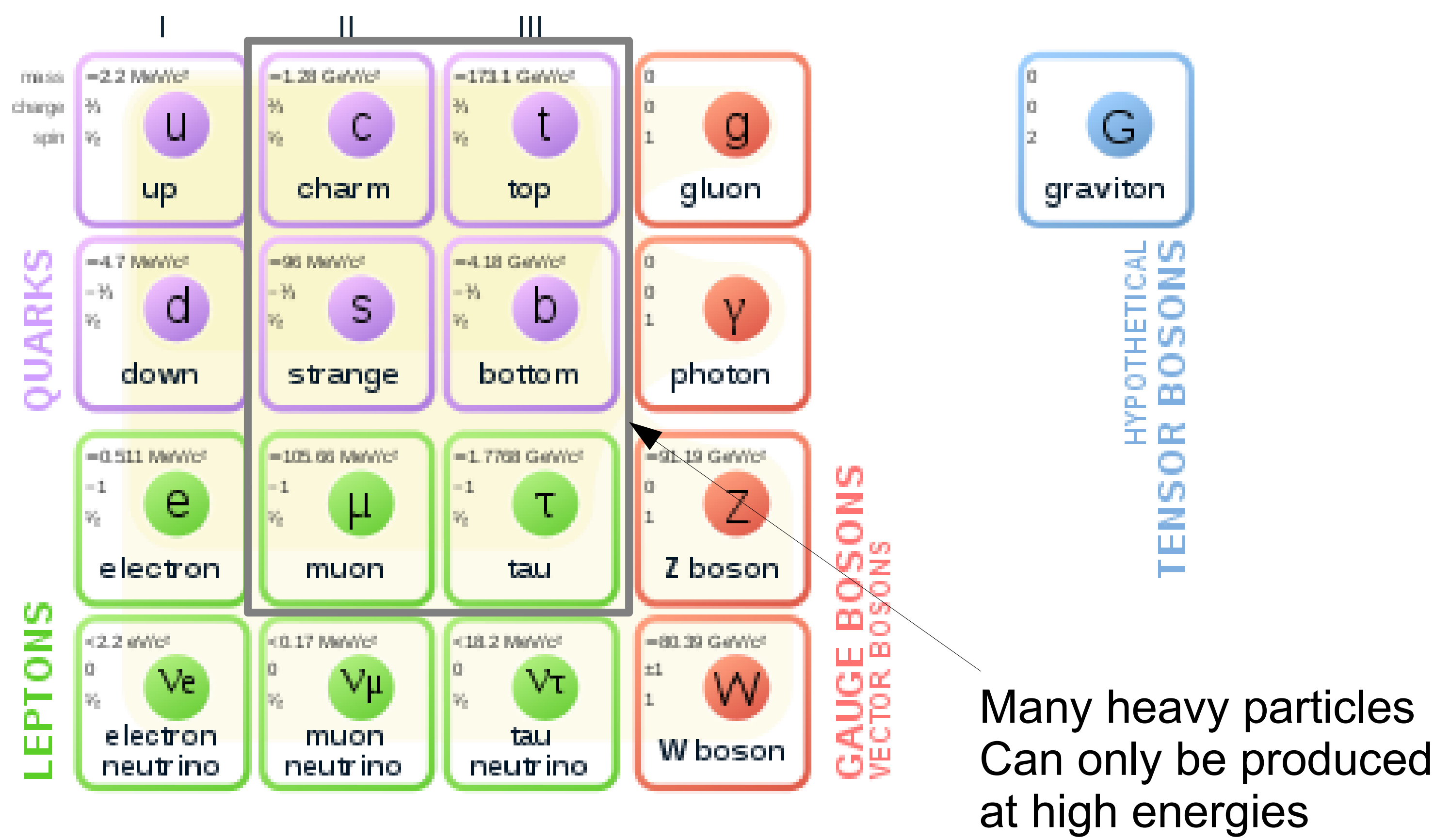


Freezes out to masses of other particles

Standard Model of Particle Physics + Gravity

- Experimentally verified paricle content + hypothetical graviton

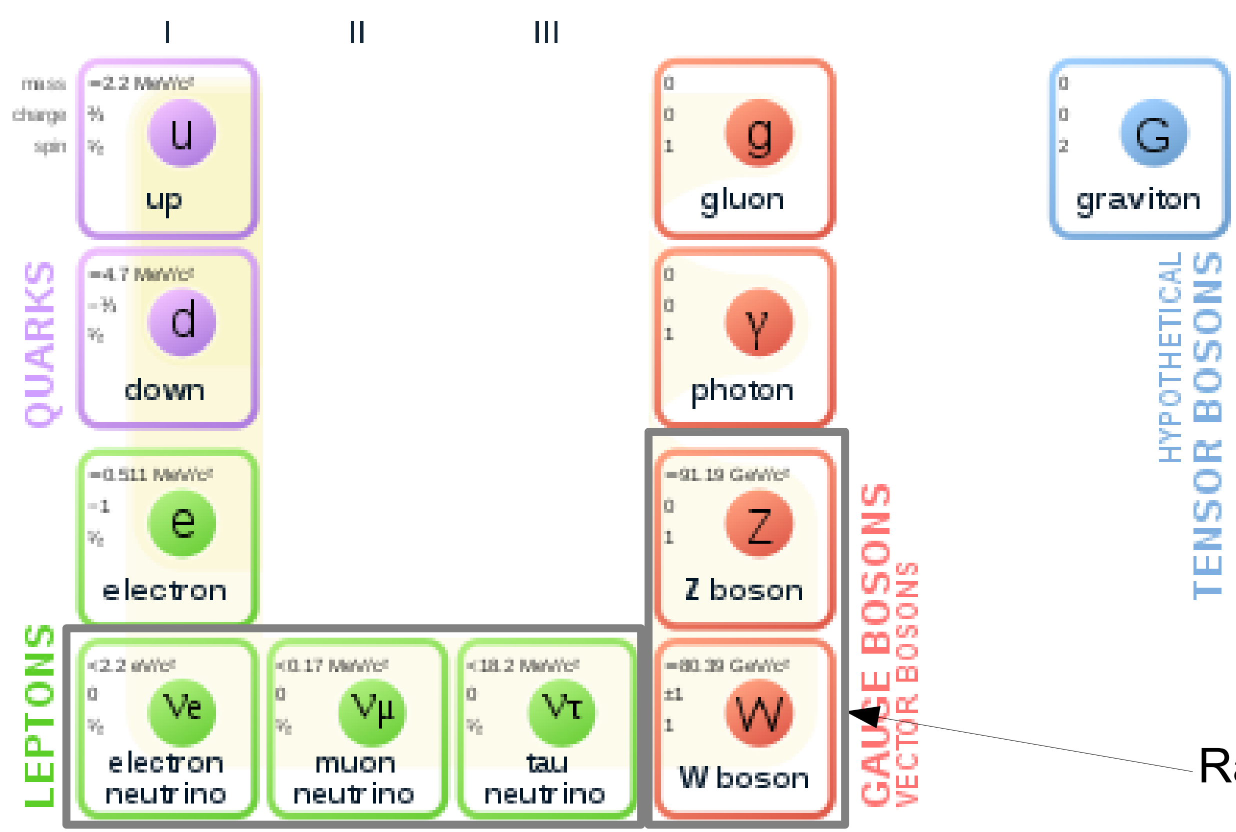
Standard Model of Elementary Particles + Gravity



Standard Model of Particle Physics + Gravity

- Experimentally verified paricle content + hypothetical graviton

Standard Model of Elementary Particles + Gravity

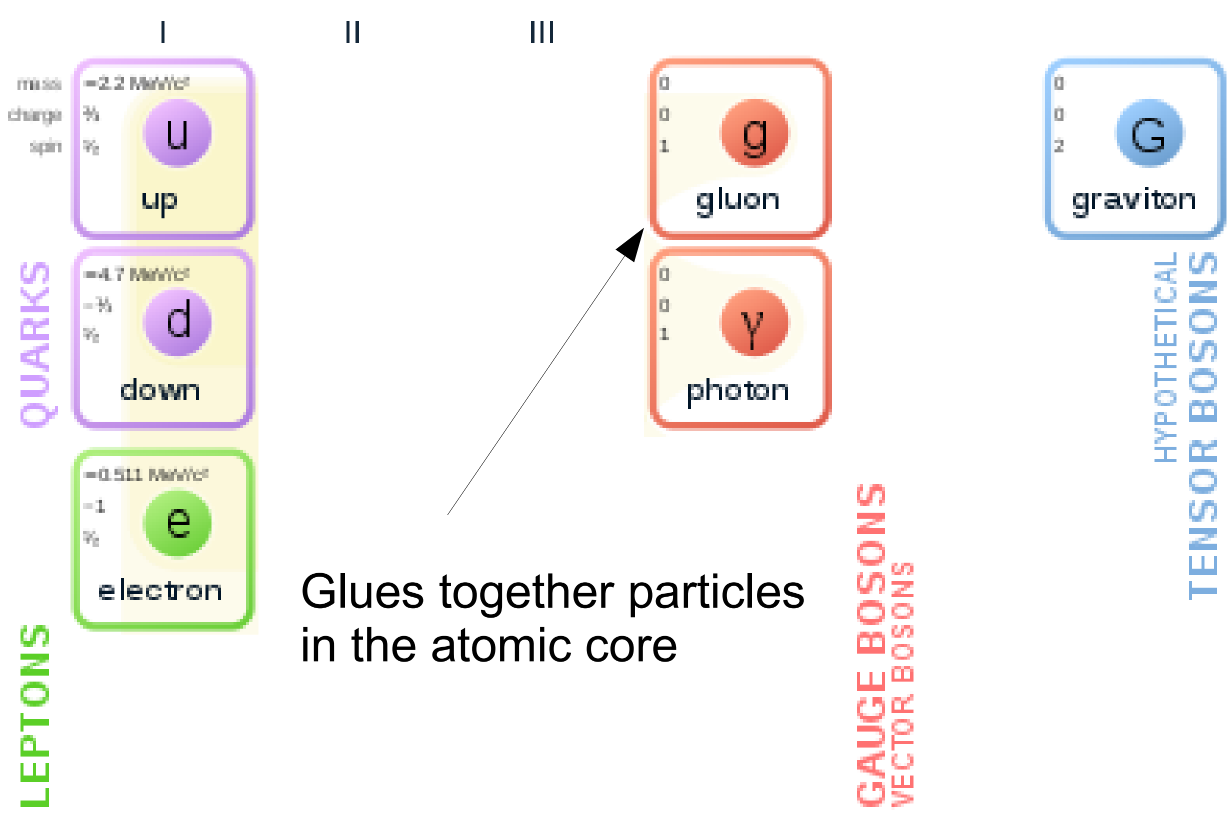


Radioactive Decay

Standard Model of Particle Physics + Gravity

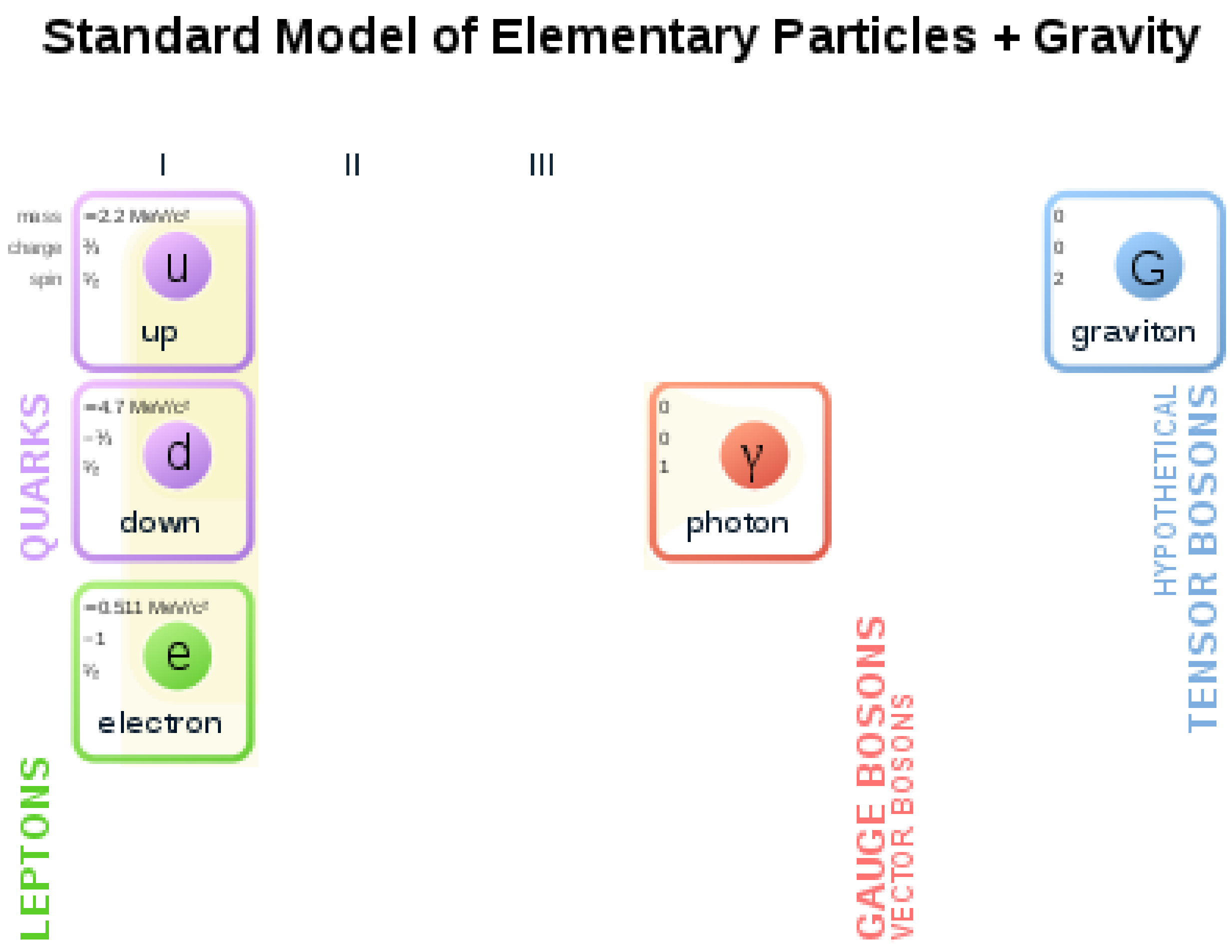
- Experimentally verified paricle content + hypothetical graviton

Standard Model of Elementary Particles + Gravity



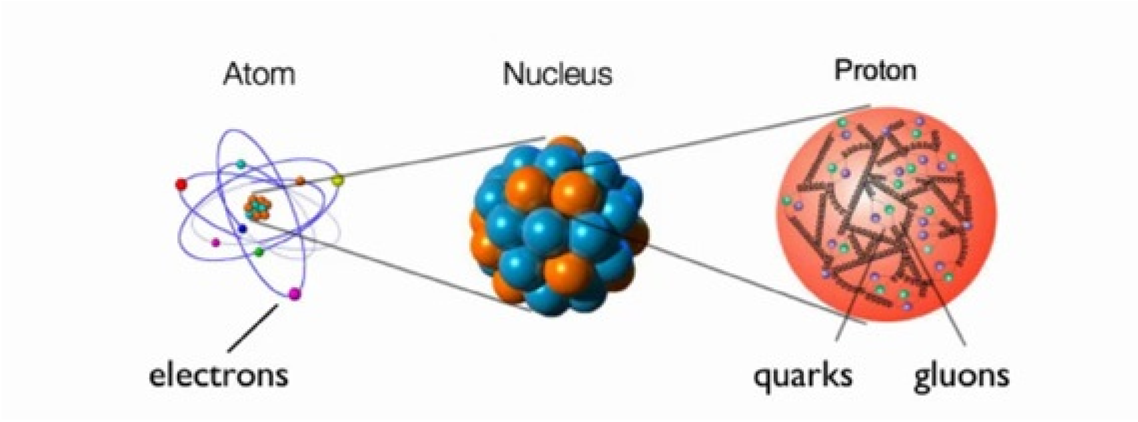
Standard Model of Particle Physics + Gravity

- Experimentally verified paricle content + hypothetical graviton



Standard Model of Particle Physics + Gravity

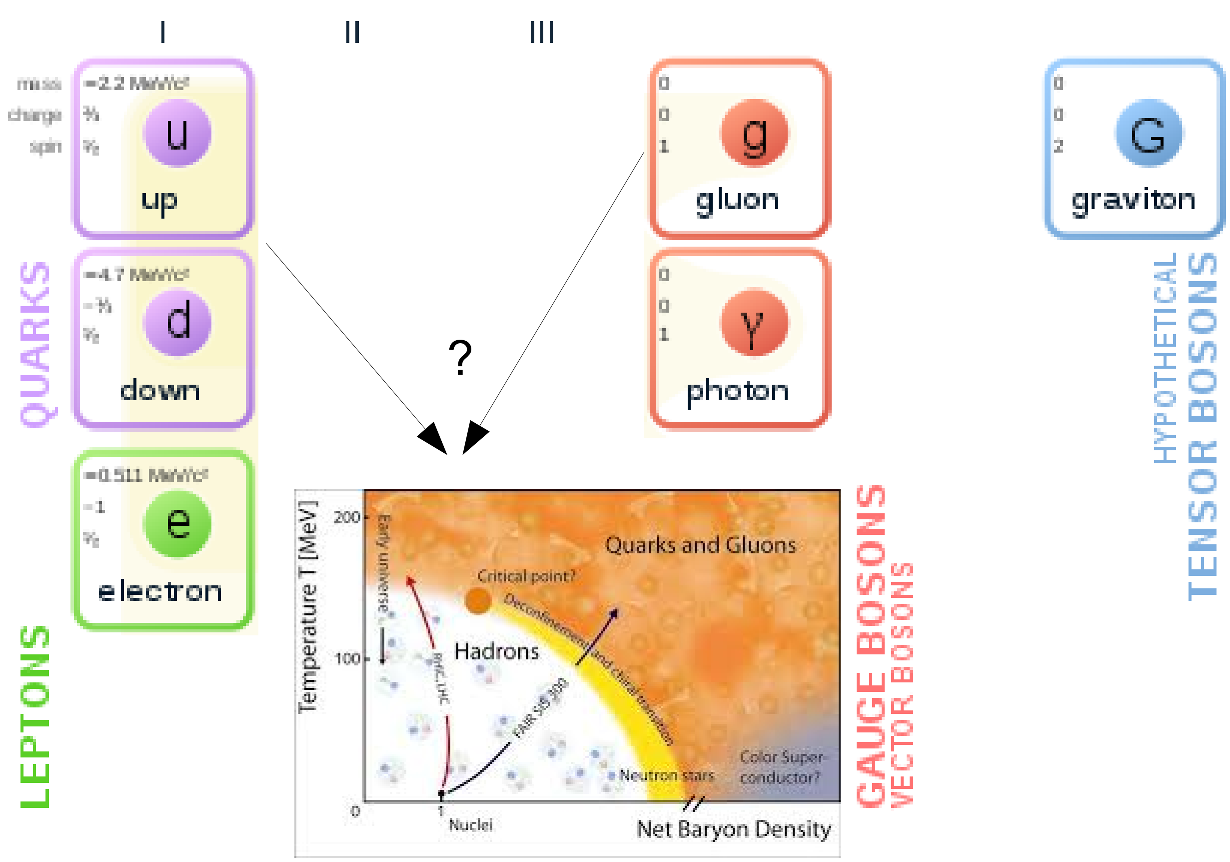
- We can only see part of the Standard Model, without
  - Heavy particles
  - Frozen Particles
  - Short Ranged Force Particles
- Visible Matter consists of up- / down- quarks and electrons



Standard Model of Particle Physics + Gravity

- Ultimate Goal of the Talk: NN reveals nature of confinement PT

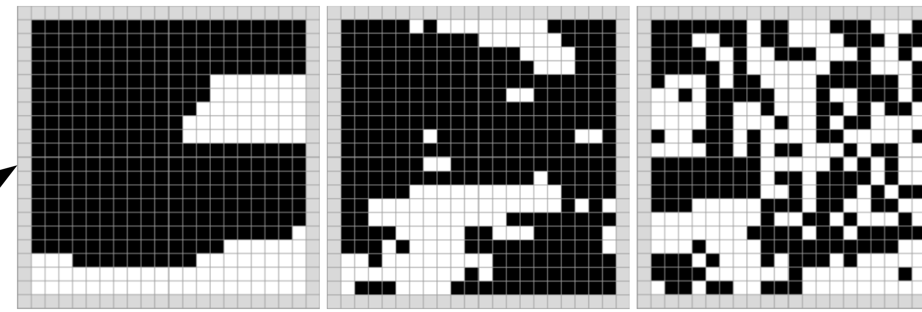
Standard Model of Elementary Particles + Gravity

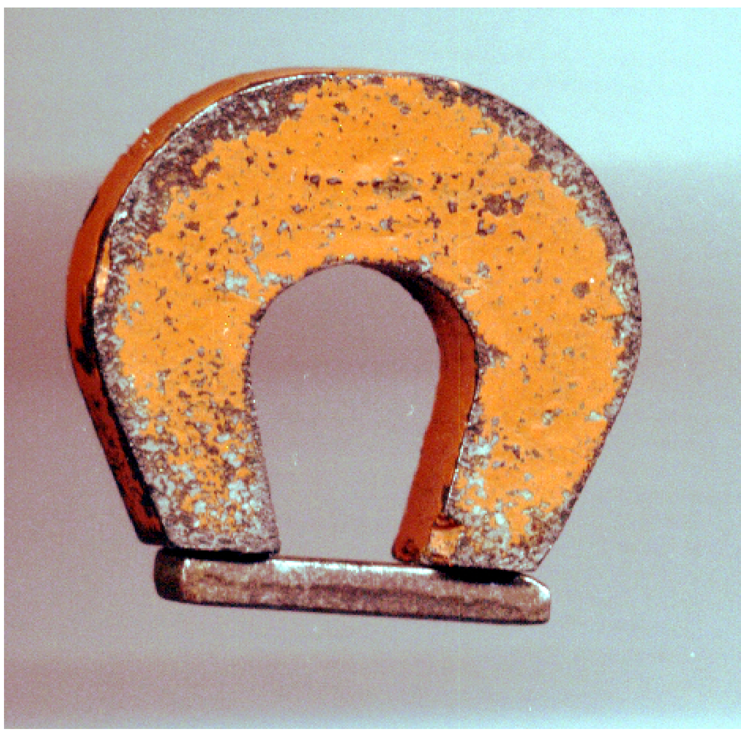


Invitation: Phase transitions from microscopic physics

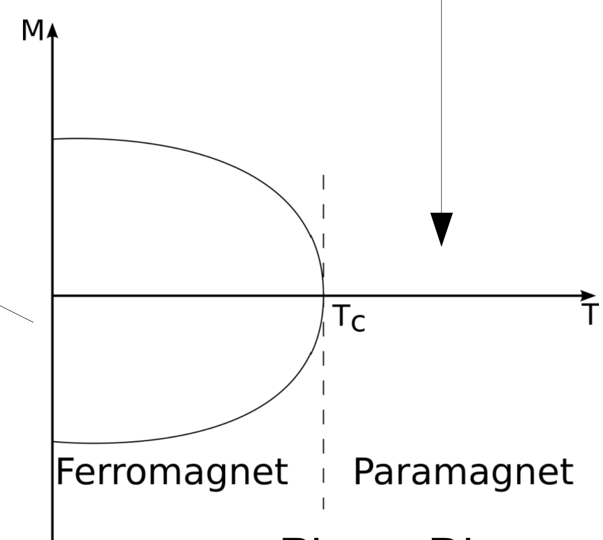
For simplicity: start with magnets

Microscopic description





Magnet



Phase Diagram

Invitation: Phase transitions from microscopic physics

Hamiltonian

$$H(S) = -J \sum_{<ij>_{NN}} s_i s_j$$

(Ising Model)

Physical Deduction

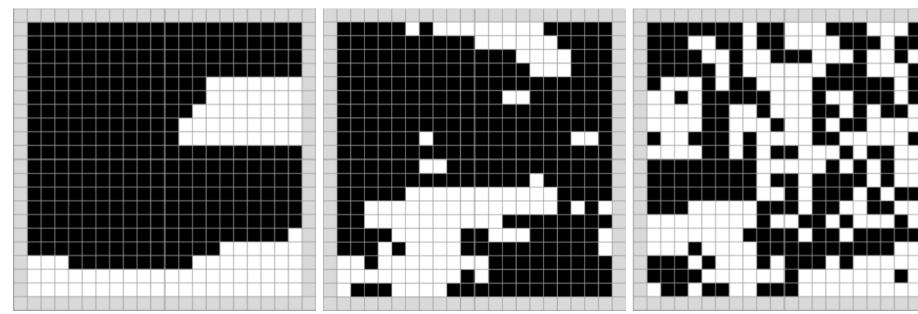
Computer

Order Parameter

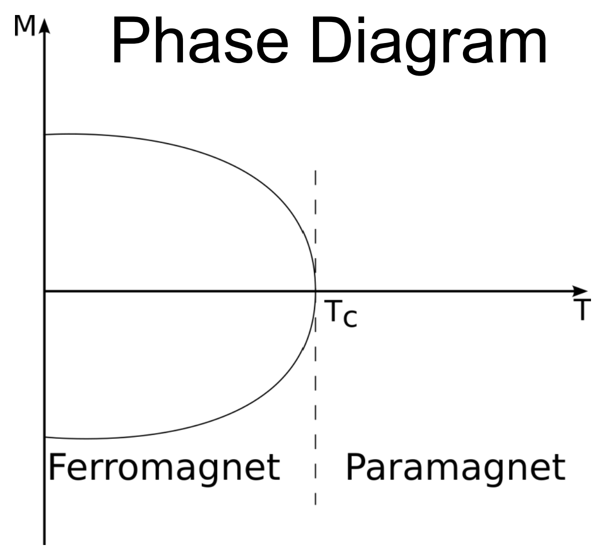
$$M(S) = \frac{1}{N} \sum_i s_i$$

$$\bar{M}(T) = \frac{1}{Z} \sum_{S \in \Lambda} |M(S)| \exp(-H(S)/T)$$

Monte-Carlo-Simulation



Phase Diagram



Invitation: Phase transitions from microscopic physics

Hamiltonian

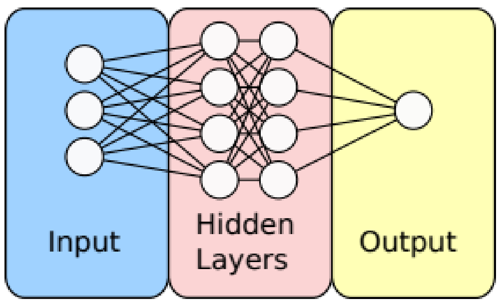
$$H(S) = -J \sum_{<ij>_{NN}} s_i s_j$$

(Ising Model)

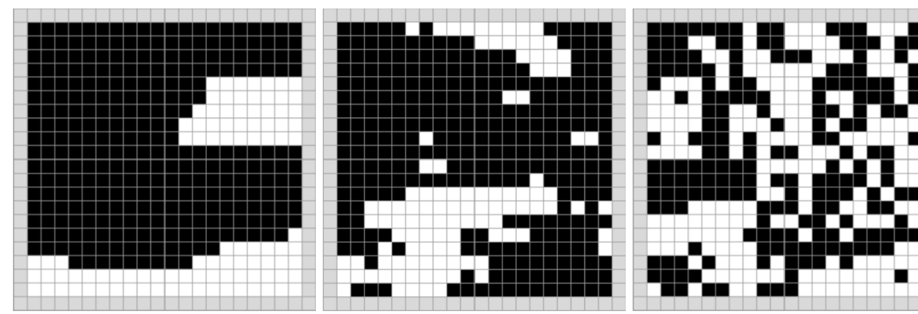
Computer

Computer

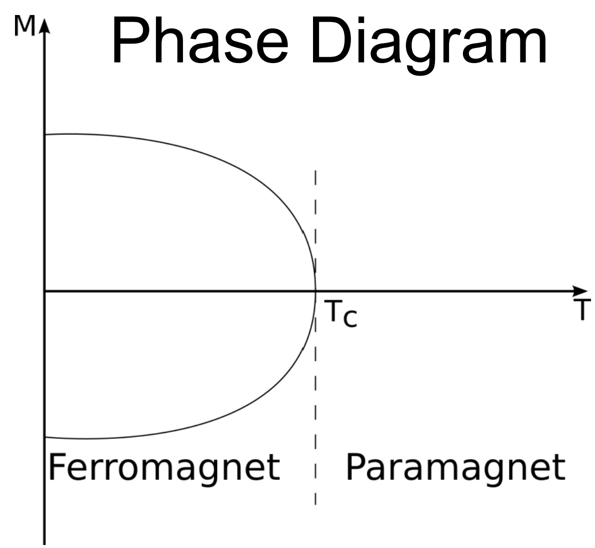
Neural Network ???



Monte-Carlo-Simulation

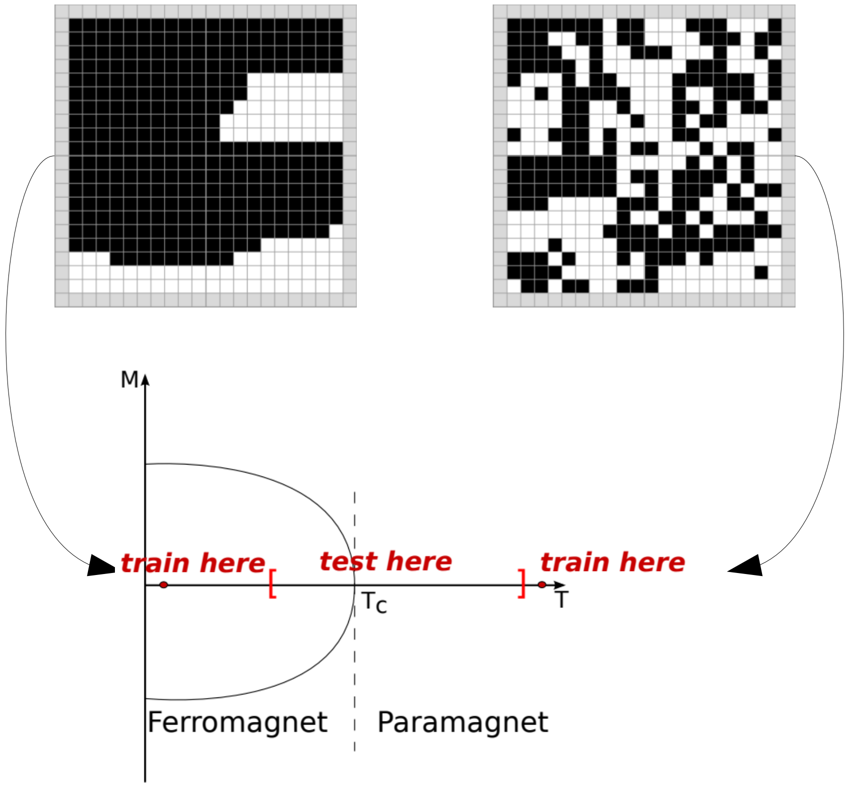


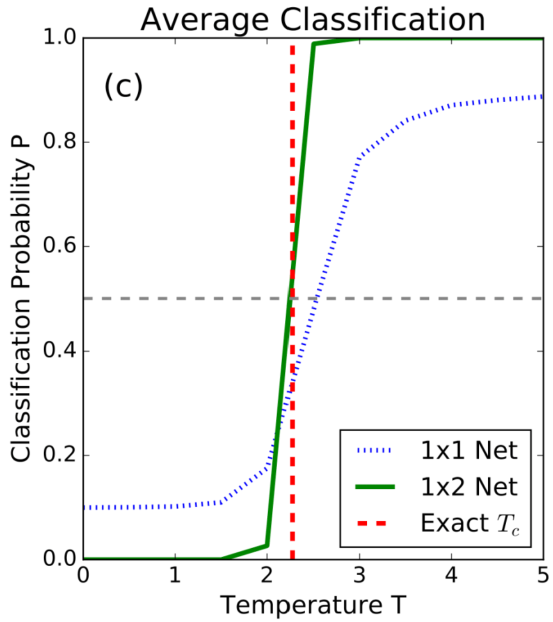
Phase Diagram



Supervised Learning
2d Ising Model

- Data: Monte Carlo samples
- Training at well known points in phase diagram
- Labels: Phase
- Testing in interval containing phase transition
- Estimate within 1% of exact value $T_c = \frac{2}{\ln(1+\sqrt{2})}$





Carrasquila, Melko, Nature 2017

Machine Learning Phases of Matter

- Starting in 2016: Rush to calculate physical phase diagrams using Neural Networks

nature physics | LETTERS

Machine learning phases of matter

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

Condensed-matter physics is the study of the collective behaviour of infinitely complex assemblies of electrons, nuclei, magnetic moments, atoms or qubits[1]. This complexity is reflected in the size of the state space, which grows exponentially with the number of particles, reminiscent of the 'curse of dimensionality' commonly encountered in machine learning[2]. Despite this curse, the machine learning community has developed techniques with remarkable abilities to recognize, classify, and characterize complex sets of data. Here, we show that modern machine learning architectures, such as fully connected and convolutional neural networks[3], can identify phases and phase transitions in a variety of condensed-matter Hamiltonians. Readily programmable through modern software libraries[4,5], neural networks can be trained to detect multiple types of order parameter, as well as highly non-trivial states with no conventional order, directly from raw state configurations sampled with Monte Carlo[6,7].

Learning phase transitions by confusion

Evert P.L. van Nieuwenburg*, Ye-Hua Liu, and Sebastian D. Huber
Institute for Theoretical Physics, ETH Zurich, 8093 Zürich, Switzerland

Classifying phases of matter is a central problem in physics. For quantum mechanical systems, this task can be daunting owing to the exponentially large Hilbert space. Thanks to the available computing power and access to ever larger data sets, classification problems are now routinely solved using machine learning techniques. Here, we propose to use a neural network based approach to find phase transitions depending on the performance of the neural network after training it with deliberately incorrectly labelled data. We demonstrate the success of this method on the topological phase transition in the Kitaev chain, the thermal phase transition in the classical Ising model, and the many-body-localization transition in a disordered quantum spin chain. Our method does not

Machine learning quantum phases of matter beyond the fermion sign problem

Peter Broecker,[1] Juan Carrasquilla,[2] Roger G. Melko,[2,3] and Simon Trebst[1]
[1]*Institute for Theoretical Physics, University of Cologne, 50937 Cologne, Germany*
[2]*Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada*
[3]*Department of Physics and Astronomy, University of Waterloo, Ontario, N2L 3G1, Canada*
(Dated: August 30, 2016)

State-of-the-art machine learning techniques promise to become a powerful tool in statistical mechanics via their capacity to distinguish different phases of matter in an automated way. Here we demonstrate that convolutional neural networks (CNN) can be optimized for quantum many-fermion systems such that they correctly identify and locate quantum phase transitions in such systems. Using auxiliary-field quantum Monte Carlo (QMC) simulations to sample the many-fermion system, we show that the Green's function (but not the auxiliary field) holds sufficient information to allow for the distinction of different fermionic phases via a CNN. We demonstrate that this QMC + machine learning approach works even for systems exhibiting a severe fermion sign problem where conventional approaches to extract information from the Green's function, e.g. in the form of equal-time correlation functions, fail. We expect that this capacity of hierarchical machine learning techniques to circumvent the fermion sign p... problems in statistical physics.

Machine learning vortices at the Kosterlitz-Thouless transition

Matthew J. S. Beach,* Anna Golubeva, and Roger G. Melko
Department of Physics and Astronomy, University of Waterloo, Waterloo N2L 3G1, Canada and Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada
(Dated: October 30, 2017)

Efficient and automated classification of phases from minimally processed data is one goal of machine learning in condensed matter and statistical physics. Supervised algorithms trained on raw samples of microstates can successfully detect conventional phase transitions via learning a bulk feature such as an order parameter. In this paper, we investigate whether neural networks can learn to classify phases based on topological defects. We address this question on the two-dimensional classical XY model which exhibits a Kosterlitz-Thouless transition. We find significant feature engineering of the raw spin states is required to convincingly claim that features of the vortex configurations are responsible for learning the transition temperature. We further show a single-layer network does not correctly classify the phases of the XY model, while a convolutional network easily performs classification by learning the global magnetization. Finally, we design a deep network capable of learning vortices without feature engineering. We demonstrate the detection of ... best classification accuracy, especially for lattices of less ... systems, it remains a difficult task to learn vortices.

Machine Learning of Explicit Order Parameters: From the Ising Model to SU(2) Lattice Gauge Theory

Sebastian J. Wetzel[1] and Manuel Scherzer[1]
[1]*Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany*

We present a procedure for reconstructing the decision function of an artificial neural network as a simple function of the input, provided the decision function is sufficiently symmetric. In this case one can easily deduce the quantity by which the neural network classifies the input. The procedure is embedded into a pipeline of machine learning algorithms able to detect the existence of different phases of matter, to determine the position of phase transitions and to find explicit expressions of the physical quantities by which the algorithm distinguishes between phases. We assume no prior knowledge about the Hamiltonian or the order parameters except Monte Carlo-sampled configurations. The method is applied to the Ising Model and SU(2) lattice gauge theory. In both systems we deduce the explicit expressions of the known order parameters from the decision functions of the neural networks.

Machine Learning of Phase Diagrams Overview

		Pro +		Con -	
Supervised	Feed Forward Neural Network	Most powerful	Conv Layer Spatial Structure	Least Interpretable	*Carrasquila, Melko, Nature 2017*
Supervised	Support Vector Machine	Interpretability		Not suitable for large datasets	*Ponte, Melko, Phys Rev B 2017*
Supervised	Recurrent Neural Network	Dynamical Systems			*Nieuwenburg, Bairey, Refael, Phys Rev B 2018*
Unsupervised	Principal Component Analysis	Interpretability	Most easy to use		*Wang, Phys Rev B 2016*
Unsupervised	Autoencoder (Neural Network)		Conv Layer Spatial Structure		*Wetzel, Phys Rev E 2017*
Hybrid	Learning by Confusion				*Nieuwenburg, Liu, Huber, Nature 2017*

Machine Learning of Phase Diagrams Overview

New physics requires Powerful ML

Problem

	Method	Pro +		Con -	Reference
Supervised	Feed Forward Neural Network	Most powerful	Conv Layer Spatial Structure	Least Interpretable	*Carrasquila, Melko, Nature 2017*
Supervised	Support Vector Machine	Interpretability		Not suitable for large datasets	*Ponte, Melko, Phys Rev B 2017*
Supervised	Recurrent Neural Network	Dynamical Systems			*Nieuwenburg, Bairey, Refael, Phys Rev B 2018*
Unsupervised	Principal Component Analysis	Interpretability	Most easy to use		*Wang, Phys Rev B 2016*
Unsupervised	Autoencoder (Neural Network)		Conv Layer Spatial Structure		*Wetzel, Phys Rev E 2017*
Hybrid	Learning by Confusion				*Nieuwenburg, Liu, Huber, Nature 2017*

Notion of Interpretability

If the neural network bases its decision on <u>one</u> single quantity/obervable $Q(S)$

- The larger the observable, the higher the classification probability.
- If two inputs have the same value of the observable, they have the same classification probability.

The Neural network can be mapped via a bijective function to the observable

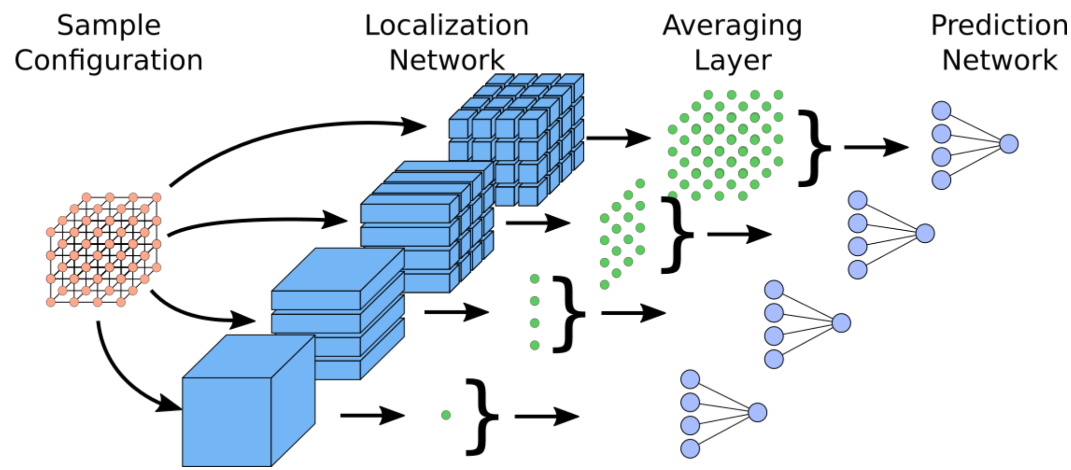
$$F(S) = f(Q(S))$$

Notion of Interpretability

- *Useful in the context of physics?*
  - In Physics often only very few quantities $Q(S)$ are characteristic features of phase transitions. (Renormalization Group: relevant parameters)
  - Physical Quantities are uniquely formulated by well defined formulas (in contrast to cars, faces ...)
  - Physical quantities are often highly symmetric: Rotation symmetry, translation symmetry

Interpretation of Neural Network

Interpretation Net:



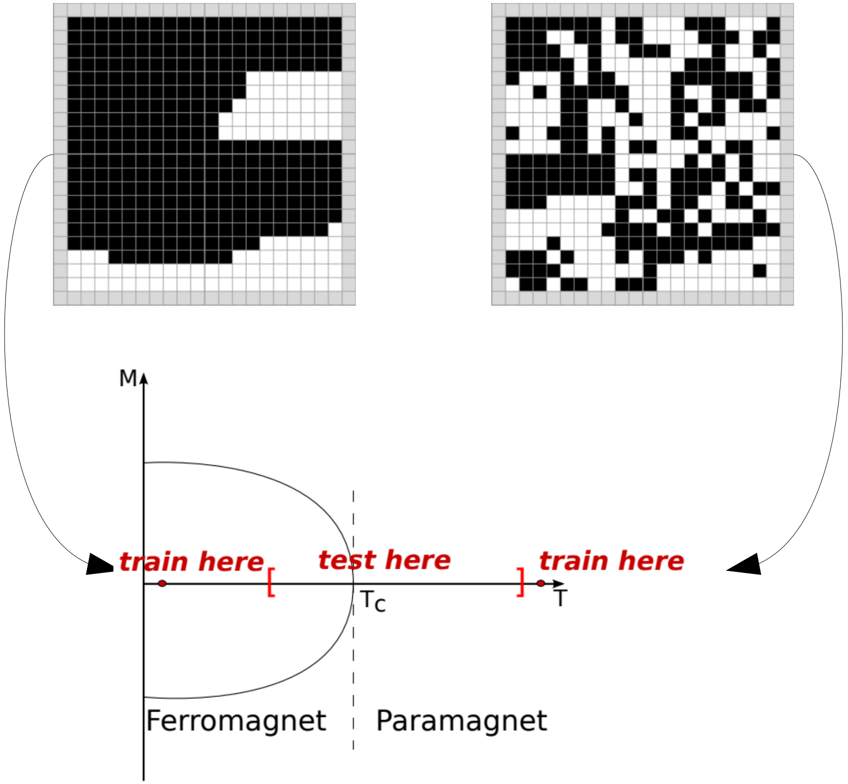
Wetzel, Scherzer, PRB 2017

- Interpretation Net interpolates between a general NN and a minimal optimal NN which has the same performance
- Interpretation by reducing the NN capacity in an ordered manner until one observes a performance drop
- Inspired by extensive physical quantities (averaging layer probes for translational invariance of the quantity $Q(S)$)

Interpretation of Neural Network
2d Ising Model

Starting Neural Network:

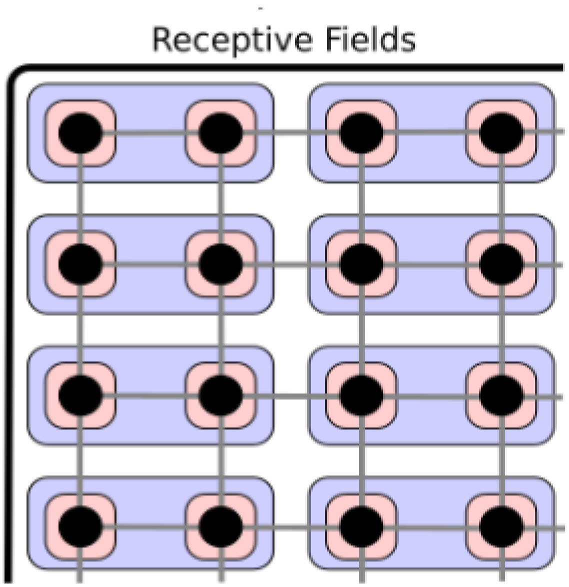
- Conv Net with full receptive field
- Training until converged
- Remember Loss value as measure of performance



Interpretation of Neural Network
2d Ising Model

Reinitialize the neural network with reduced receptive field sizes

- Train again until converged and compare the loss to the previous network
- Observe drop in performance from 1x2 to 1x1 and from 1x1 to baseline
- Dominant contributions must contain functions of spins and neighboring spins



Receptive Field Size	Train Loss	Validation Loss
28×28	$6.1588e-04$	0.0232
1×2	**1.2559e-04**	**1.2105e-07**
1×1	**0.2015**	**0.1886**
baseline	0.6931	0.6931

Interpretation of Neural Network
2d Ising Model

1st Network: 1x1 receptive field

- Express the full neural network in 1x1 form

$$F(S) = F(\frac{1}{N}\sum_i f(s_i)) = \text{sigmoid}(\xi(\frac{1}{N}\sum_i f(s_i)))$$

- Taylor expansion eliminates all higher order terms

$$f(s_i) = f_0 + f_1 \; s_i + f_2 \underbrace{s_i^2}_{1} + f_3 \underbrace{s_i^3}_{s_i} + ...$$

- Regression on a single variable yields explicit form

$$F(S) \approx \text{sigmoid}\left(w \left|\frac{1}{N}\sum_i s_i\right| + b\right)$$

- Where f_0, f_1 have been absorbed into weights and bias w, b

Interpretation of Neural Network
2d Ising Model

1st Network: 1x1 receptive field

- Express the full neural network in 1x1 form

$$F(S) = F(\frac{1}{N}\sum_i f(s_i)) = \mathrm{sigmoid}(\xi(\frac{1}{N}\sum_i f(s_i)))$$

- Taylor expansion eliminates all higher order terms

$$f(s_i) = f_0 + f_1 \; s_i + f_2 \; \underbrace{s_i^2}_{1} + f_3 \; \underbrace{s_i^3}_{s_i} + \ldots$$

- Regression on a single variable yields explicit form

$$F(S) \approx \mathrm{sigmoid}\left(w \left|\frac{1}{N}\sum_i s_i\right| + b\right)$$

Magnetization

- Where f_0, f_1 have been absorbed into weights and bias w, b

Interpretation of Neural Network
2d Ising Model

2nd Network: 1x2 receptive field

- Express the full neural network in 1x2 form

$$F(S) = F\left(\frac{1}{N}\sum_{<i,j>_T} f(s_i, s_j)\right)$$

- Taylor expansion contains only one addition to 1x1 case

$$f(s_i, s_j) = f_{0,0} + f_{1,0}\, s_i + f_{0,1}\, s_j + f_{2,0}\, s_i^2 + \boxed{f_{1,1}\, s_i\, s_j} + f_{0,2}\, s_j^2 + \ldots$$

- Regression yields explicit form

$$D(S) \approx \mathrm{sigmoid}\left(w\left(\frac{1}{N}\sum_{<i,j>_T} s_i s_j\right) + b\right)$$

Interpretation of Neural Network
2d Ising Model

2nd Network: 1x2 receptive field

- Express the full neural network in 1x2 form

$$F(S) = F\left(\frac{1}{N}\sum_{<i,j>_T} f(s_i, s_j)\right)$$

- Taylor expansion contains only one addition to 1x1 case

$$f(s_i, s_j) = f_{0,0} + f_{1,0}\, s_i + f_{0,1}\, s_j + f_{2,0}\, s_i^2 + \boxed{f_{1,1}\, s_i\, s_j} + f_{0,2}\, s_j^2 + \ldots$$

- Regression yields explicit form

$$D(S) \approx \text{sigmoid}\left(w\left(\frac{1}{N}\sum_{<i,j>_T} s_i s_j\right) + b\right)$$

Energy / 2

- Only half the energy since we dont sum over all neighbors

Interpretation of Neural Network
2d Ising Model

Receptive Field Size	Train Loss	Validation Loss
28×28	$6.1588e-04$	0.0232
1×2	**1.2559e-04**	**1.2105e-07**
1×1	**0.2015**	**0.1886**
baseline	0.6931	0.6931

Decision functions

$$F(S) = \mathrm{sigmoid}(w\,Q(S) + b)$$

- $Q(S) = |1/N \sum_i s_i|$: Magnetization

Kashiwa, Kikuchi, Tomiya, arxiv 2019

Kim, Kim, Phys Rev E 2018

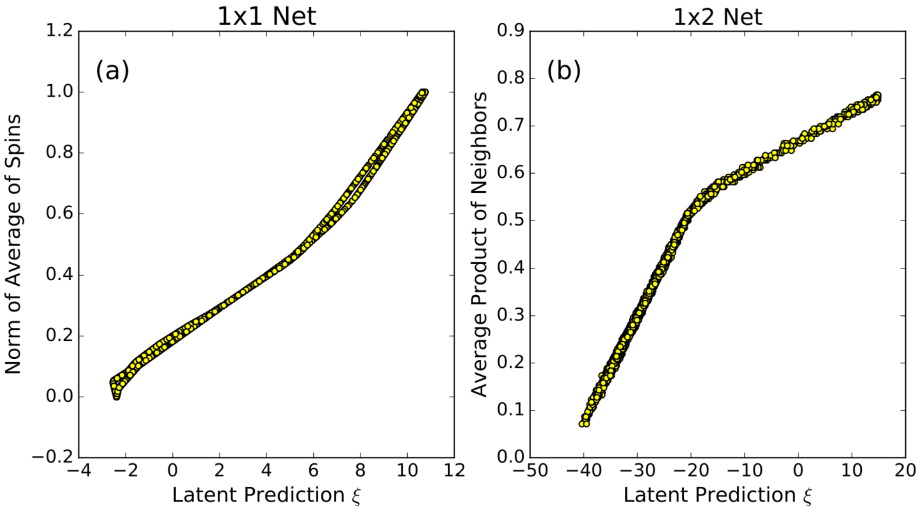
- $Q(S) = \frac{1}{N} \sum_{<i,j>nn} s_i s_j$: Expected Energy per site

Deduction easily confirmed:

- Perfect correlation

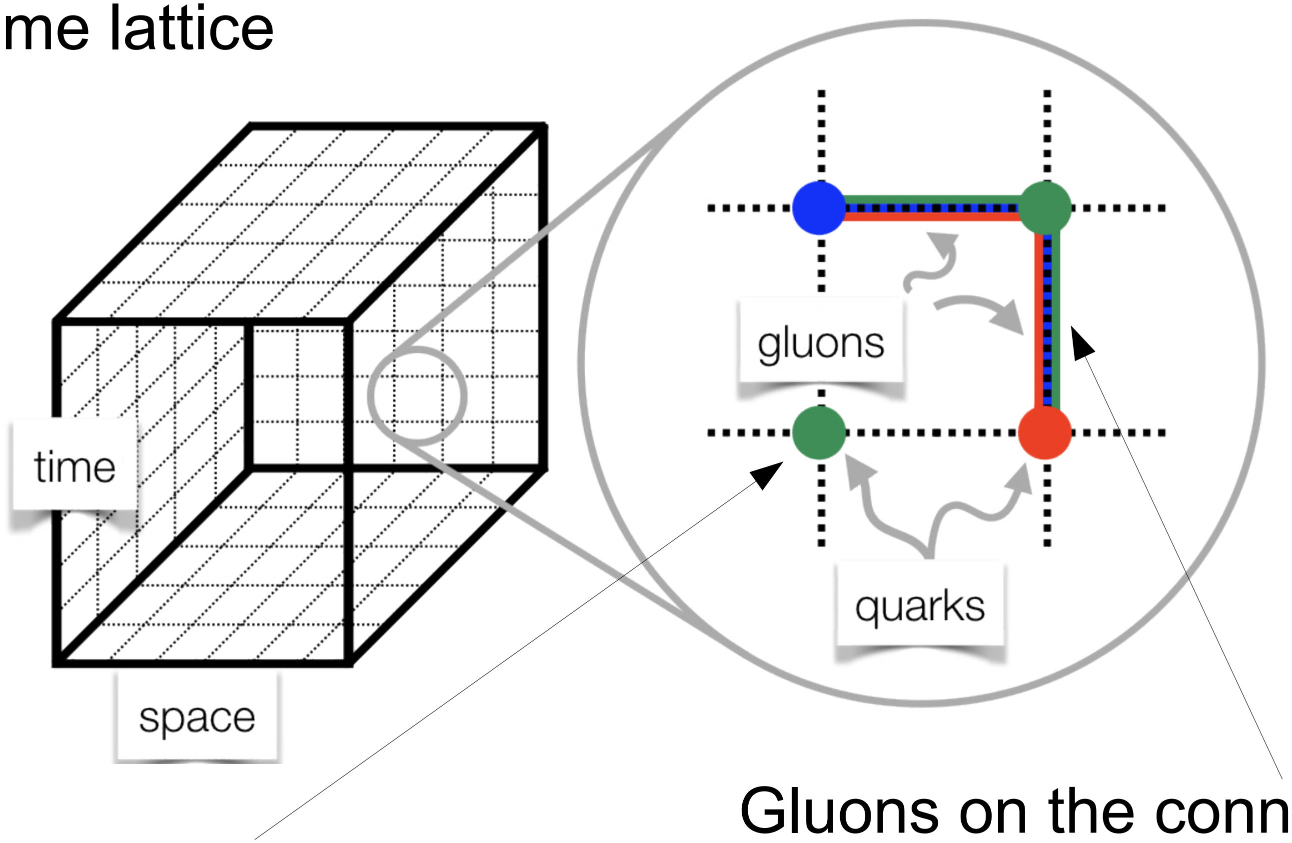
Note:

1x2 Network also has the Magnetization minimum which is easier to find!



Back to Gluons SU(2) Lattice Gauge Theory

Space time lattice



Quarks on heavy static lattice sites.

Gluons on the connections between lattice sites are described by Matrices

$$U^x_\mu \in SU(2)$$

SU(2) Lattice Gauge Theory

Describes smallest loop on the lattice

$$S_{\mathrm{Wilson}}[U] = \beta_{\mathrm{latt}} \sum_{x} \sum_{\mu<\nu} \mathrm{Re\ tr}\left(1 - U^{x}_{\mu\nu}\right)$$

$$U^{x}_{\mu\nu} = U^{x}_{\mu} U^{x+\hat{\mu}}_{\nu} U^{x+\hat{\mu}+\hat{\nu}}_{-\mu} U^{x+\hat{\nu}}_{-\nu}$$

Each Matrix connects two lattice sites

$$U^{x}_{\mu} \in SU(2)$$

$$U^{x}_{\mu} = a^{x}_{\mu} 1 + i\left(b^{x}_{\mu}\sigma_1 + c^{x}_{\mu}\sigma_2 + d^{x}_{\mu}\sigma_3\right)$$

- Each Matrix is parametrized by 4 real numbers.

  We performed a MC simulation on a lattice of size 8x8x8x2 as input for the Neural Network

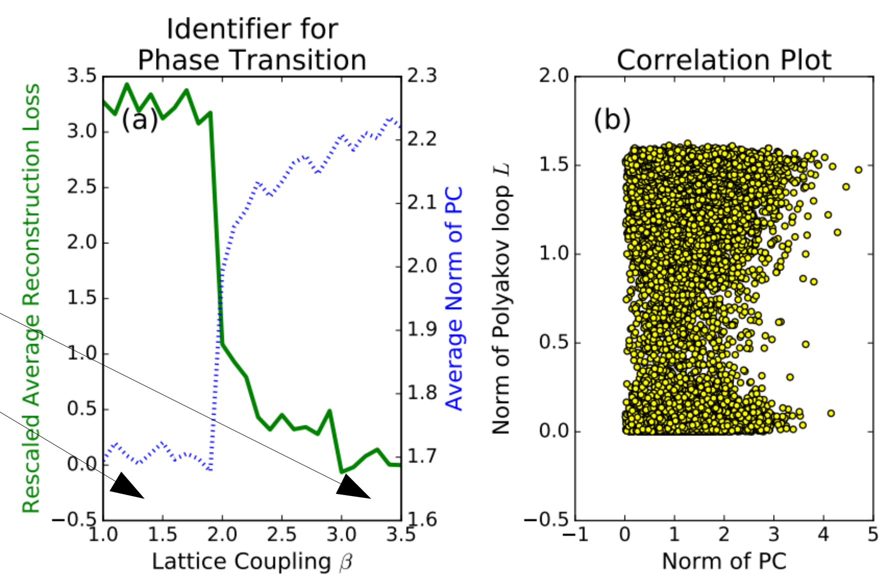
- Toy model for confinement of particles in atomic cores.
- Polyakov Loop is Order Parameter for in the limit of infinitely heavy quarks.
- Perfect Testing Ground: Polyakov Loop Order Parameter is non-linear and non-local.

Unsupervised Learning (PCA)
SU(2) Lattice Gauge Theory

$$S_{\text{Wilson}}[U] = \beta_{\text{latt}} \sum_x \sum_{\mu<\nu} \text{Re tr}\left(1 - U^x_{\mu\nu}\right)$$

- Latent parameter does not correspond to order parameter
- PCA + Reconstruction loss can be used to infer different phases

Training at phase indications from unsupervised learning (wait for next slide)



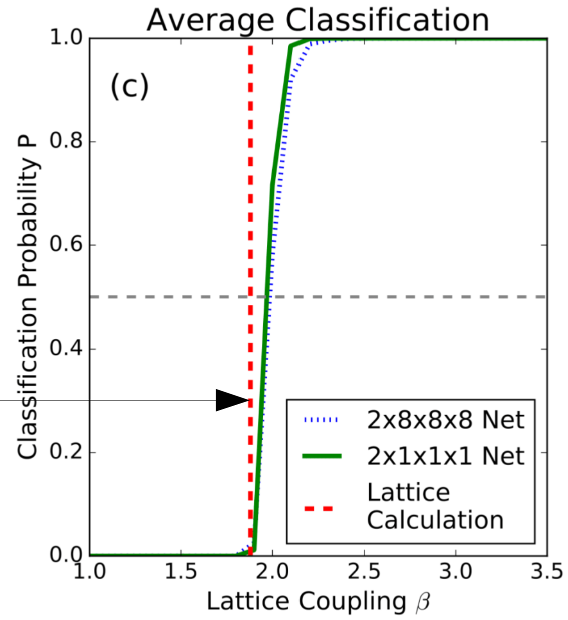
Supervised Learning
SU(2) Lattice Gauge Theory

$$S_{\text{Wilson}}[U] = \beta_{\text{latt}} \sum_x \sum_{\mu<\nu} \text{Re tr}\left(1 - U^x_{\mu\nu}\right)$$

- Find phase transition close to lattice calculation
- Prediction is inaccurate: Monte Caro Simulations not thermalized

Training at phase indications from unsupervised learning

Testing in interval containing phase transition



Interpretation of Neural Network
SU(2) Gauge Theory (2x8x8x8 Lattice)

General decision function:

$$F(S) = \text{sigmoid}(w\,Q(S) + b)$$

2x1x1x1 Decision function:

Receptive Field Size	Train Loss	Validation Loss
$2 \times 8 \times 8 \times 8$	$1.0004e-04$	$2.6266e-04$
$2 \times 1 \times 1 \times 1$	**8.8104e-08**	**6.8276e-08**
$2 \times 1 \times 1 \times 1^*$	**2.2292e-07**	**4.2958e-07**
$1 \times 1 \times 1 \times 1$	0.6620	0.9482
baseline	0.6931	0.6931

$$F(S) \approx \text{sigmoid}\left(w \left(\frac{2}{N} \sum_{\vec{x}} f(\{U_\mu^{x_0,\vec{x}}\}) \right) + b \right)$$

Regression yields 561 terms:

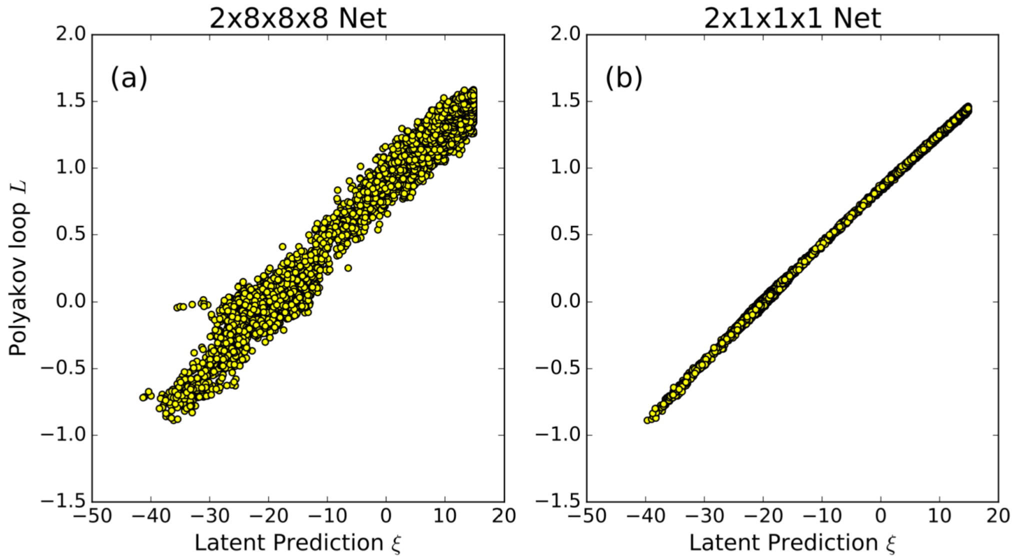
$$\begin{aligned} f(\{U_\mu^{x_0}\}) \approx &+ 7.3816\ a_\tau^0 a_\tau^1 + 0.2529\ a_\tau^1 b_\tau^1 + \ \ldots \\ &- 0.2869\ d_\tau^0 c_\tau^1 - 7.2279\ b_\tau^0 b_\tau^1 \\ &- 7.3005\ c_\tau^0 c_\tau^1 - 7.4642\ d_\tau^0 d_\tau^1\ . \end{aligned}$$

$$f(\{U_\mu^{x_0}\}) = a_\tau^0 a_\tau^1 - b_\tau^0 b_\tau^1 - c_\tau^0 c_\tau^1 - d_\tau^0 d_\tau^1 = \text{tr}\left(U_\tau^0 U_\tau^1\right)$$

- Neural Network uses **Polyakov Loop** to distinguish between phases.

Interpretation of Neural Network
SU(2) Gauge Theory (2x8x8x8 Lattice)

Deduction confirmed by perfect correlation between NN output and Polyakov Loop order parameter



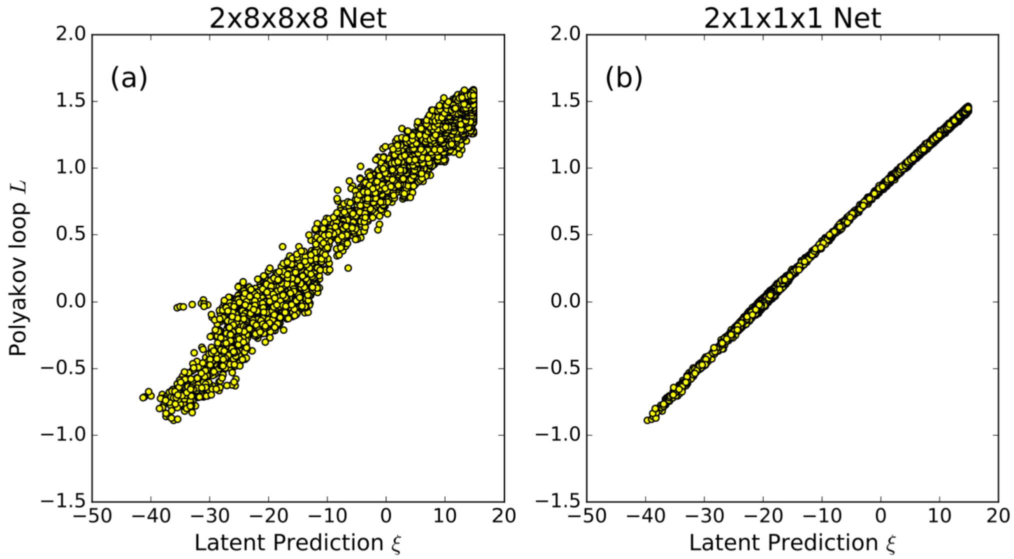
$$F(S) \approx \mathrm{sigmoid}\left(w \left(\frac{2}{N} \sum_{\vec{x}} f(\{U_\mu^{x_0,\vec{x}}\}) \right) + b \right)$$

$$f(\{U_\mu^{x_0}\}) = a_\tau^0 a_\tau^1 - b_\tau^0 b_\tau^1 - c_\tau^0 c_\tau^1 - d_\tau^0 d_\tau^1 = \mathrm{tr}\left(U_\tau^0 U_\tau^1\right)$$

- Polyakov Loop

Interpretation of Neural Network
SU(2) Gauge Theory (2x8x8x8 Lattice)

Deduction confirmed by perfect correlation between NN output and Polyakov Loop order parameter



$$F(S) \approx \mathrm{sigmoid}\left(w \left(\frac{2}{N} \sum_{\vec{x}} f(\{U_\mu^{x_0,\vec{x}}\}) \right) + b \right)$$

$$f(\{U_\mu^{x_0}\}) = a_\tau^0 a_\tau^1 - b_\tau^0 b_\tau^1 - c_\tau^0 c_\tau^1 - d_\tau^0 d_\tau^1 = \mathrm{tr}\left(U_\tau^0 U_\tau^1\right)$$

- Polyakov Loop

Note: We have constructed the PL without prior knowledge!

Conclusion

Neural Networks are capable of producing phase diagrams for many physical systems.

- NNs are no longer a black box algorithm in the context of order parameter based phase transitions.
- Neural Networks learn the same physical quantities that we humans use (Landau/Ehrenfest)
  - In the spirit of the conference: robust features
- In some cases we can determine the nature of phases by constructively interpreting what neural networks learn.