Structure and Dynamics of Deep Neural Networks

A Perspective from Geometry and Physics

Docent Lecture

Jan E. Gerken

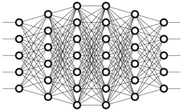
CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Machine learning with neural networks

- Neural networks

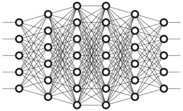


$\Rightarrow \quad \mathcal{N}_\theta : \mathbb{R}^m \to \mathbb{R}^n, \quad \theta \in \mathbb{R}^N$: parameters

e.g. $\mathcal{N}_\theta$: picture $\mapsto P(\text{cat})$

Machine learning with neural networks

- Neural networks



$$\Rightarrow \qquad \mathcal{N}_\theta : \mathbb{R}^m \to \mathbb{R}^n , \qquad \theta \in \mathbb{R}^N\text{: parameters}$$

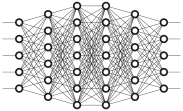
e.g. $\mathcal{N}_\theta : \text{picture} \mapsto P(\text{cat})$

- Training data
  examples $x \mapsto y$ of target function, e.g.

$\mathcal{D} = \Big\{$  $\mapsto 1.0 ,$  $\mapsto 0.0 , \quad \ldots \Big\}$

Machine learning with neural networks

- Neural networks



$$\Rightarrow \quad \mathcal{N}_\theta : \mathbb{R}^m \to \mathbb{R}^n, \qquad \theta \in \mathbb{R}^N\text{: parameters}$$

e.g. $\mathcal{N}_\theta$: picture $\mapsto P(\text{cat})$

- Training data
  examples $x \mapsto y$ of target function, e.g.

$\mathcal{D} = \{$  $\mapsto 1.0,$  $\mapsto 0.0, \quad \ldots \}$

- Optimize θ so that $\mathcal{N}_\theta$ matches the target function

Neural networks

Neural networks

- Consider a very simple function

$$f_{W,b} : \mathbb{R}^m \to \mathbb{R}^n \qquad f_{W,b}(x) = \quad Wx + b$$

Neural networks

- Consider a very simple function

$$f_{W,b} : \mathbb{R}^m \to \mathbb{R}^n \qquad f_{W,b}(x) = \quad Wx + b$$

- Compose many of these

$$\mathcal{N}_\theta = f_{W^{(L)},b^{(L)}} \circ f_{W^{(L-1)},b^{(L-1)}} \circ \cdots \circ f_{W^{(1)},b^{(1)}}$$

Neural networks

- Consider a very simple function

$$\sigma(x) = \max(0, x)$$

$$f_{W,b} : \mathbb{R}^m \to \mathbb{R}^n \qquad f_{W,b}(x) = \sigma(Wx + b)$$

- Compose many of these

$$\mathcal{N}_\theta = f_{W^{(L)},b^{(L)}} \circ f_{W^{(L-1)},b^{(L-1)}} \circ \cdots \circ f_{W^{(1)},b^{(1)}}$$

Neural networks

- Consider a very simple function

$\sigma(x) = \max(0, x)$

$$f_{W,b} : \mathbb{R}^m \to \mathbb{R}^n \qquad f_{W,b}(x) = \sigma(Wx + b)$$

- Compose many of these

$$\mathcal{N}_\theta = f_{W^{(L)},b^{(L)}} \circ f_{W^{(L-1)},b^{(L-1)}} \circ \cdots \circ f_{W^{(1)},b^{(1)}}$$

- Optimize all parameters

$$\theta = \left\{ \left(W^{(\ell)}, b^{(\ell)}\right) \,\middle|\, \ell = 1, \ldots, L \right\}$$

Neural networks

- Consider a very simple function

$\sigma(x) = \max(0, x)$

$$f_{W,b} : \mathbb{R}^m \to \mathbb{R}^n \qquad f_{W,b}(x) = \sigma(Wx + b)$$

- Compose many of these

$$\mathcal{N}_\theta = f_{W^{(L)},b^{(L)}} \circ f_{W^{(L-1)},b^{(L-1)}} \circ \cdots \circ f_{W^{(1)},b^{(1)}}$$

- Optimize all parameters

$$\theta = \left\{ \left(W^{(\ell)}, b^{(\ell)}\right) \,\middle|\, \ell = 1, \ldots, L \right\}$$

What can we say about $\mathcal{N}_\theta$ mathematically?

Surprising properties of neural networks

Neural networks can approximate any function

Surprising properties of neural networks

Neural networks can approximate any function

Consider a network $\mathcal{N}_\theta : \mathbb{R} \to \mathbb{R}$ with one hidden layer of width 4

$$f_i^{(1)}(x) = \sigma\left(W_i^{(1)} x + b_i^{(1)}\right), \quad i = 1, 2, 3, 4$$

$$\mathcal{N}_\theta(x) = \sum_{i=1}^{4} W_i^{(2)} f_i^{(1)}(x) + b^{(2)}$$

Surprising properties of neural networks

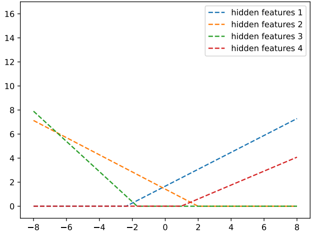
Neural networks can approximate any function

Consider a network $\mathcal{N}_\theta : \mathbb{R} \to \mathbb{R}$ with one hidden layer of width 4

$$f_i^{(1)}(x) = \sigma\left(W_i^{(1)} x + b_i^{(1)}\right), \quad i = 1, 2, 3, 4$$

$$\mathcal{N}_\theta(x) = \sum_{i=1}^{4} W_i^{(2)} f_i^{(1)}(x) + b^{(2)}$$

For random parameters, we get



Surprising properties of neural networks

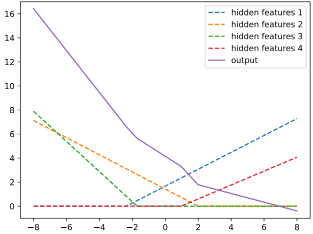
Neural networks can approximate any function

Consider a network $\mathcal{N}_\theta : \mathbb{R} \to \mathbb{R}$ with one hidden layer of width 4

$$f_i^{(1)}(x) = \sigma\left(W_i^{(1)} x + b_i^{(1)}\right), \quad i = 1, 2, 3, 4$$

$$\mathcal{N}_\theta(x) = \sum_{i=1}^{4} W_i^{(2)} f_i^{(1)}(x) + b^{(2)}$$

For random parameters, we get



Surprising properties of neural networks

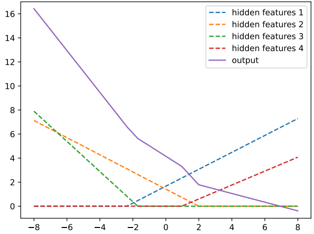
Neural networks can approximate any function

Consider a network $\mathcal{N}_\theta : \mathbb{R} \to \mathbb{R}$ with one hidden layer of width 4

$$f_i^{(1)}(x) = \sigma\left(W_i^{(1)} x + b_i^{(1)}\right), \quad i = 1, 2, 3, 4$$

$$\mathcal{N}_\theta(x) = \sum_{i=1}^{4} W_i^{(2)} f_i^{(1)}(x) + b^{(2)}$$

For random parameters, we get



→ number of linear pieces exponential in depth

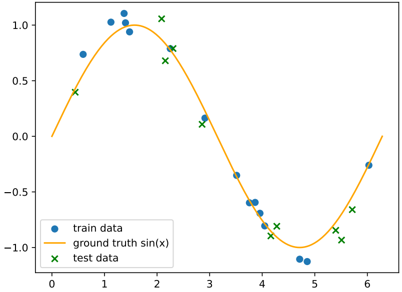
Surprising properties of neural networks

Neural networks work very well!

Surprising properties of neural networks

Neural networks work very well!

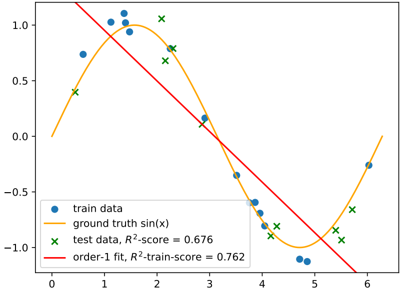
Consider a simple machine-learning problem



Surprising properties of neural networks

Neural networks work very well!

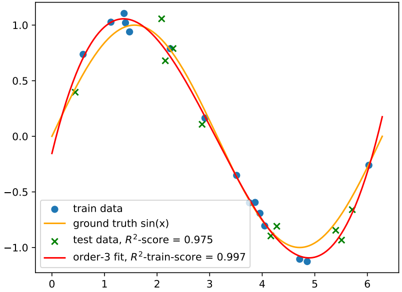
Consider a simple machine-learning problem



Surprising properties of neural networks

Neural networks work very well!

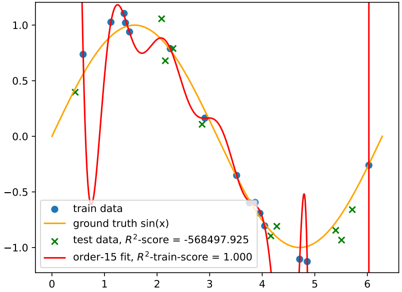
Consider a simple machine-learning problem



Surprising properties of neural networks

Neural networks work very well!

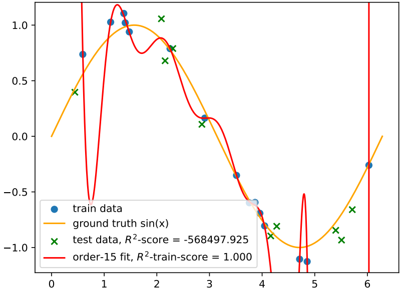
Consider a simple machine-learning problem

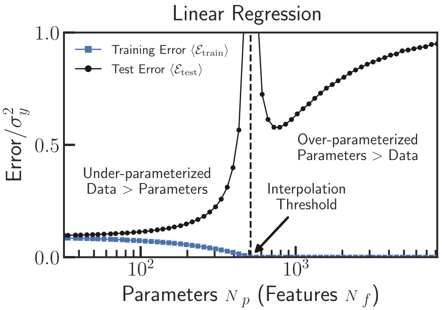


Surprising properties of neural networks

Neural networks work very well!

Consider a simple machine-learning problem



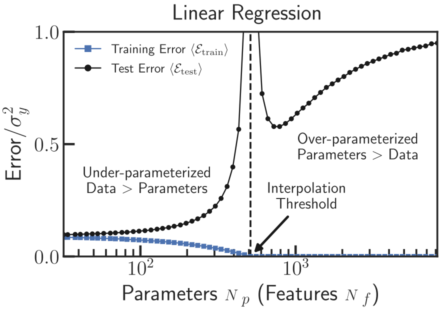


[Rocks, Mehta 2020]

Surprising properties of neural networks

Neural networks work very well!

Neural networks show double decent

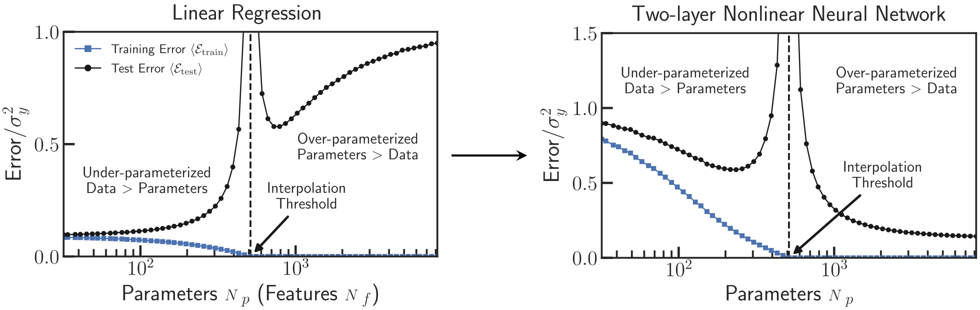


[Rocks, Mehta 2020]

Surprising properties of neural networks

Neural networks work very well!

Neural networks show double decent



[Rocks, Mehta 2020]

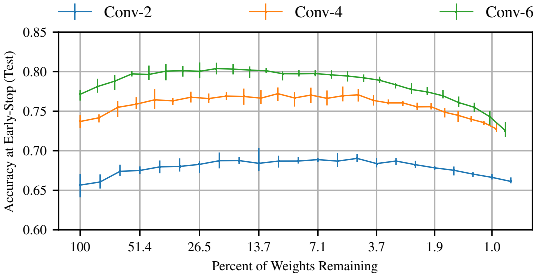
Surprising properties of neural networks

Only a small subnetwork matters

Surprising properties of neural networks

Only a small subnetwork matters

Possible to find tiny subnetworks which have almost the same performance as the full network



[Frankle, Carbin 2019]

Hyperparameter tuning

Hyperparameter tuning

- We use some algorithm to optimize the parameters θ

Hyperparameter tuning

- We use some algorithm to optimize the parameters θ
- Hyperparameters are not optimized
  - Parameters of the optimization algorithm
  - Sizes of intermediate vector spaces
  - How many layers, which non-linear function…

Hyperparameter tuning

- We use some algorithm to optimize the parameters θ
- Hyperparameters are not optimized
  - Parameters of the optimization algorithm
  - Sizes of intermediate vector spaces
  - How many layers, which non-linear function…
- In practice, huge amount of compute spent on trial and error

Geometry and Physics in Neural Networks

Geometry and Physics in Neural Networks

- Can we exploit the geometry of the data distribution?
- Can we exploit symmetries of the target function?
- Can we use methods from theoretical physics to understand the learning process?

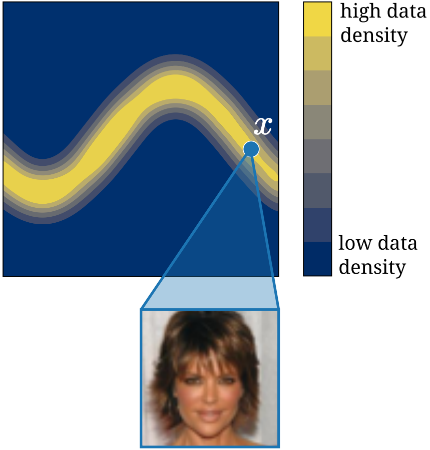
Geometry and Physics in Neural Networks

Geometric Deep Learning

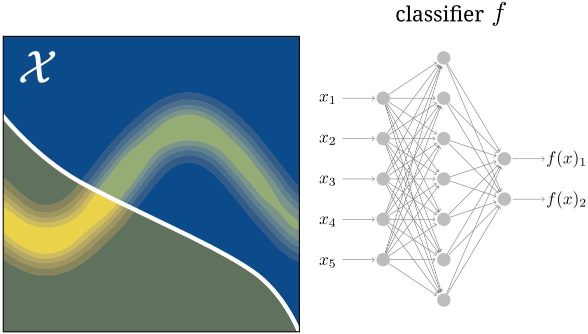
- Can we exploit the geometry of the data distribution?
- Can we exploit symmetries of the target function?
- Can we use methods from theoretical physics to understand the learning process?

Geometry of the Data Distribution

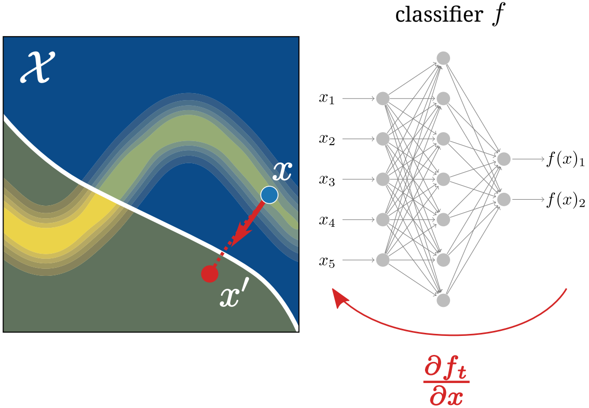
Manifold Hypothesis



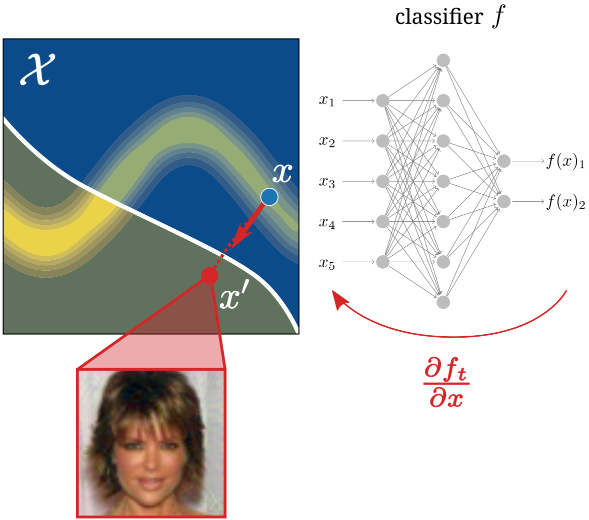
Adversarial Examples



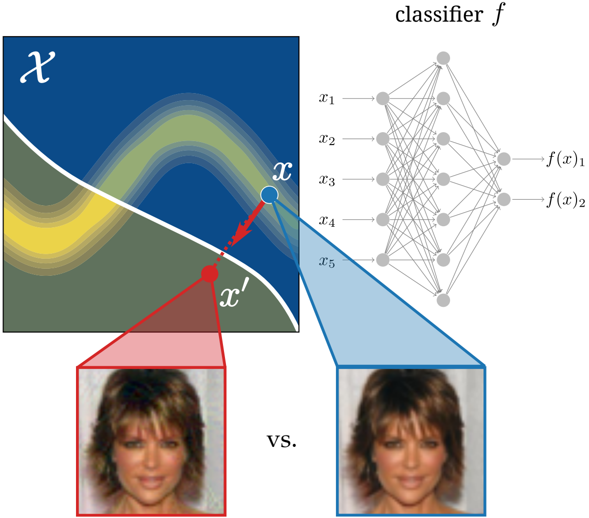
Adversarial Examples



Adversarial Examples



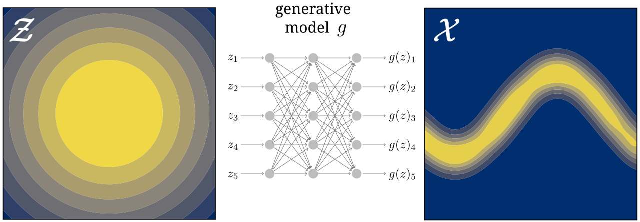
Adversarial Examples



Geometry of the Data Manifold

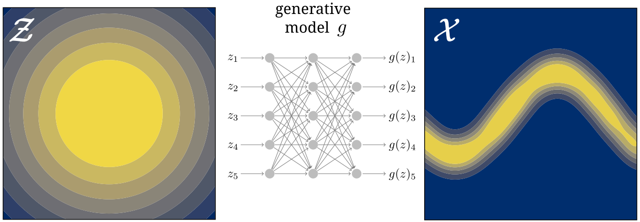
Geometry of the Data Manifold

Use model to learn diffeomorphism to normal distribution



Geometry of the Data Manifold

Use model to learn diffeomorphism to normal distribution



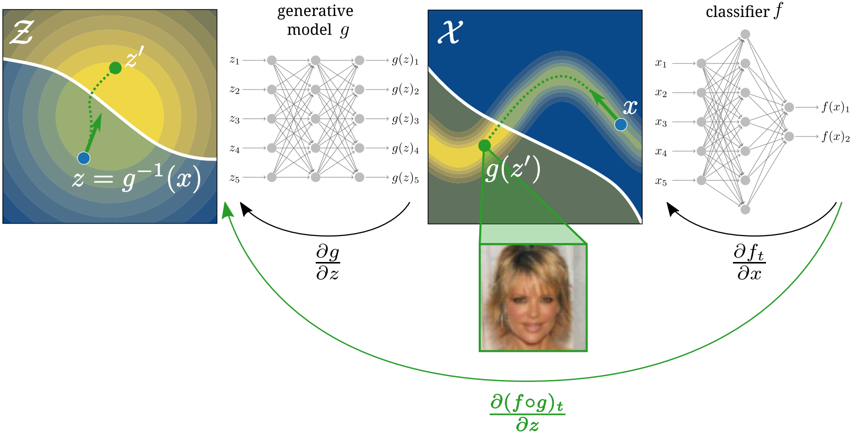
Euclidean metric δ

induced metric γ

Optimize along the Data Manifold

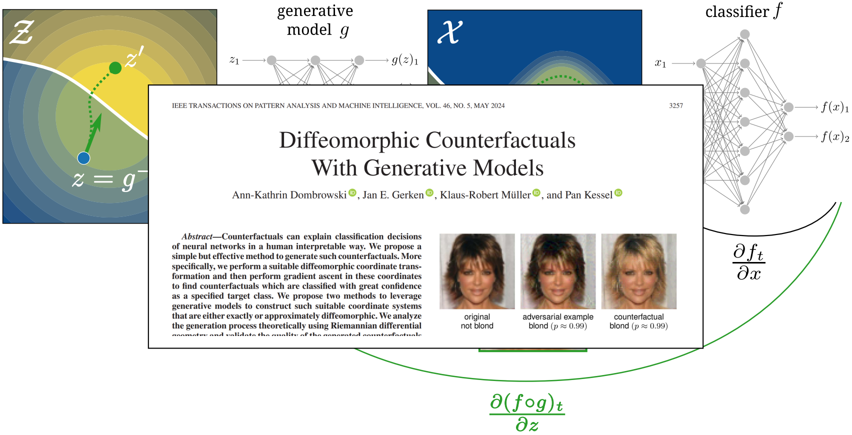
Optimize along the Data Manifold

Follow the gradient in the normalized coordinates



Optimize along the Data Manifold

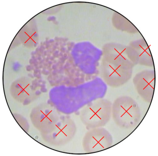
Follow the gradient in the normalized coordinates



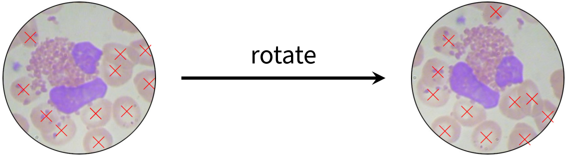
Symmetries

Symmetry of the Data

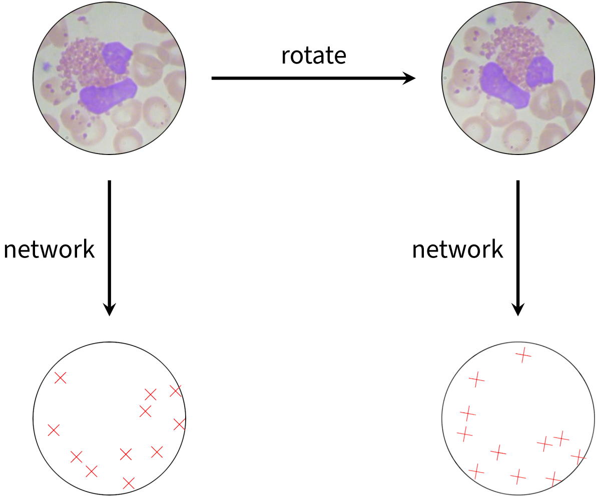
Symmetry of the Data



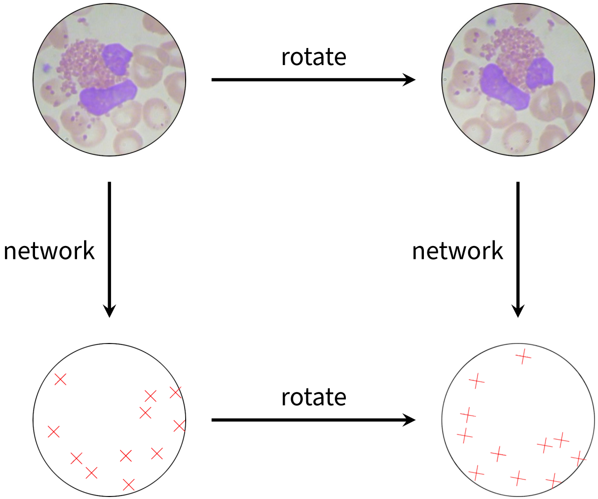
Symmetry of the Data



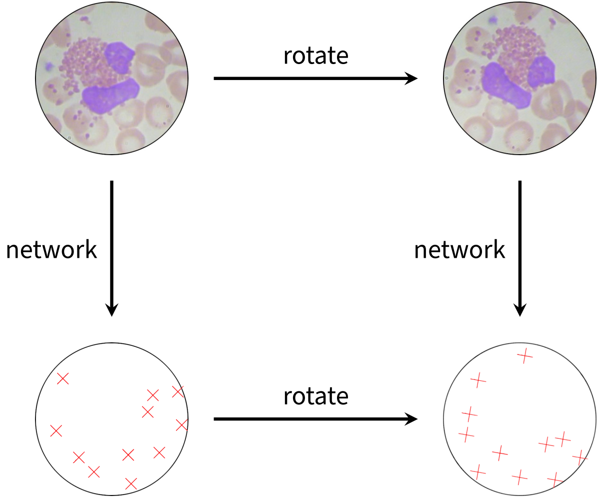
Symmetry of the Data



Symmetry of the Data



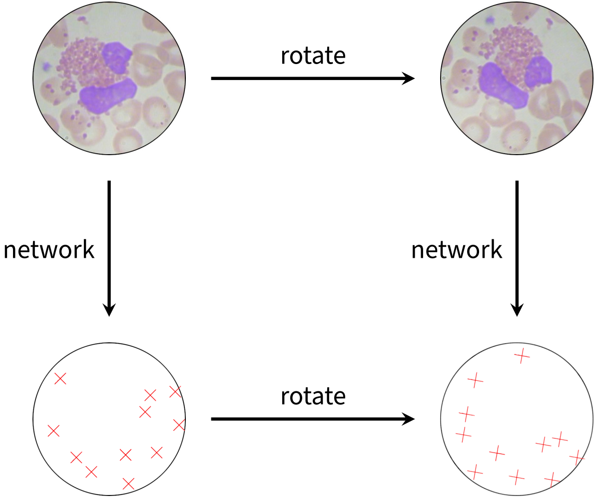
Symmetry of the Data



$$\Longleftrightarrow \quad \mathcal{N}(\rho_{\text{in}}(g)x) = \rho_{\text{out}}(g)\mathcal{N}(x)$$

rotation (ρ_{in}) · image (x)

Symmetry of the Data $\Rightarrow$ Equivariant Networks



$$\Longleftrightarrow \quad \mathcal{N}(\rho_{\text{in}}(g)x) = \rho_{\text{out}}(g)\mathcal{N}(x)$$

rotation — image

Equivariant neural networks

Equivariant neural networks

Group Equivariant Convolutional Networks

Taco S. Cohen T.S.COHEN@UVA.NL
University of Amsterdam

Max Welling M.WELLING@UVA.NL
University of Amsterdam
University of California Irvine
Canadian Institute for Advanced Research

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symme-

Convolution layers can be used effectively in a *deep* network because all the layers in such a network are *translation equivariant*: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature maps (at least up to edge-effects). In

Equivariant neural networks

Group Equivariant Convolutional Networks

Taco S. Cohen T.S.COHEN@UVA.NL
University of Amsterdam

Max Welling M.WELLING@UVA.NL
University of Amsterdam
University of California Irvine
Canadian Institute for Advanced Research

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symme-

Convolution layers can be used effectively in a *deep* network because all the layers in such a network are *translation equivariant*: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature maps (at least up to edge-effects). In

Equivariant Transformer Networks

Kai Sheng Tai [1] **Peter Bailis** [1] **Gregory Valiant** [1]

Abstract

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivariant Transformers (ETs), a family of differentiable image-to-image mappings that improve the robustness of models towards pre-defined continuous transformation groups. Through the use of specially-derived canonical coordinate systems, ETs incorporate functions that

scaling to each training image). While data augmentation typically helps reduce the test error of CNN-based models, there is no guarantee that transformation invariance will be enforced for data not seen during training.

In contrast to training time approaches like data augmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016; Dieleman et al., 2016; Marcos et al., 2017; Worrall et al., 2017; Henriques & Vedaldi, 2017; Cohen et al., 2018) has explored new CNN architectures that are guaranteed to respond predictably to particular transforma-

Equivariant neural networks

Group Equivariant Convolutional Networks

Taco S. Cohen T.S.COHEN@UVA.NL
University of Amsterdam

Max Welling M.WELLING@UVA.NL
University of Amsterdam
University of California Irvine
Canadian Institute for Advanced Research

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symme-

Convolution layers can be used effectively in a *deep* network because all the layers in such a network are *translation equivariant*: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature maps (at least up to edge-effects). In

Equivariant Transformer Networks

Kai Sheng Tai [1] **Peter Bailis** [1] **Gregory Valiant** [1]

Abstract

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivariant Transformers (ETs), a family of differentiable image-to-image mappings that improve the robustness of models towards pre-defined continuous transformation groups. Through the use of specially-derived canonical coordinate systems, ETs incorporate functions that

scaling to each training image). While data augmentation typically helps reduce the test error of CNN-based models, there is no guarantee that transformation invariance will be enforced for data not seen during training.

In contrast to training time approaches like data augmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016; Dieleman et al., 2016; Marcos et al., 2017; Worrall et al., 2017; Henriques & Vedaldi, 2017; Cohen et al., 2018) has explored new CNN architectures that are guaranteed to respond predictably to particular transforma-

Theory for Equivariant Quantum Neural Networks

Quynh T. Nguyen,[1,2] Louis Schatzki,[3,4] Paolo Braccia,[1,5] Michael Ragone,[1,6] Patrick J. Coles,[1] Frédéric Sauvage,[1] Martín Larocca,[1,7] and M. Cerezo[3]

[1] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[2] *School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*
[3] *Information Sciences, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[4] *Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[5] *Dipartimento di Fisica e Astronomia, Università di Firenze, Sesto Fiorentino (FI), 50019 , Italy*
[6] *Department of Mathematics, University of California Davis, Davis, California 95616, USA*
[7] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

Quantum neural network architectures that have little-to-no inductive biases are known to face trainability and generalization issues. Inspired by a similar problem, recent breakthroughs in machine learning address this challenge by creating models encoding the symmetries of the learning task. This is materialized through the usage of equivariant neural networks whose action commutes with that of the symmetry. In this work, we import these ideas to the quantum realm by

Equivariant neural networks

Group Equivariant Convolutional Networks

Taco S. Cohen T.S.COHEN@UVA.NL
University of Amsterdam

Max Welling M.WELLING@UVA.NL
University of Amsterdam
University of California Irvine
Canadian Institute for Advanced Research

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symme-

Convolution layers can be used effectively in a *deep* network because all the layers in such a network are *translation equivariant*: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting

Equivariant Transformer Networks

Kai Sheng Tai [1] **Peter Bailis** [1] **Gregory Valiant** [1]

Abstract

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivariant Transformers (ETs), a family of differentiable image-to-image mappings that improve the robustness of models towards pre-defined continuous transformation groups. Through the use of specially-derived canonical co-

scaling to each training image). While data augmentation typically helps reduce the test error of CNN-based models, there is no guarantee that transformation invariance will be enforced for data not seen during training.

In contrast to training time approaches like data augmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016; Dieleman et al., 2016; Marcos et al., 2017; Worrall et al., 2017; Henriques & Vedaldi, 2017; Cohen et al., 2018) has explored new CNN architectures that are

Theory for Equivariant Quantum Neural Networks

Quynh T. Nguyen,[1,2] Louis Schatzki,[3,4] Paolo Braccia,[1,5] Michael Ragone,[1,6] Patrick J. Coles,[1] Frédéric Sauvage,[1] Martín Larocca,[1,7] and M. Cerezo[3]

[1] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[2] *School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*
[3] *Information Sciences, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[4] *Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[5] *Dipartimento di Fisica e Astronomia, Università di Firenze, Sesto Fiorentino (FI), 50019 , Italy*
[6] *Department of Mathematics, University of California Davis, Davis, California 95616, USA*
[7] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

Quantum neural network architectures that have little-to-no inductive biases are known to face trainability and generalization issues. Inspired by a similar problem, recent breakthroughs in machine learning address this challenge by creating models encoding the symmetries of the learning task. This is materialized through the usage of equivariant neural networks whose action com-

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging

Shiqi Gong[a,e,1] **Qi Meng**[b] **Jue Zhang**[b] **Huilin Qu**[c] **Congqiao Li**[d] **Sitian Qian**[d] **Weitao Du**[a] **Zhi-Ming Ma**[a] **Tie-Yan Liu**[b]

[a] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Zhongguancun East Road, Beijing 100190, China*
[b] *Microsoft Research Asia, Danling Street, Beijing 100080, China*
[c] *CERN, EP Department, CH-1211 Geneva 23, Switzerland*
[d] *School of Physics, Peking University, Chengfu Road, Beijing 100871, China*

Equivariant neural networks

Group Equivariant Convolutional Networks

Taco S. Cohen T.S.COHEN@UVA.NL
University of Amsterdam
Max Welling M.WELLING@UVA.NL
University of Amsterdam
University of California Irvine
Canadian Institute for Advanced Research

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symme-

Convolution layers can be used effectively in a *deep* network because all the layers in such a network are *translation equivariant*: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature maps (at least up to edge-effects). In

Equivariant Transformer Networks

Kai Sheng Tai [1] **Peter Bailis** [1] **Gregory Valiant** [1]

Abstract

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivariant Transformers (ETs), a family of differentiable image-to-image mappings that improve the robustness of models towards pre-defined continuous transformation groups. Through the use of specially-derived canonical co-

scaling to each training image). While data augmentation typically helps reduce the test error of CNN-based models, there is no guarantee that transformation invariance will be enforced for data not seen during training.

In contrast to training time approaches like data augmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016; Dieleman et al., 2016; Marcos et al., 2017; Worrall et al., 2017; Henriques & Vedaldi, 2017; Cohen et al., 2018) has explored new CNN architectures that are

Theory for Equivariant Quantum Neural Networks

Quynh T. Nguyen,[1,2] Louis Schatzki,[3,4] Paolo Braccia,[1,5] Michael Ragone,[1,6] Patrick J. Coles,[1] Frédéric Sauvage,[1] Martín Larocca,[1,7] and M. Cerezo[3]

[1] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[2] *School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*
[3] *Information Sciences, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[4] *Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[5] *Dipartimento di Fisica e Astronomia, Università di Firenze, Sesto Fiorentino (FI), 50019 , Italy*
[6] *Department of Mathematics, University of California Davis, Davis, California 95616, USA*
[7] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

Quantum neural network architectures that have little-to-no inductive biases are known to face trainability and generalization issues. Inspired by a similar problem, recent breakthroughs in machine learning address this challenge by creating models encoding the symmetries of the learning task. This is materialized through the usage of equivariant neural networks whose action com-

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging

Shiqi Gong[a,e,1] **Qi Meng**[b] **Jue Zhang**[b] **Huilin Qu**[c] **Congqiao Li**[d] **Sitian Qian**[d] **Weitao Du**[a] **Zhi-Ming Ma**[a] **Tie-Yan Liu**[b]

[a] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Zhongguancun East Road, Beijing 100190, China*
[b] *Microsoft Research Asia, Danling Street, Beijing 100080, China*
[c] *CERN, EP Department, CH-1211 Geneva 23, Switzerland*
[d] *School of Physics, Peking University, Chengfu Road, Beijing 100871, China*

E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials

Simon Batzner*,[1] Albert Musaelian,[1] Lixin Sun,[1] Mario Geiger,[2] Jonathan P. Mailoa,[3] Mordechai Kornbluth,[3] Nicola Molinari,[1] Tess E. Smidt,[4,5] and Boris Kozinsky*[1,3]

[1] *John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA*
[2] *École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*
[3] *Robert Bosch Research and Technology Center, Cambridge, MA 02139, USA*
[4] *Computational Research Division and Center for Advanced Mathematics for Energy Research Applications, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[5] *Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA 02142, USA*

This work presents Neural Equivariant Interatomic Potentials (NequIP), an E(3)-equivariant neural network approach for learning interatomic potentials from *ab-initio* calculations for molecular dynamics simulations. While most contemporary symmetry-aware models use invariant convolutions and only act on scalars, NequIP employs E(3)-equivariant convolutions for interactions of geometric tensors, resulting in a more information-rich and faithful representation of atomic environments. The method achieves state-of-the-art accuracy on a challenging and diverse set of molecules and

Equivariant neural networks

Group Equivariant Convolutional Networks

Taco S. Cohen T.S.COHEN@UVA.NL
University of Amsterdam

Max Welling M.WELLING@UVA.NL
University of Amsterdam
University of California Irvine
Canadian Institute for Advanced Research

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symme-

Convolution layers can be used effectively in a *deep* network because all the layers in such a network are *translation equivariant*: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature maps (at least up to edge-effects). In

Equivariant Transformer Networks

Kai Sheng Tai [1] **Peter Bailis** [1] **Gregory Valiant** [1]

Abstract

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivariant Transformers (ETs), a family of differentiable image-to-image mappings that improve the robustness of models towards pre-defined continuous transformation groups. Through the use of specially-derived canonical coordinate systems, ETs incorporate functions that

scaling to each training image). While data augmentation typically helps reduce the test error of CNN-based models, there is no guarantee that transformation invariance will be enforced for data not seen during training.

In contrast to training time approaches like data augmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016; Dieleman et al., 2016; Marcos et al., 2017; Worrall et al., 2017; Henriques & Vedaldi, 2017; Cohen et al., 2018) has explored new CNN architectures that are guaranteed to respond predictably to particular transforma-

Theory for Equivariant Quantum Neural Networks

Quynh T. Nguyen,[1,2] Louis Schatzki,[3,4] Paolo Braccia,[1,5] Michael Ragone,[1,6]
Patrick J. Coles,[1] Frédéric Sauvage,[1] Martín Larocca,[1,7] and M. Cerezo[3]

[1] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[2] *School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*
[3] *Information Sciences, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[4] *Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[5] *Dipartimento di Fisica e Astronomia, Università di Firenze, Sesto Fiorentino (FI), 50019 , Italy*
[6] *Department of Mathematics, University of California Davis, Davis, California 95616, USA*
[7] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

Quantum neural network architectures that have little-to-no inductive biases are known to face trainability and generalization issues. Inspired by a similar problem, recent breakthroughs in machine learning address this challenge by creating models encoding the symmetries of the learning task. This is materialized through the usage of equivariant neural networks whose action com-

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging

Shiqi Gong[a,e,1] Qi Meng[b] Jue Zhang[b] Huilin Qu[c] Congqiao Li[d] Sitian Qian[d] Weitao Du[a] Zhi-Ming Ma[a] Tie-Yan Liu[b]

[a] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Zhongguancun East Road, Beijing 100190, China*
[b] *Microsoft Research Asia, Danling Street, Beijing 100080, China*
[c] *CERN, EP Department, CH-1211 Geneva 23, Switzerland*
[d] *School of Physics, Peking University, Chengfu Road, Beijing 100871, China*

E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials

Simon Batzner*,[1] Albert Musaelian,[1] Lixin Sun,[1] Mario Geiger,[2] Jonathan P. Mailoa,[3] Mordechai Kornbluth,[3] Nicola Molinari,[1] Tess E. Smidt,[4,5] and Boris Kozinsky*[1,3]

[1] *John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA*
[2] *École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*
[3] *Robert Bosch Research and Technology Center, Cambridge, MA 02139, USA*
[4] *Computational Research Division and Center for Advanced Mathematics for Energy Research Applications, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[5] *Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA 02142, USA*

This work presents Neural Equivariant Interatomic Potentials (NequIP), an E(3)-equivariant neural network approach for learning interatomic potentials from *ab-initio* calculations for molecular dynamics simulations. While most contemporary symmetry-aware models use invariant convolutions and only act on scalars, NequIP employs E(3)-equivariant convolutions for interactions of geometric tensors, resulting in a more information-rich and faithful representation of atomic environments. The method achieves state-of-the-art accuracy on a challenging and diverse set of molecules and

HIERARCHICAL, ROTATION-EQUIVARIANT NEURAL NETWORKS TO SELECT STRUCTURAL MODELS OF PROTEIN COMPLEXES

Stephan Eismann* Department of Applied Physics, Stanford University, seismann@stanford.edu
Raphael J.L. Townshend* Department of Computer Science, Stanford University, raphael@cs.stanford.edu
Nathaniel Thomas* Department of Physics, Stanford University, nthomas103@gmail.com
Milind Jagota Department of Electrical Engineering, Stanford University, mjagota@stanford.edu
Bowen Jing Department of Computer Science, Stanford University, bjing@stanford.edu
Ron O. Dror Department of Computer Science, Stanford University, rondror@cs.stanford.edu

ABSTRACT

Predicting the structure of multi-protein complexes is a grand challenge in biochemistry, with major implications for basic science and drug discovery. Computational structure prediction methods generally leverage pre-defined structural features to distinguish accurate structural models from less accurate ones. This raises the question of whether it is possible to learn characteristics of accurate models directly from atomic coordinates of protein complexes, with no prior assumptions. Here we introduce a machine learning method that learns directly from the 3D positions of all atoms to

Equivariant Neural Networks

What is the general mathematical formulation of equivariant neural networks?

Geometric deep learning and equivariant neural networks

Jan E. Gerken[1,2,3] · Jimmy Aronsson[1] · Oscar Carlsson[1] · Hampus Linander[4] · Fredrik Ohlsson[5] · Christoffer Petersson[1,6] · Daniel Persson[1]

Accepted: 1 May 2023 / Published online: 4 June 2023
© The Author(s) 2023

Abstract
We survey the mathematical foundations of geometric deep learning, focusing on group equivariant and gauge equivariant neural networks. We develop gauge equivariant convolutional neural networks on arbitrary manifolds $\mathcal{M}$ using principal bundles with structure

Equivariant Neural Networks

What is the general mathematical formulation of equivariant neural networks?

Check for updates

Geometric deep learning and equivariant neural networks

Jan E. Gerken[1,2,3] · Jimmy Aronsson[1] · Oscar Carlsson[1] · Hampus Linander[4] · Fredrik Ohlsson[5] · Christoffer Petersson[1,6] · Daniel Persson[1]

Accepted: 1 May 2023 / Published online: 4 June 2023
© The Author(s) 2023

Abstract
We survey the mathematical foundations of geometric deep learning, focusing on group equivariant and gauge equivariant neural networks. We develop gauge equivariant convolutional neural networks on arbitrary manifolds $\mathcal{M}$ using principal bundles with structure

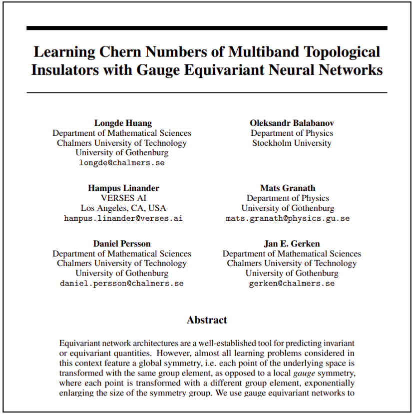
Fiber bundles

Representation theory

Spherical harmonics, Wigner matrices…

Equivariant Neural Networks

Use gauge equivariant network to learn topological invariants



Learning Chern Numbers of Multiband Topological Insulators with Gauge Equivariant Neural Networks

Longde Huang
Department of Mathematical Sciences
Chalmers University of Technology
University of Gothenburg
longde@chalmers.se

Oleksandr Balabanov
Department of Physics
Stockholm University

Hampus Linander
VERSES AI
Los Angeles, CA, USA
hampus.linander@verses.ai

Mats Granath
Department of Physics
University of Gothenburg
mats.granath@physics.gu.se

Daniel Persson
Department of Mathematical Sciences
Chalmers University of Technology
University of Gothenburg
daniel.persson@chalmers.se

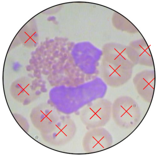
Jan E. Gerken
Department of Mathematical Sciences
Chalmers University of Technology
University of Gothenburg
gerken@chalmers.se

Abstract

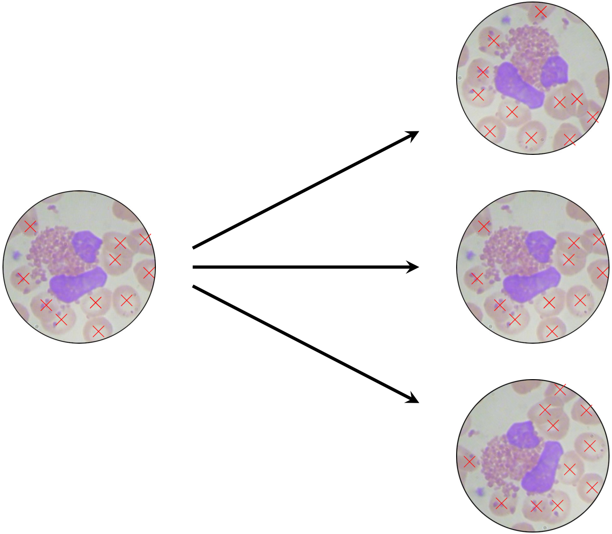
Equivariant network architectures are a well-established tool for predicting invariant or equivariant quantities. However, almost all learning problems considered in this context feature a global symmetry, i.e. each point of the underlying space is transformed with the same group element, as opposed to a local *gauge* symmetry, where each point is transformed with a different group element, exponentially enlarging the size of the symmetry group. We use gauge equivariant networks to

Learning symmetries

Learning symmetries



Learning symmetries



Impose symmetries or learn them?

Impose symmetries or learn them?

Article

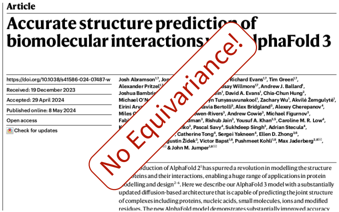
Accurate structure prediction of biomolecular interactions with AlphaFold 3

https://doi.org/10.1038/s41586-024-07487-w
Received: 19 December 2023
Accepted: 29 April 2024
Published online: 8 May 2024
Open access
Check for updates

Josh Abramson[1,7], Jonas Adler[1,7], Jack Dunger[1,7], Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7], Olaf Ronneberger[1,7], Lindsay Willmore[1,7], Andrew J. Ballard[1], Joshua Bambrick[2], Sebastian W. Bodenstein[1], David A. Evans[1], Chia-Chun Hung[2], Michael O'Neill[1], David Reiman[1], Kathryn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arvaniti[3], Charles Beattie[3], Ottavia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles Congreve[4], Alexander I. Cowen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fabian B. Fuchs[3], Hannah Gladman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], Kuba Perlin[3], Anna Potapenko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], Ashok Thillaisundaram[3], Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], Michal Zielinski[3], Augustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8] ✉, Demis Hassabis[1,2,8] ✉ & John M. Jumper[1,8] ✉

The introduction of AlphaFold 2[1] has spurred a revolution in modelling the structure of proteins and their interactions, enabling a huge range of applications in protein modelling and design[2–6]. Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved accuracy

Impose symmetries or learn them?



Article

Accurate structure predictio[illegible] of biomolecular interactions [illegible]lphaFold 3

https://doi.org/10.1038/s41586-024-07487-w
Received: 19 December 2023
Accepted: 29 April 2024
Published online: 8 May 2024
Open access
Check for updates

Josh Abramson[1,7], Jo[illegible] Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7] [illegible]say Willmore[1,7], Andrew J. Ballard[1], Joshua Bambri[illegible]ein[1], David A. Evans[1], Chia-Chun Hung[2], Michael O'N[illegible]yn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arv[illegible]avia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles [illegible]owen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fab[illegible]dman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], [illegible]nko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], [illegible], Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], [illegible]gustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8], [illegible] & John M. Jumper[1,8]

[illegible]oduction of AlphaFold 2[1] has spurred a revolution in modelling the structure [illegible]oteins and their interactions, enabling a huge range of applications in protein [illegible]odelling and design[2–6]. Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved accuracy

Impose symmetries or learn them?

Article

Accurate structure predictio of biomolecular interactions phaFold 3

No Equivariance!

https://doi.org/10.1038/s41586-024-07487-w
Received: 19 December 2023
Accepted: 29 April 2024
Published online: 8 May 2024
Open access
Check for updates

Josh Abramson[1,7], Jo Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7] say Willmore[1,7], Andrew J. Ballard[1], Joshua Bambri in[1], David A. Evans[1], Chia-Chun Hung[2], Michael O'N yn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arv avia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles wen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fab dman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], nko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], gustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8], & John M. Jumper[1,8]

oduction of AlphaFold 2[1] has spurred a revolution in modelling the structure oteins and their interactions, enabling a huge range of applications in protein odelling and design[2–6]. Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved accuracy

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric Qu
UC Berkeley
ericqu@berkeley.edu

Aditi S. Krishnapriyan
UC Berkeley, LBNL
aditik1@berkeley.edu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. It involves how a model's performance changes with increases in model size or input data, as well as how efficiently computational resources are utilized to support this growth. Despite successes in scaling other types of machine learning models, the study of scaling in Neural Network Interatomic Potentials (NNIPs) remains limited. NNIPs act as

Impose symmetries or learn them?

Article

Accurate structure predictio of biomolecular interactions phaFold 3

No Equivariance!

https://doi.org/10.1038/s41586-024-07487-w
Received: 19 December 2023
Accepted: 29 April 2024
Published online: 8 May 2024
Open access
Check for updates

Josh Abramson[1,7], Jon Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7] say Willmore[1,7], Andrew J. Ballard[1], Joshua Bambri ein[1], David A. Evans[1], Chia-Chun Hung[2], Michael O'N yn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arv avia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles C owen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fab dman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], nko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], gustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8], & John M. Jumper[1,8]

oduction of AlphaFold 2[1] has spurred a revolution in modelling the structure oteins and their interactions, enabling a huge range of applications in protein odelling and design[2–6]. Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved accuracy

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric Qu
UC Berkeley
ericqu@berkeley.edu

Aditi S. Krishnapriyan
UC Berkeley, LBNL
aditik1@berkeley.edu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. It involves how a model's performance changes with increases in model size or input data, as well as how efficiently computational resources are utilized to support this growth. Despite successes in scaling other types of machine learning models, the study of scaling in Neural Network Interatomic Potentials (NNIPs) remains limited. NNIPs act as

Swallowing the Bitter Pill: Simplified Scalable Conformer Generation

Yuyang Wang[1] Ahmed A. Elhag[1,2] Navdeep Jaitly[1] Joshua M. Susskind[1] Miguel Ángel Bautista[1]

Abstract

We present a novel way to predict molecular conformers through a simple formulation that sidesteps many of the heuristics of prior works and achieves state of the art results by using the advantages of scale. By training a diffusion generative model directly on 3D atomic positions without making assumptions about the explicit structure of molecules (*e.g.* modeling torsional angles) we are able to radically simplify struc-

is the vast complexity of the 3D structure space, encompassing factors such as bond lengths and torsional angles. Despite the molecular graph dictating potential 3D conformers through specific constraints, such as bond types and spatial arrangements determined by chiral centers, the conformational space experiences exponential growth with the expansion of the graph size and the number of rotatable bonds (Axelrod & Gomez-Bombarelli, 2022). This complicates brute force and exhaustive approaches, making them virtually unfeasible for even moderately small molecules.

Systematic methods, like OMEGA (Hawkins et al., 2010),

Impose symmetries or learn them?

Article

Accurate structure prediction of biomolecular interactions ... AlphaFold 3

No Equivariance!

https://doi.org/10.1038/s41586-024-07487-w
Received: 19 December 2023
Accepted: 29 April 2024
Published online: 8 May 2024
Open access
Check for updates

Josh Abramson[1,7], Jon... Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7], ... Lindsay Willmore[1,7], Andrew J. Ballard[1], Joshua Bambrick... ...ein[1], David A. Evans[1], Chia-Chun Hung[2], Michael O'N... ...yn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arv... ...avia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles C... ...owen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fabi... ...dman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], K... ...nko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], ... Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], ...gustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8], ... & John M. Jumper[1,8]

...oduction of AlphaFold 2[1] has spurred a revolution in modelling the structure ...oteins and their interactions, enabling a huge range of applications in protein ...odelling and design[2–6]. Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved accuracy

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric Qu
UC Berkeley
ericqu@berkeley.edu

Aditi S. Krishnapriyan
UC Berkeley, LBNL
aditik1@berkeley.edu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. It involves how a model's performance changes with increases in model size or input data, as well as how efficiently computational resources are utilized to support this growth. Despite successes in scaling other types of machine learning models, the study of scaling in Neural Network Interatomic Potentials (NNIPs) remains limited. NNIPs act as

Swallowing the Bitter Pill: Simplified Scalable Conformer Generation

Yuyang Wang[1] Ahmed A. Elhag[1,2] Navdeep Jaitly[1] Joshua M. Susskind[1] Miguel Ángel Bautista[1]

Abstract

We present a novel way to predict molecular conformers through a simple formulation that sidesteps many of the heuristics of prior works and achieves state of the art results by using the advantages of scale. By training a diffusion generative model directly on 3D atomic positions without making assumptions about the explicit structure of molecules (*e.g.* modeling torsional angles) we are able to radically simplify struc-

is the vast complexity of the 3D structure space, encompassing factors such as bond lengths and torsional angles. Despite the molecular graph dictating potential 3D conformers through specific constraints, such as bond types and spatial arrangements determined by chiral centers, the conformational space experiences exponential growth with the expansion of the graph size and the number of rotatable bonds (Axelrod & Gomez-Bombarelli, 2022). This complicates brute force and exhaustive approaches, making them virtually unfeasible for even moderately small molecules.

Systematic methods, like OMEGA (Hawkins et al., 2010),

Probing the effects of broken symmetries in machine learning

Marcel F Langer, Sergey N Pozdnyakov and Michele Ceriotti*

Laboratory of Computational Science and Modeling and National Centre for Computational Design and Discovery of Novel Materials MARVEL, Institute of Materials, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
* Author to whom any correspondence should be addressed.

E-mail: michele.ceriotti@epfl.ch

Keywords: machine learning, symmetry-constrained models, atomistic modeling, molecular simulations

Supplementary material for this article is available online

Abstract

Symmetry is one of the most central concepts in physics, and it is no surprise that it has also been widely adopted as an inductive bias for machine-learning models applied to the physical sciences. This is especially true for models targeting the properties of matter at the atomic scale. Both established and state-of-the-art approaches, with almost no exceptions, are built to be exactly equivariant to translations, permutations, and rotations of the atoms. Incorporating symmetries—rotations in particular—constrains the model design space and implies more complicated architectures that are often also computationally demanding. There are indications

Impose symmetries or learn them?

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

No Equivariance!

https://doi.org/10.1038/s41586-024-07487-w
Received: 19 December 2023
Accepted: 29 April 2024
Published online: 8 May 2024
Open access
Check for updates

Josh Abramson[1,7], Jonas Adler[1,7], Jack Dunger[1,7], Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7], Olaf Ronneberger[1,7], Lindsay Willmore[1,7], Andrew J. Ballard[1], Joshua Bambrick[2], Sebastian W. Bodenstein[1], David A. Evans[1], Chia-Chun Hung[2], Michael O'Neill[1], David Reiman[1], Kathryn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arvaniti[3], Charles Beattie[3], Ottavia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles Congreve[4], Alexander I. Cowen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fabian B. Fuchs[3], Hannah Gladman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], Kuba Perlin[3], Anna Potapenko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], Ashok Thillaisundaram[3], Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], Michal Zielinski[3], Augustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8], Demis Hassabis[1,2,8] & John M. Jumper[1,8]

The introduction of AlphaFold 2[1] has spurred a revolution in modelling the structure of proteins and their interactions, enabling a huge range of applications in protein modelling and design[2–6]. Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved accuracy

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric Qu
UC Berkeley
ericqu@berkeley.edu

Aditi S. Krishnapriyan
UC Berkeley, LBNL
aditik1@berkeley.edu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. It involves how a model's performance changes with increases in model size or input data, as well as how efficiently computational resources are utilized to support this growth. Despite successes in scaling other types of machine learning models, the study of scaling in Neural Network Interatomic Potentials (NNIPs) remains limited. NNIPs act as

Swallowing the Bitter Pill: Simplified Scalable Conformer Generation

Yuyang Wang[1] Ahmed A. Elhag[1,2] Navdeep Jaitly[1] Joshua M. Susskind[1] Miguel Ángel Bautista[1]

Abstract

We present a novel way to predict molecular conformers through a simple formulation that sidesteps many of the heuristics of prior works and achieves state of the art results by using the advantages of scale. By training a diffusion generative model directly on 3D atomic positions without making assumptions about the explicit structure of molecules (*e.g.* modeling torsional angles) we are able to radically simplify structure learning, and make it trivial to scale up the

is the vast complexity of the 3D structure space, encompassing factors such as bond lengths and torsional angles. Despite the molecular graph dictating potential 3D conformers through specific constraints, such as bond types and spatial arrangements determined by chiral centers, the conformational space experiences exponential growth with the expansion of the graph size and the number of rotatable bonds (Axelrod & Gomez-Bombarelli, 2022). This complicates brute force and exhaustive approaches, making them virtually unfeasible for even moderately small molecules.

Systematic methods, like OMEGA (Hawkins et al., 2010),

Probing the effects of broken symmetries in machine learning

Marcel F Langer, Sergey N Pozdnyakov and Michele Ceriotti*
Laboratory of Computational Science and Modeling and National Centre for Computational Design and Discovery of Novel Materials MARVEL, Institute of Materials, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
* Author to whom any correspondence should be addressed.
E-mail: michele.ceriotti@epfl.ch

Keywords: machine learning, symmetry-constrained models, atomistic modeling, molecular simulations
Supplementary material for this article is available online

Abstract

Symmetry is one of the most central concepts in physics, and it is no surprise that it has also been widely adopted as an inductive bias for machine-learning models applied to the physical sciences. This is especially true for models targeting the properties of matter at the atomic scale. Both established and state-of-the-art approaches, with almost no exceptions, are built to be exactly equivariant to translations, permutations, and rotations of the atoms. Incorporating symmetries—rotations in particular—constrains the model design space and implies more complicated architectures that are often also computationally demanding. There are indications

Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics

Marloes Arts,[*,†,‡,@] Victor Garcia Satorras,[*,¶,@] Chin-Wei Huang,[¶] Daniel Zügner,[§] Marco Federici,[†,‖] Cecilia Clementi,[§,⊥] Frank Noé,[§] Robert Pinsler,[#] and Rianne van den Berg[¶]

†*Work done during an internship at Microsoft Research (Amsterdam).*
‡*University of Copenhagen, Department of Computer Science, Universitetsparken 1, Copenhagen, 2100, Denmark.*
¶*AI4Science, Microsoft Research, Evert van de Beekstraat 354, Amsterdam, 1118 CZ, The Netherlands.*
§*AI4Science, Microsoft Research, Karl-Liebknecht-Straße 32, Berlin, 10178, Germany.*
‖*University of Amsterdam, Informatics Institute, Science Park 904, Amsterdam, 1098 XH, The Netherlands.*
⊥*Freie Universität Berlin, Department of Physics, Arnimalle 12, Berlin, 14195, Germany.*
#*AI4Science, Microsoft Research, 21 Station Road, Cambridge, CB1 2FB, United Kingdom.*
@*Equal contribution.*

* E-mail: ma@di.ku.dk; victorgar@microsoft.com

Abstract

Coarse-grained (CG) molecular dynamics enables the study of biological processes at temporal and spatial scales that would be intractable at an atomistic resolution. However, accurately learning a CG force field remains a challenge. In this work, we leverage connections between score-based generative models, force fields and molecular

Impose symmetries or learn them?

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

No Equivariance!

https://doi.org/10.1038/s41586-024-07487-w
Received: 19 December 2023
Accepted: 29 April 2024
Published online: 8 May 2024
Open access
Check for updates

Josh Abramson[1,7], Jo… Richard Evans[1,7], Tim Green[1,7], Alexander Pritzel[1,7] … Lindsay Willmore[1,7], Andrew J. Ballard[1], Joshua Bambrick … in[1], David A. Evans[1], Chia-Chun Hung[2], Michael O'N… Kathryn Tunyasuvunakool[1], Zachary Wu[1], Akvilė Žemgulytė[1], Eirini Arv… Octavia Bertolli[3], Alex Bridgland[3], Alexey Cherepanov[4], Miles C… Bowen-Rivers[3], Andrew Cowie[3], Michael Figurnov[3], Fabi… dman[3], Rishub Jain[3], Yousuf A. Khan[3,5], Caroline M. R. Low[4], … ko[3], Pascal Savy[4], Sukhdeep Singh[3], Adrian Stecula[4], … Catherine Tong[4], Sergei Yakneen[4], Ellen D. Zhong[3,6], … Augustin Žídek[3], Victor Bapst[1,8], Pushmeet Kohli[1,8], Max Jaderberg[2,8], … & John M. Jumper[1,8]

The introduction of AlphaFold 2[1] has spurred a revolution in modelling the structure of proteins and their interactions, enabling a huge range of applications in protein modelling and design[2–6]. Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved accuracy

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric Qu
UC Berkeley
ericqu@berkeley.edu

Aditi S. Krishnapriyan
UC Berkeley, LBNL
aditik1@berkeley.edu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. It involves how a model's performance changes with increases in model size or input data, as well as how efficiently computational resources are utilized to support this growth. Despite successes in scaling other types of machine learning models, the study of scaling in Neural Network Interatomic Potentials (NNIPs) remains limited. NNIPs act as

Swallowing the Bitter Pill: Simplified Scalable Conformer Generation

Yuyang Wang[1] Ahmed A. Elhag[1,2] Navdeep Jaitly[1] Joshua M. Susskind[1] Miguel Ángel Bautista[1]

Abstract

We present a novel way to predict molecular conformers through a simple formulation that sidesteps many of the heuristics of prior works and achieves state of the art results by using the advantages of scale. By training a diffusion generative model directly on 3D atomic positions without making assumptions about the explicit structure of molecules (*e.g.* modeling torsional angles) we are able to radically simplify struc-

is the vast complexity of the 3D structure space, encompassing factors such as bond lengths and torsional angles. Despite the molecular graph dictating potential 3D conformers through specific constraints, such as bond types and spatial arrangements determined by chiral centers, the conformational space experiences exponential growth with the expansion of the graph size and the number of rotatable bonds (Axelrod & Gomez-Bombarelli, 2022). This complicates brute force and exhaustive approaches, making them virtually unfeasible for even moderately small molecules.

Systematic methods, like OMEGA (Hawkins et al., 2010)

Probing the effects of broken symmetries in machine learning

Marcel F Langer, Sergey N Pozdnyakov and Michele Ceriotti*

Laboratory of Computational Science and Modeling and National Centre for Computational Design and Discovery of Novel Materials MARVEL, Institute of Materials, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
* Author to whom any correspondence should be addressed.

E-mail: michele.ceriotti@epfl.ch

Keywords: machine learning, symmetry-constrained models, atomistic modeling, molecular simulations

Supplementary material for this article is available online

Abstract

Symmetry is one of the most central concepts in physics, and it is no surprise that it has also been widely adopted as an inductive bias for machine-learning models applied to the physical sciences. This is especially true for models targeting the properties of matter at the atomic scale. Both established and state-of-the-art approaches, with almost no exceptions, are built to be exactly equivariant to translations, permutations, and rotations of the atoms. Incorporating symmetries—rotations in particular—constrains the model design space and implies more complicated architectures that are often also computationally demanding. There are indications

Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics

Marloes Arts,*,†,‡,@ Victor Garcia Satorras,*,¶,@ Chin-Wei Huang,¶ Daniel Zügner,§ Marco Federici,†,|| Cecilia Clementi,§,⊥ Frank Noé,§ Robert Pinsler,# and Rianne van den Berg¶

†*Work done during an internship at Microsoft Research (Amsterdam).*
‡*University of Copenhagen, Department of Computer Science, Universitetsparken 1, Copenhagen, 2100, Denmark.*
¶*AI4Science, Microsoft Research, Evert van de Beekstraat 354, Amsterdam, 1118 CZ, The Netherlands.*
§*AI4Science, Microsoft Research, Karl-Liebknecht-Straße 32, Berlin, 10178, Germany.*
||*University of Amsterdam, Informatics Institute, Science Park 904, Amsterdam, 1098 XH, The Netherlands.*
⊥*Freie Universität Berlin, Department of Physics, Arnimalle 12, Berlin, 14195, Germany.*
#*AI4Science, Microsoft Research, 21 Station Road, Cambridge, CB1 2FB, United Kingdom.*
@*Equal contribution.*

* E-mail: ma@di.ku.dk; victorgar@microsoft.com

Abstract

Coarse-grained (CG) molecular dynamics enables the study of biological processes at temporal and spatial scales that would be intractable at an atomistic resolution. However, accurately learning a CG force field remains a challenge. In this work, we leverage connections between score-based generative models, force fields and molecular

DOES EQUIVARIANCE MATTER AT SCALE?

Johann Brehmer* Sönke Behrends Pim de Haan* Taco Cohen*
Qualcomm AI Research†
mail@johannbrehmer.de

ABSTRACT

Given large data sets and sufficient compute, is it beneficial to design neural architectures for the structure and symmetries of each problem? Or is it more efficient to learn them from data? We study empirically how equivariant and non-equivariant networks scale with compute and training samples. Focusing on a benchmark problem of rigid-body interactions and on general-purpose transformer architectures, we perform a series of experiments, varying the model size, training steps, and dataset size. We find evidence for three conclusions. First, equivariance improves data efficiency, but training non-equivariant models with data augmentation can close this gap given sufficient epochs. Second, scaling with compute follows a power law, with equivariant models outperforming non-equivariant ones at each tested compute budget. Finally, the optimal allocation of a compute budget onto model size and training duration differs between equivariant and non-equivariant models.

Impose symmetries or learn them?

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

No Equivariance!

https://doi.org/10.1038/s41586-024-07487-w
Received: 19 December 2023
Accepted: 29 April 2024
Published online: 8 May 2024
Open access
Check for updates

Josh Abramson, ... Richard Evans, Tim Green, ... Alexander Pritzel, ... Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, ... David A. Evans, Chia-Chun Hung, Michael O'Neill, ... Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, ... Olivia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, ... Owen-Rivers, Andrew Cowie, Michael Figurnov, ... Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, ... Pascal Savy, Sukhdeep Singh, Adrian Stecula, ... Catherine Tong, Sergei Yakneen, Ellen D. Zhong, ... Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, ... & John M. Jumper

The introduction of AlphaFold 2 has spurred a revolution in modelling the structure of proteins and their interactions, enabling a huge range of applications in protein modelling and design. Here we describe our AlphaFold 3 model with a substantially updated diffusion-based architecture that is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions and modified residues. The new AlphaFold model demonstrates substantially improved ...

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric Qu
UC Berkeley
ericqu@berkeley.edu

Aditi S. Krishnapriyan
UC Berkeley, LBNL
aditik1@berkeley.edu

Swallowing the Bitter Pill: Simplified Scalable Conformer Generation

Yuyang Wang[1] Ahmed A. Elhag[1 2] Navdeep Jaitly[1] Joshua M. Susskind[1] Miguel Ángel Bautista[1]

Abstract

We present a novel way to predict molecular conformers through a simple formulation that ... many of the heuristics of prior works ... achieves state of the art results by using the ... advantages of scale. By training a diffusion generative model directly on 3D atomic positions ... making assumptions about the explicit ... of molecules (e.g. modeling torsional ...) we are able to radically simplify structure ...

is the vast complexity of the 3D structure space, encompassing factors such as bond lengths and torsional angles. Despite the molecular graph dictating potential 3D conformers through specific constraints, such as bond types and spatial arrangements determined by chiral centers, the conformational space experiences exponential growth with the expansion of the graph size and the number of rotatable bonds (Axelrod & Gomez-Bombarelli, 2022). This complicates brute force and exhaustive approaches, making them virtually unfeasible for even moderately small molecules.

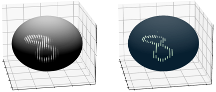
Systematic methods, like OMEGA (Hawkins et al., 2010)

Equivariance versus Augmentation for Spherical Images

Jan E. Gerken[1 2 3] Oscar Carlsson[1] Hampus Linander[4] Fredrik Ohlsson[5] Christoffer Petersson[6 1] Daniel Persson[1]

Abstract

We analyze the role of rotational equivariance in convolutional neural networks (CNNs) applied to spherical images. We compare the performance of the group equivariant networks known as S2CNNs and standard non-equivariant CNNs trained with an increasing amount of data augmentation. The chosen architectures can be consid-



Probing the effects of broken symmetries in machine learning

Marcel F Langer, Sergey N Pozdnyakov and Michele Ceriotti*

Laboratory of Computational Science and Modeling and National Centre for Computational Design and Discovery of Novel Materials MARVEL, Institute of Materials, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
* Author to whom any correspondence should be addressed.

E-mail: michele.ceriotti@epfl.ch

Keywords: machine learning, symmetry-constrained models, atomistic modeling, molecular simulations

Supplementary material for this article is available online

Abstract

Symmetry is one of the most central concepts in physics, and it is no surprise that it has also been widely adopted as an inductive bias for machine-learning models applied to the physical sciences. This is especially true for models targeting the properties of matter at the atomic scale. Both established and state-of-the-art approaches, with almost no exceptions, are built to be exactly equivariant to translations, permutations, and rotations of the atoms. Incorporating symmetries—rotations in particular—constrains the model design space and implies more complicated architectures that are often also computationally demanding. There are indications that unconstrained models can easily learn symmetries from data, and that doing so can even be ...

Netherlands.
§AI4Science, Microsoft Research, Karl-Liebknecht-Straße 32, Berlin, 10178, Germany.
‖University of Amsterdam, Informatics Institute, Science Park 904, Amsterdam, 1098 XH, The Netherlands.
⊥Freie Universität Berlin, Department of Physics, Arnimalle 12, Berlin, 14195, Germany.
#AI4Science, Microsoft Research, 21 Station Road, Cambridge, CB1 2FB, United Kingdom.
@Equal contribution.

* E-mail: ma@di.ku.dk; victorgar@microsoft.com

Abstract

Coarse-grained (CG) molecular dynamics enables the study of biological processes at temporal and spatial scales that would be intractable at an atomistic resolution. However, accurately learning a CG force field remains a challenge. In this work, we leverage connections between score-based generative models, force fields and molecular dynamics to learn a CG force field without requiring any force inputs during training ...

Does Equivariance Matter at Scale?

Johann Brehmer* Sönke Behrends Pim de Haan* Taco Cohen*
Qualcomm AI Research†
mail@johannbrehmer.de

Abstract

Given large data sets and sufficient compute, is it beneficial to design neural architectures for the structure and symmetries of each problem? Or is it more efficient to learn them from data? We study empirically how equivariant and non-equivariant networks scale with compute and training samples. Focusing on a benchmark problem of rigid-body interactions and on general-purpose transformer architectures, we perform a series of experiments, varying the model size, training steps, and dataset size. We find evidence for three conclusions. First, equivariance improves data efficiency, but training non-equivariant models with data augmentation can close this gap given sufficient epochs. Second, scaling with compute follows a power law, with equivariant models outperforming non-equivariant ones at each tested compute budget. Finally, the optimal allocation of a compute budget onto model size and training duration differs between equivariant and non-equivariant models.

Practitioners like data augmentation

- 👍 Easy to implement
- 👍 No specialized architecture necessary

Practitioners like data augmentation

- 👍 Easy to implement
- 👍 No specialized architecture necessary
- 👎 No exact equivariance

Practitioners like data augmentation

- 👍 Easy to implement
- 👍 No specialized architecture necessary
- 👎 No exact equivariance

Can we understand data augmentation theoretically?

Infinite-Width Networks

Empirical NTK

Training dynamics under continuous gradient descent:

$$\frac{\mathrm{d}\mathcal{N}_\theta(x)}{\mathrm{d}t} = -\frac{\eta}{N}\sum_{i=1}^{N} \Theta_\theta(x, x_i) \frac{\partial L}{\partial \mathcal{N}(x_i)}$$

Empirical NTK

Training dynamics under continuous gradient descent:

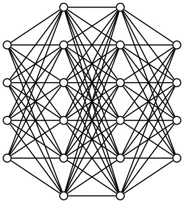
$$\frac{\mathrm{d}\mathcal{N}_\theta(x)}{\mathrm{d}t} = -\frac{\eta}{N}\sum_{i=1}^{N} \Theta_\theta(x, x_i) \frac{\partial L}{\partial \mathcal{N}(x_i)}$$

with the empirical neural tangent kernel (NTK)

$$\Theta_\theta(x, x') = \sum_\mu \frac{\partial \mathcal{N}(x)}{\partial \theta_\mu} \frac{\partial \mathcal{N}(x')}{\partial \theta_\mu}$$

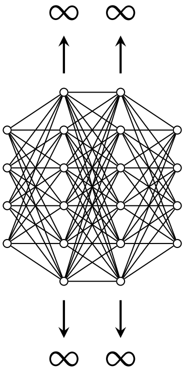
Infinite width limit

[Jacot et al. 2018]



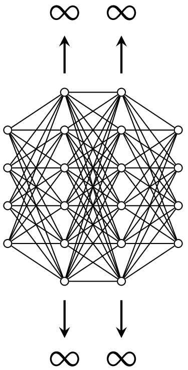
Infinite width limit

[Jacot et al. 2018]



Infinite width limit

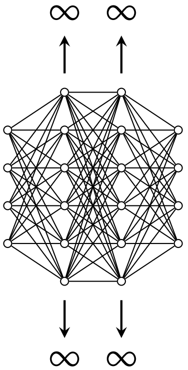
[Jacot et al. 2018]



👍 NTK becomes independent of initialization

Infinite width limit

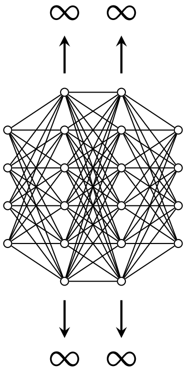
[Jacot et al. 2018]



- 👍 NTK becomes independent of initialization
- 👍 NTK becomes constant in training

Infinite width limit

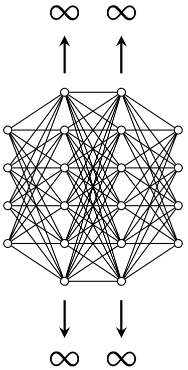
[Jacot et al. 2018]



- 👍 NTK becomes independent of initialization
- 👍 NTK becomes constant in training
- 👍 NTK can be computed for most networks

Infinite width limit

[Jacot et al. 2018]



- NTK becomes independent of initialization
- NTK becomes constant in training
- NTK can be computed for most networks
- ✓ Training dynamics can be solved

Mean prediction from NTK

[Jacot et al. 2018]

ⓘ At infinite width, the mean prediction is given by

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

Mean prediction from NTK

[Jacot et al. 2018]

At infinite width, the mean prediction is given by

neural tangent kernel

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

Mean prediction from NTK

[Jacot et al. 2018]

ⓘ At infinite width, the mean prediction is given by

neural tangent kernel

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

train data

Mean prediction from NTK

[Jacot et al. 2018]

ⓘ At infinite width, the mean prediction is given by

neural tangent kernel

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

learning rate

train data

Mean prediction from NTK

[Jacot et al. 2018]

At infinite width, the mean prediction is given by

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

neural tangent kernel

train labels

learning rate

train data

Data augmentation at infinite width

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

Data augmentation at infinite width

$$\mu_t(x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

augmented data

augmented labels

Data augmentation at infinite width

group transformation

$$\mu_t(\rho(g)x) = \Theta(\rho(g)x, X)\,\Theta(X, X)^{-1}\left(\mathbb{1} - e^{-\eta\Theta(X,X)t}\right)Y$$

augmented data

augmented labels

Data augmentation at infinite width

group transformation

for augmented data

$$\mu_t(\rho(g)x) = \Theta(\rho(g)x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})Y$$

augmented data

augmented labels

Data augmentation at infinite width

group transformation

$$\mu_t(\rho(g)x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\rho(g)Y$$

augmented data

augmented labels

Data augmentation at infinite width

group transformation

augmented labels

$$\mu_t(\rho(g)x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{1} - e^{-\eta\Theta(X,X)t})\underbrace{\rho(g)Y}_{=Y}$$

for invariance

Data augmentation at infinite width

group transformation

$$\mu_t(\rho(g)x) = \Theta(x, X)\Theta(X, X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\underbrace{\rho(g)Y}_{=Y}$$

$$= \mu_t(x)$$

for invariance

Mean prediction

$\mu_t(x)$

Mean prediction

$$\mu_t(x) = \mathbb{E}_{\theta_0 \sim \text{initializations}}[\mathcal{N}_{\theta_t}(x)]$$

Mean prediction

$$\mu_t(x) = \mathbb{E}_{\theta_0 \sim \text{initializations}}[\mathcal{N}_{\theta_t}(x)] = \lim_{n \to \infty} \frac{1}{n} \sum_{\theta_0 = \text{init}_1}^{\text{init}_n} \mathcal{N}_{\theta_t}(x)$$

Mean prediction

$$\mu_t(x) = \mathbb{E}_{\theta_0 \sim \text{initializations}}[\mathcal{N}_{\theta_t}(x)] = \lim_{n \to \infty} \underbrace{\frac{1}{n} \sum_{\theta_0 = \text{init}_1}^{\text{init}_n} \mathcal{N}_{\theta_t}(x)}_{\text{mean prediction of deep ensemble}}$$

Equivariant Ensembles

Ensembles of networks become
exactly equivariant under data augmentation

Emergent Equivariance in Deep Ensembles

Jan E. Gerken [*1] **Pan Kessel** [*2]

Abstract

We show that deep ensembles become equivariant for all inputs and at all training times by simply using full data augmentation. Crucially, equivariance holds off-manifold and for any architecture emergent: while the prediction of the ensemble is equivariant, the predictions of its members are not. In particular, the ensemble members are not required to have an equivariant architecture.

We rigorously derive this surprising emergent equivariance

Equivariant Ensembles

Ensembles of networks become
exactly equivariant under data augmentation

Emergent Equivariance in Deep Ensembles

Jan E. Gerken [*1] **Pan Kessel** [*2]

Abstract

We show that deep ensembles become equivariant for all inputs and at all training times by simply using full data augmentation. Crucially, equivariance holds off-manifold and for any architecture emergent: while the prediction of the ensemble is equivariant, the predictions of its members are not. In particular, the ensemble members are not required to have an equivariant architecture.

We rigorously derive this surprising emergent equivariance

Also true for finite-width ensembles: [Nordenfors, Flinth 2024]

Equivariant NTKs

Extend NTK framework to equivariant models

Equivariant Neural Tangent Kernels

Philipp Misof [a] Pan Kessel [b] Jan E. Gerken [a]

Abstract

Little is known about the training dynamics of equivariant neural networks, in particular how it compares to data augmented training of their non-equivariant counterparts. Recently, neural tangent kernels (NTKs) have emerged as a powerful tool to analytically study the training dynamics of wide neural networks. In this work, we take an important step towards a theoretical understanding of

Could show that an ensemble of augmented MLPs corresponds to an ensemble of GCNNs.

Methods from Physics

Non-Gaussian Corrections from Physics

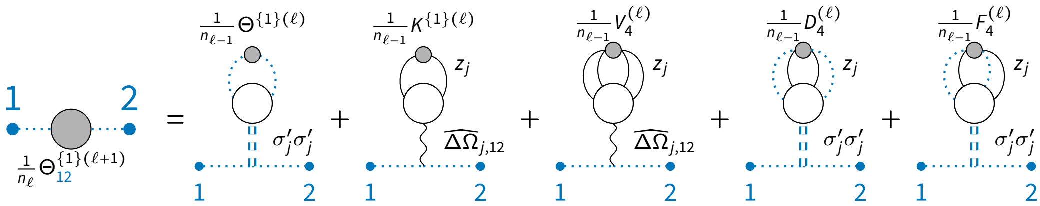
- 👎 Gaussians are limiting
- 👍 Taylor-expand in 1/width
- 👍 Use techniques from quantum field theory

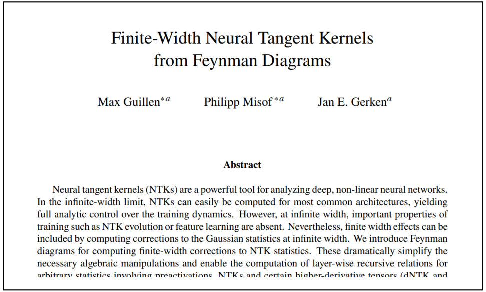
Non-Gaussian Corrections from Physics

- 👎 Gaussians are limiting
- 👍 Taylor-expand in 1/width
- 👍 Use techniques from quantum field theory

Neural Networks	Quantum Field Theory
infinite width	no interactions
Gaussian distribution	free fields
finite-width	interactions

Feynman diagrams for Neural Networks





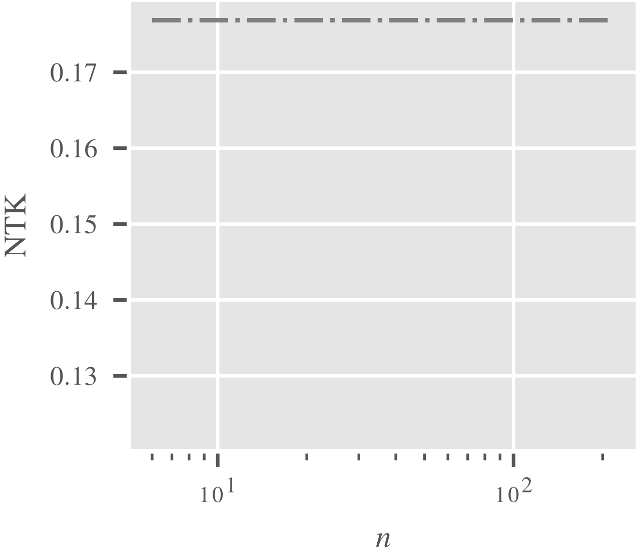
Finite-Width Neural Tangent Kernels from Feynman Diagrams

Max Guillen[*a] Philipp Misof[*a] Jan E. Gerken[a]

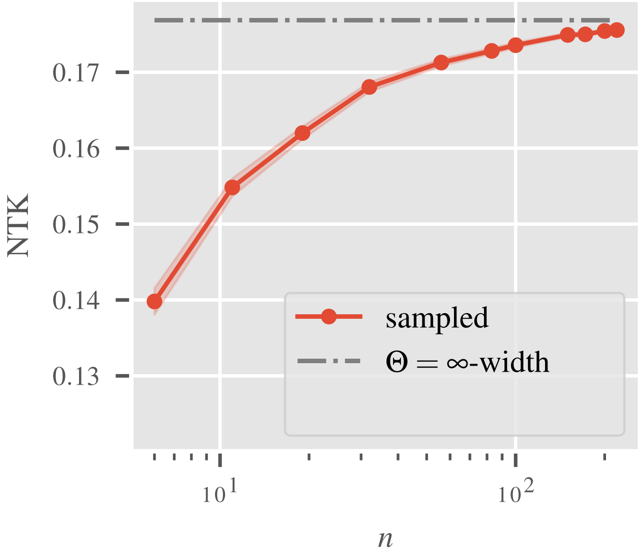
Abstract

Neural tangent kernels (NTKs) are a powerful tool for analyzing deep, non-linear neural networks. In the infinite-width limit, NTKs can easily be computed for most common architectures, yielding full analytic control over the training dynamics. However, at infinite width, important properties of training such as NTK evolution or feature learning are absent. Nevertheless, finite width effects can be included by computing corrections to the Gaussian statistics at infinite width. We introduce Feynman diagrams for computing finite-width corrections to NTK statistics. These dramatically simplify the necessary algebraic manipulations and enable the computation of layer-wise recursive relations for arbitrary statistics involving preactivations, NTKs and certain higher-derivative tensors (dNTK and

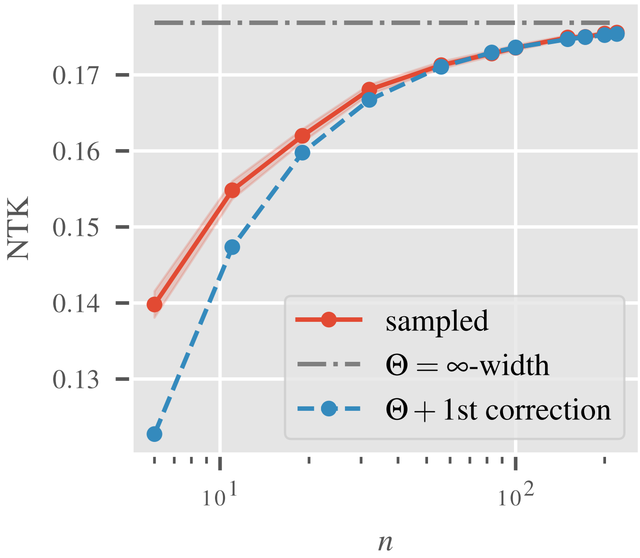
Finite-width corrections



Finite-width corrections



Finite-width corrections



Outlook

- Understanding training dynamics of neural networks remains challenging

Outlook

- Understanding training dynamics of neural networks remains challenging

⇨ Use symmetric case as simplified approach to study training

Outlook

- Understanding training dynamics of neural networks remains challenging

⇨ Use symmetric case as simplified approach to study training

- How does data augmentation affect the training dynamics?

Outlook

- Understanding training dynamics of neural networks remains challenging

⇨ Use symmetric case as simplified approach to study training

- How does data augmentation affect the training dynamics?

- How do the training dynamics of equivariant models differ from those of non-equivariant models?

Outlook

- Understanding training dynamics of neural networks remains challenging

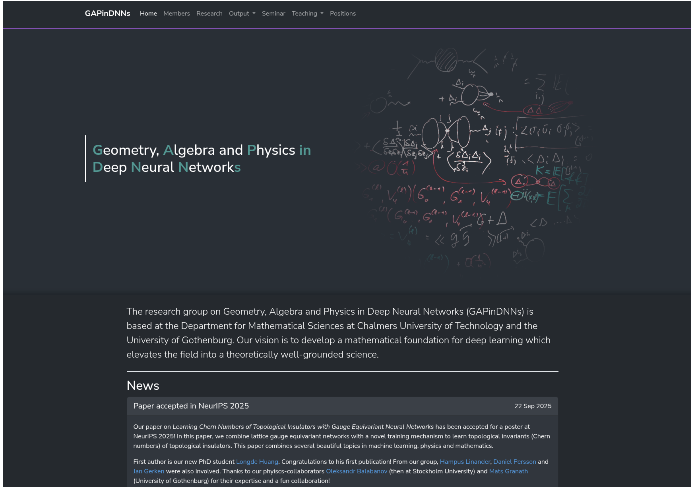
⇨ Use symmetric case as simplified approach to study training

- How does data augmentation affect the training dynamics?

- How do the training dynamics of equivariant models differ from those of non-equivariant models?

- Use insights from physics to make progress

Website



Thank you!