Toward the systematic generation of hypothetical atomic structures:

Neural networks and geometric motifs

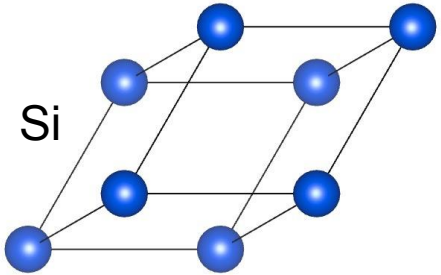
LBL CSSS Talk
2019.07.19

Tess Smidt
Luis W. Alvarez Postdoctoral Fellow
Computational Research Division
Lawrence Berkeley National Lab

What a computational materials physicist does:

Given an atomic structure,

Si

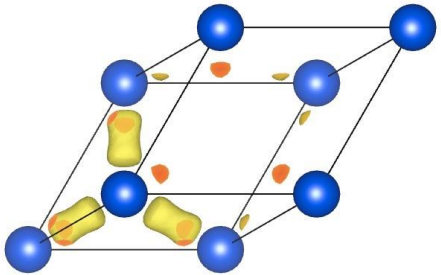


...use quantum theory and supercomputers to determine...

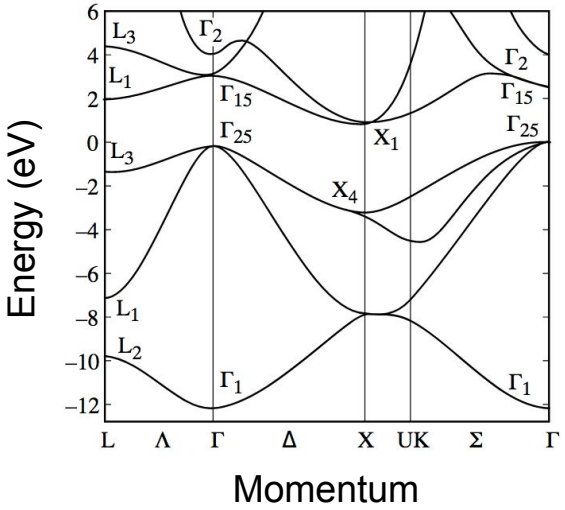
$$\hat{H}\,|\psi\rangle = E\,|\psi\rangle$$

NERSC
National Energy Research Scientific Computing Center

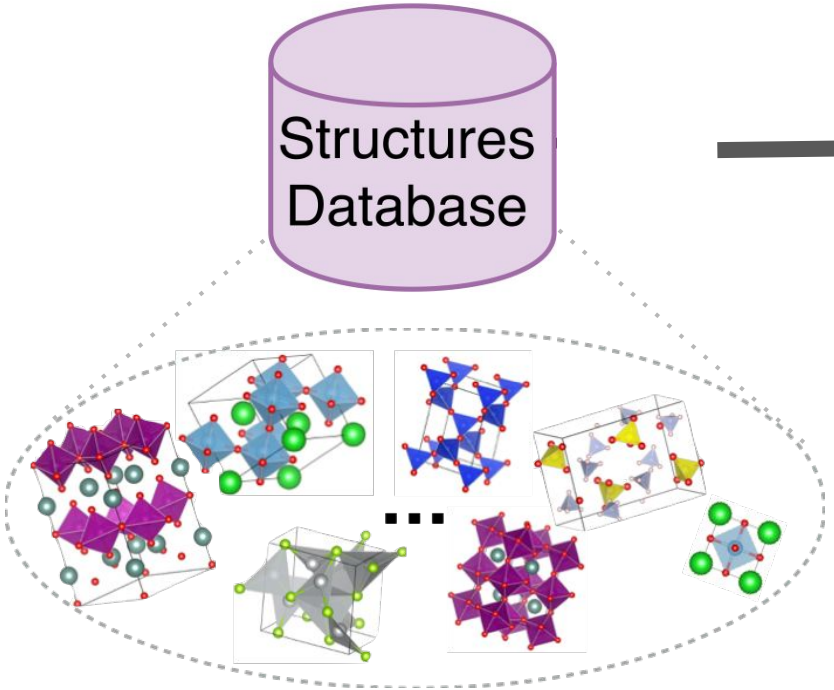
...where the electrons are...



...and what the electrons are doing.



Workflows are automated recipes that encode best practices for calculating materials properties. We use them to screen materials for specific properties and applications.



tens of thousands of structures

Properties

Elasticity
Thermal properties
Band gap
Electron mobility
Piezoelectricity
Polarization
...

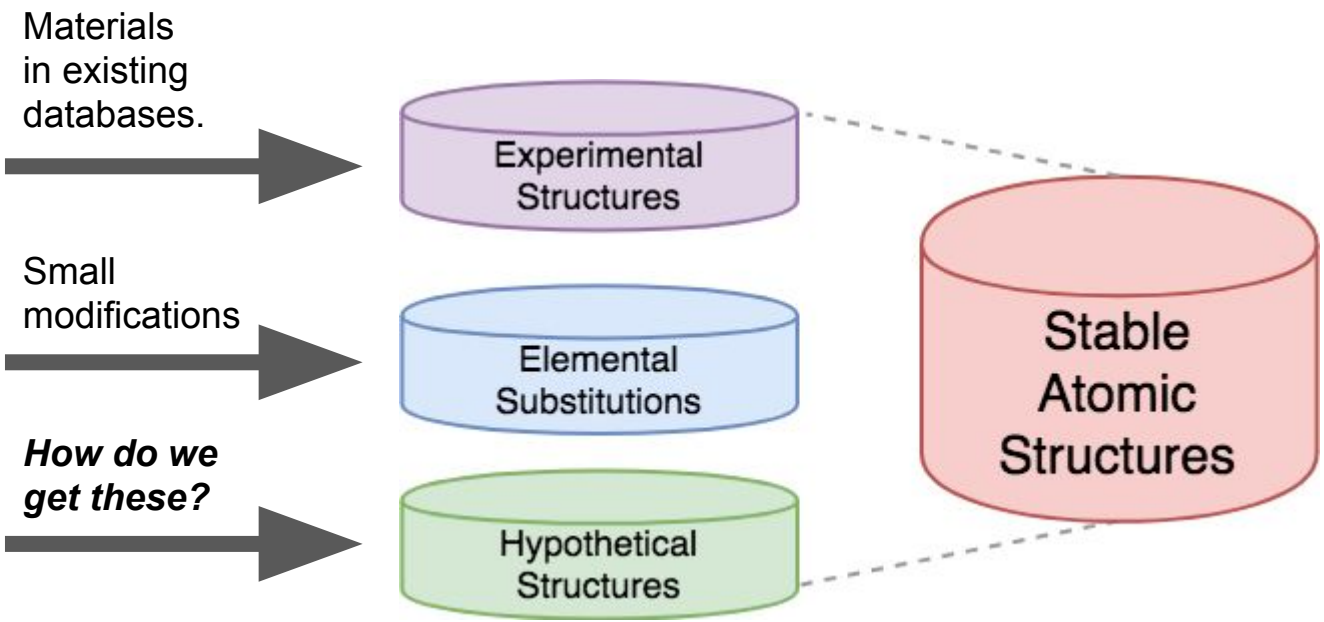
Photovoltaics

Magnetic materials

Batteries

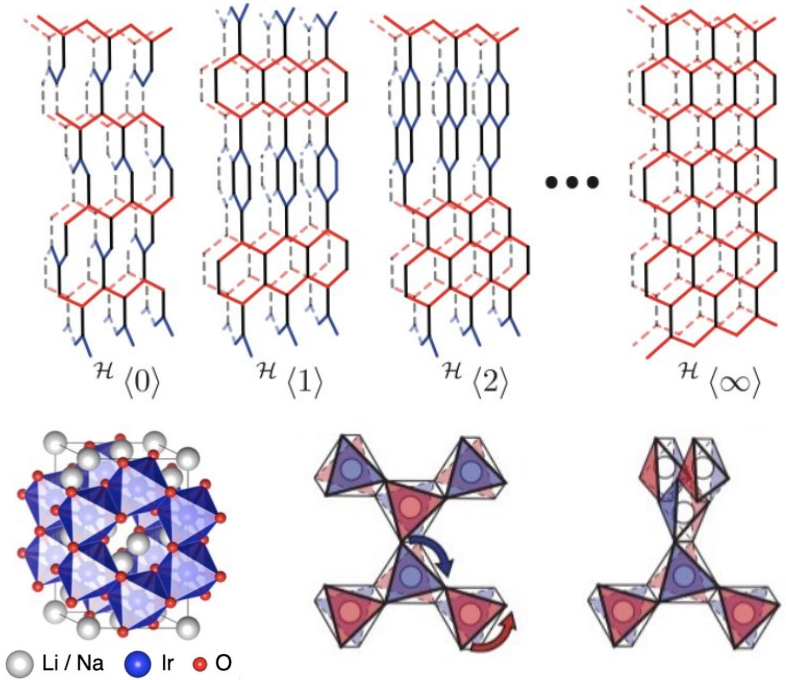
Ferroelectrics

However, screening is bottlenecked by our ability to propose hypothetical atomic structures.

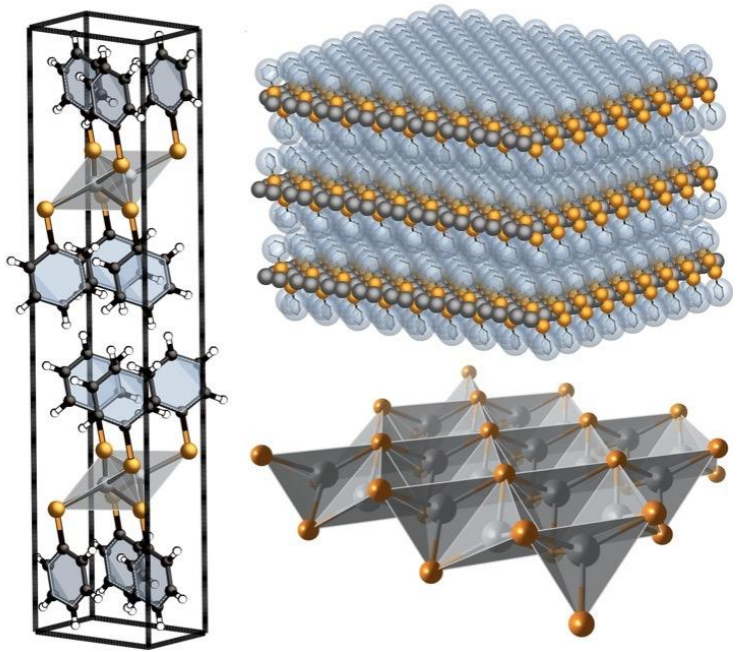


Experimentalists are making new structures every day!
These structures are not in existing databases.

Harmonic honeycomb iridates:
Frustrated quantum magnets



Metal-organic chalcogenide assemblies (MOChAs):
2D electronic properties in a 3D crystal



T. Smidt, S. Griffin, and J. B. Neaton, *Ab initio Studies of Structural and Energetic Trends in the Harmonic Honeycomb Iridates*, In preparation for submission to Physical Review: B (2018).

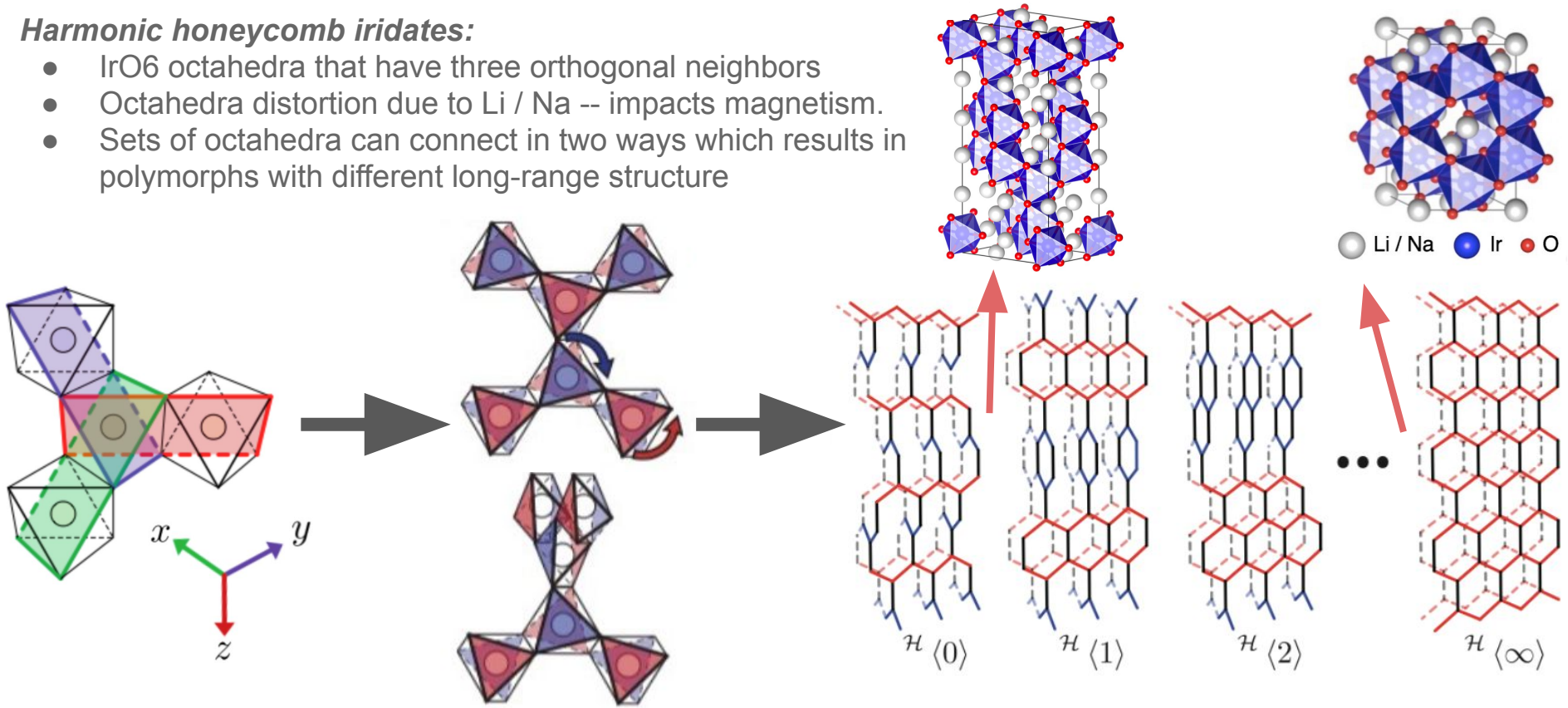
K. Modic, T. Smidt, I. Kimchi et al., *Realization of a three-dimensional spin-anisotropic harmonic honeycomb iridate*, Nature Communications 5 (2014). (arXiv:1402.3254)

J.N. Hohman, M. Collins, and T. Smidt, *Mithrene and methods of fabrication of mithrene*, (2017). International Patent App. PCT/US20I7/045609. Filed August 4, 2017.

Geometric motifs at different length scales determine electronic properties.

Harmonic honeycomb iridates:

- IrO6 octahedra that have three orthogonal neighbors
- Octahedra distortion due to Li / Na -- impacts magnetism.
- Sets of octahedra can connect in two ways which results in polymorphs with different long-range structure

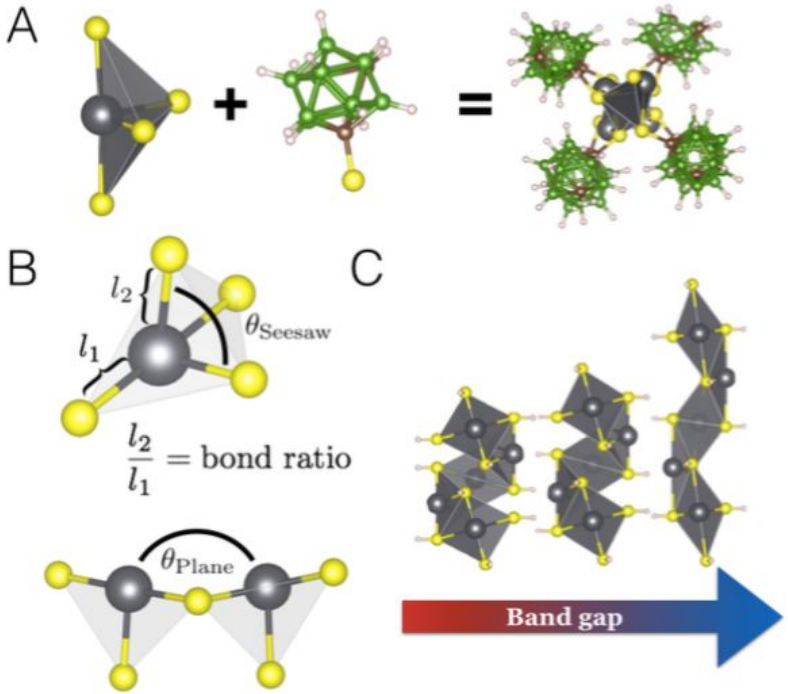


T. Smidt, S. Griffin, and J. B. Neaton, *Ab initio Studies of Structural and Energetic Trends in the Harmonic Honeycomb Iridates*, In preparation for submission to Physical Review: B (2018).

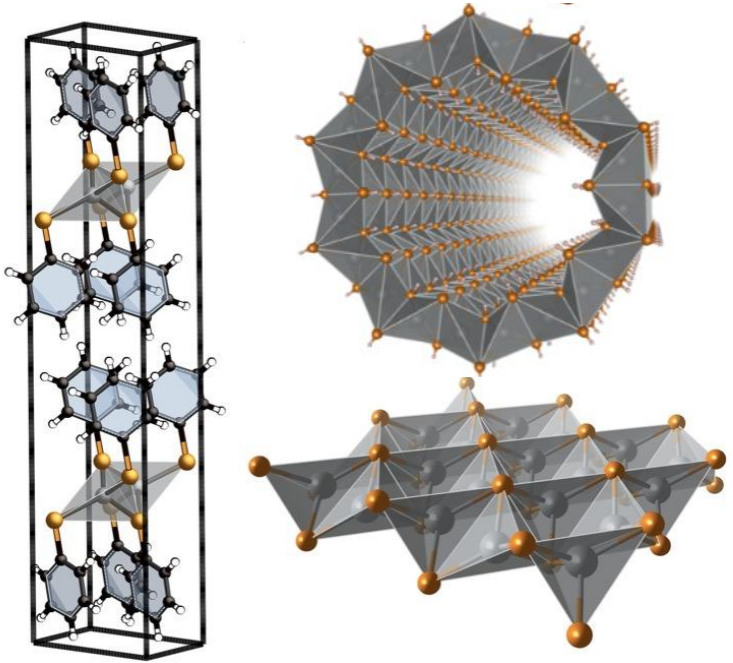
K. Modic, T. Smidt, I. Kimchi et al., *Realization of a three-dimensional spin-anisotropic harmonic honeycomb iridate*, Nature Communications 5 (2014). (arXiv:1402.3254)

Materials are challenging to design because their 3D geometry and interactions are complex.

Ex: Hypothetical materials that I designed by hand (with parametric models).



Distort subunits to tune properties.

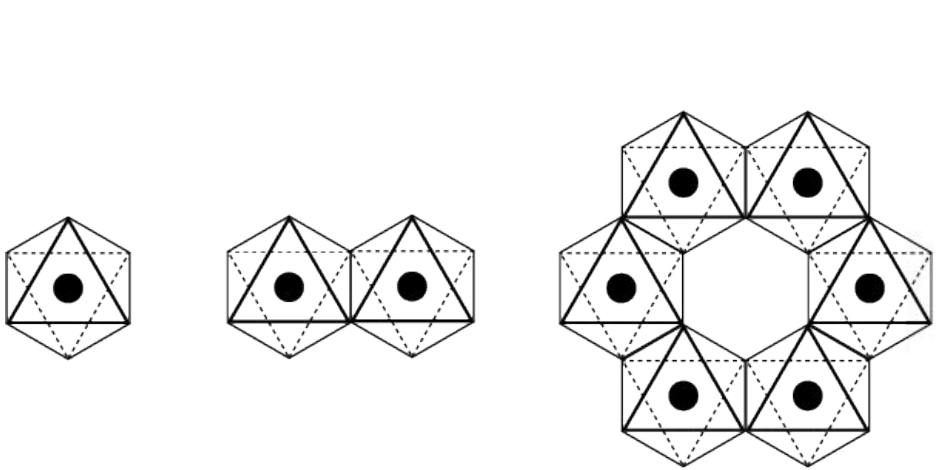


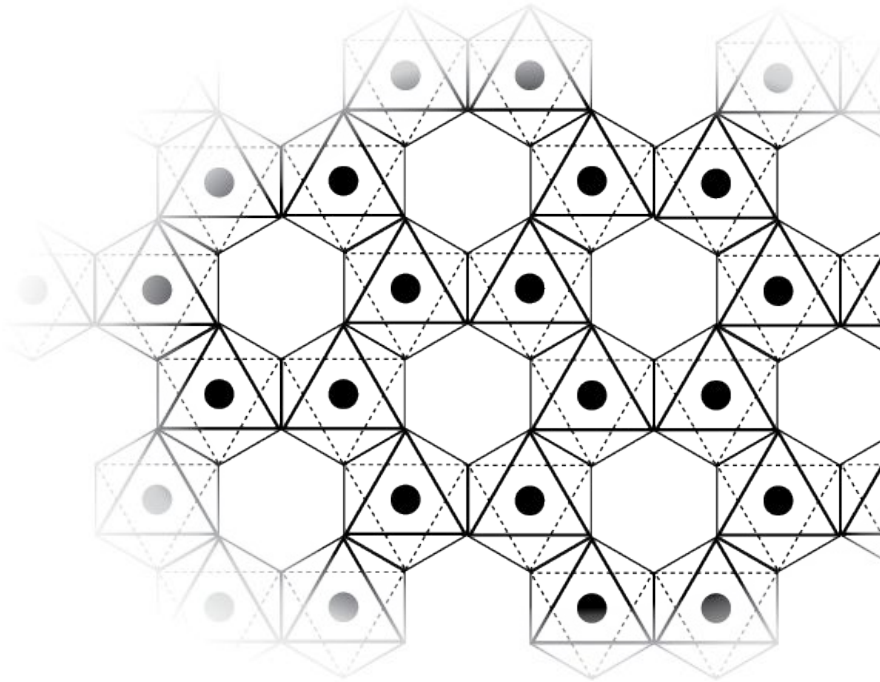
Produce new topologies that are chemically viable.

We need better tools to systematically generate and guide the design of new hypothetical atomic structures.

We need the right abstractions to design well.

The design space of stable atomic systems is much more limited than all possible arrangements of points in 3D space. Atoms in materials form geometric patterns and simple recurring arrangements.





Can we use patterns we've seen in existing materials to propose new structures that may be synthesized in the lab?

Deep learning shows promise for learning abstractions from data…

A brief primer on deep learning

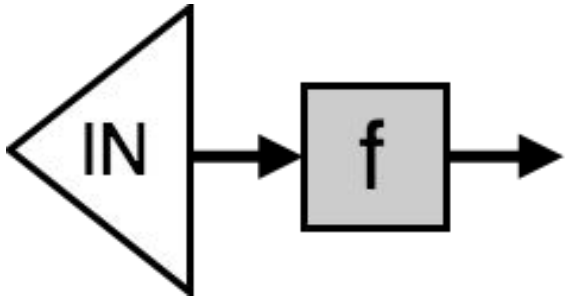
deep learning $\subset$ machine learning $\subset$ artificial intelligence

A brief primer on deep learning

model:

Function with learnable parameters.

$$y = f(x)$$



A brief primer on deep learning

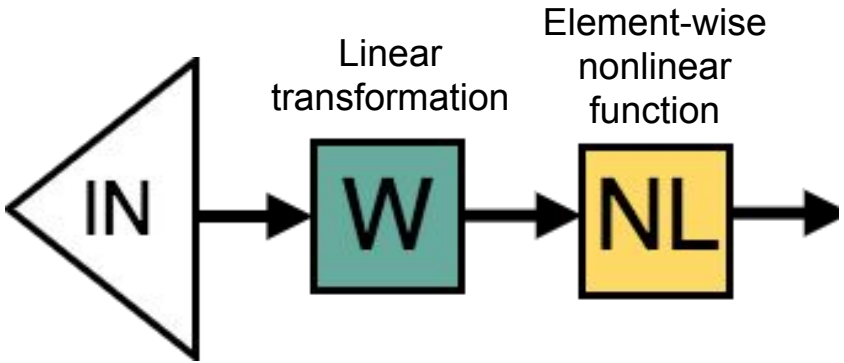
model:

Function with learnable parameters.

Ex: "Fully-connected" network

$$y = \tanh(Wx + b)$$

Learned Parameters



A brief primer on deep learning

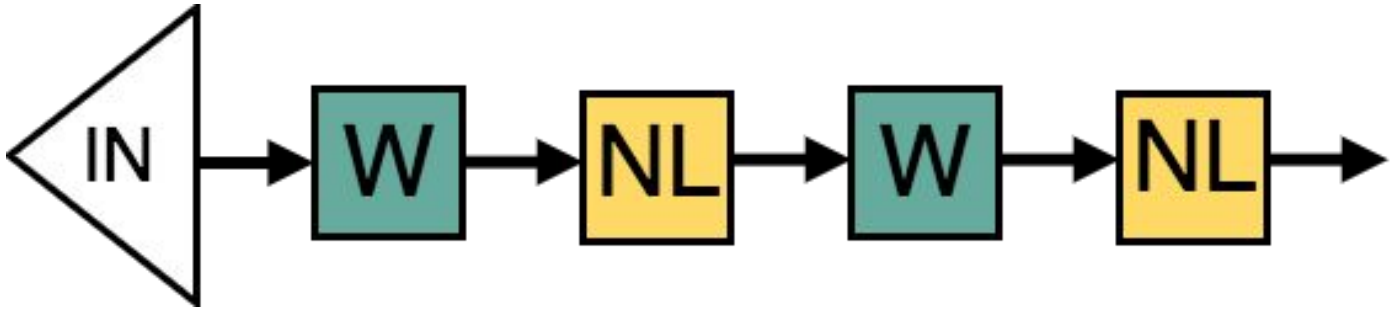
model:

Function with learnable parameters.

Ex: "Fully-connected" network

$$y = \tanh(W_2 \tanh(W_1 x + b_1) + b_2)$$

Learned Parameters

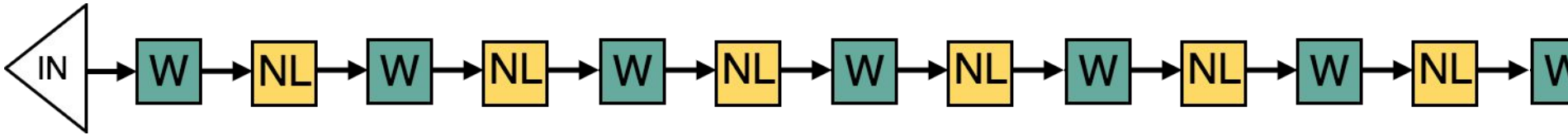


Neural networks with multiple layers can learn more complicated functions.

A brief primer on deep learning

deep learning:

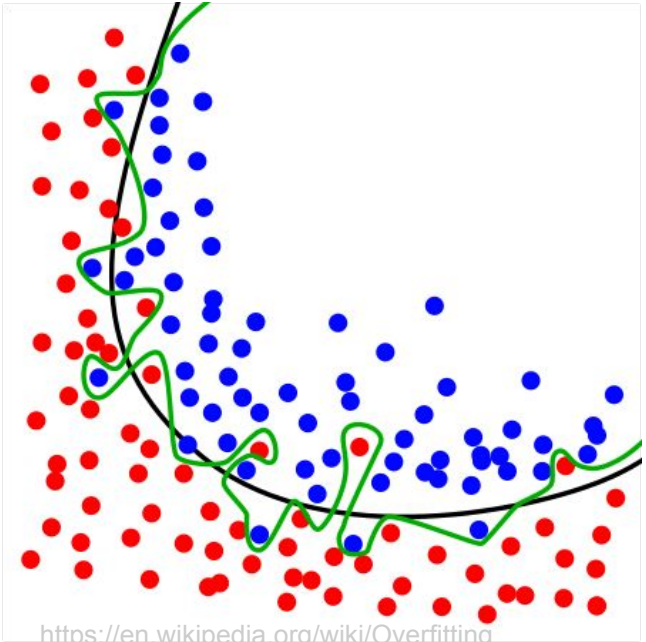
Add more layers.



A brief primer on deep learning

data:

Want lots of it. Model has many parameters. Don't want to easily overfit.



https://en.wikipedia.org/wiki/Overfitting

A brief primer on deep learning

cost function:

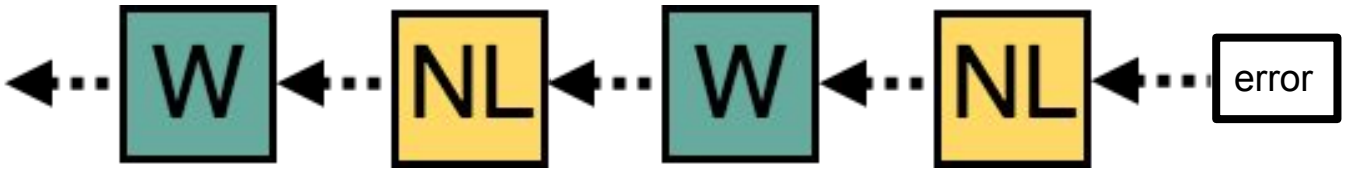
A metric to assess how well the model is performing.
The cost function is evaluated on the output of the model.
Also called the **loss** or **error**.

A brief primer on deep learning

way to update parameters:

Construct a model that is differentiable.
Take derivatives of the cost function (loss or error) wrt to learnable parameters.
This is called backpropogation (aka the chain rule).

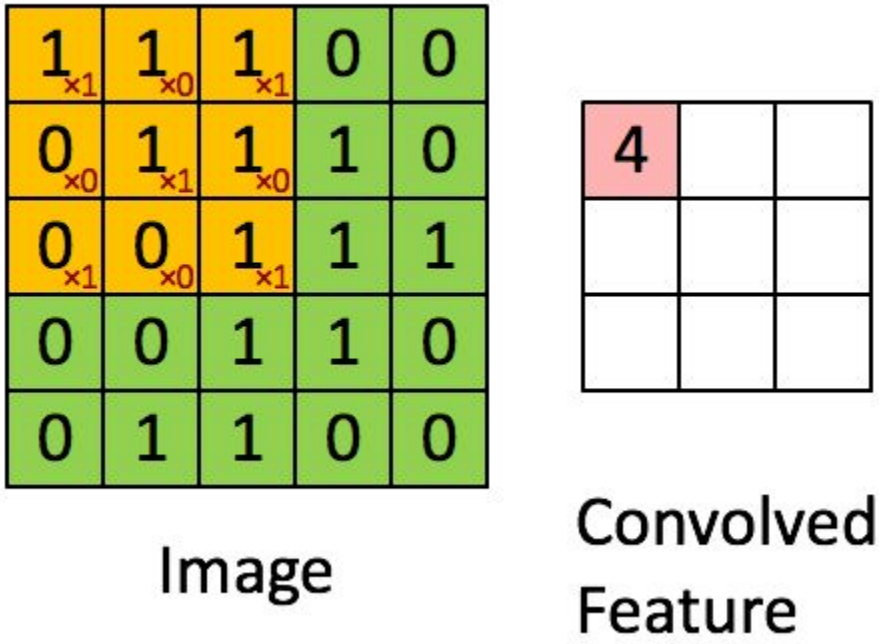
$$\Delta W_{ij} = -\eta \frac{\partial \text{ error}(f(W, x), y)}{\partial W_{ij}}$$



A brief primer on deep learning

convolutional neural networks:

Used for images. In each layer, scan over image with learned filters.



http://deeplearning.stanford.edu/wiki/index.php/Feature_extraction_using_convolution

A brief primer on deep learning

convolutional neural networks:

Used for images. In each layer, scan over image with learned filters.

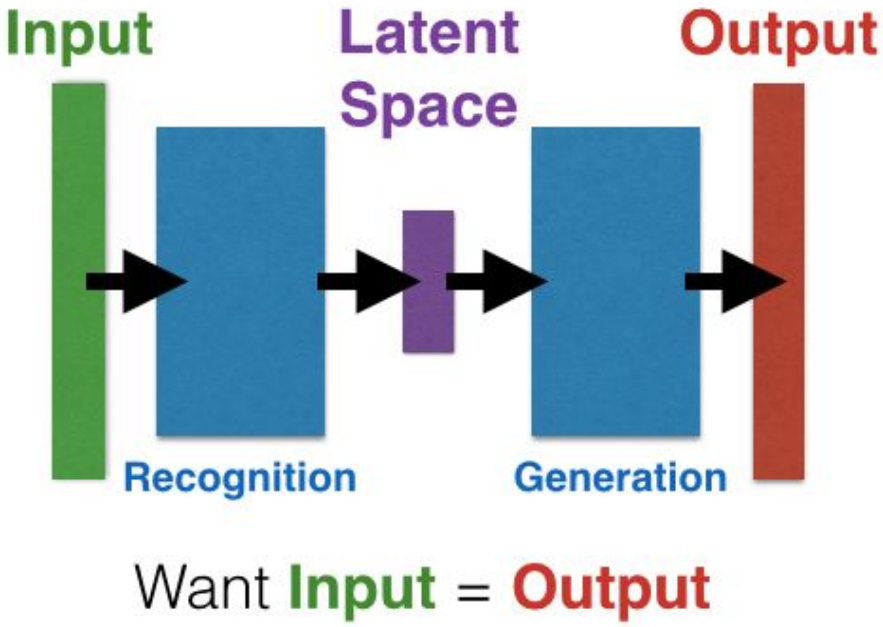
http://cs.nyu.edu/~fergus/tutorials/deep_learning_cvpr12/

Deep learning shows promise for learning abstractions from data…

Deep learning shows promise for learning abstractions from data…

Autoencoders can learn how map data in its original representation to a new representation and back again.

The learned representation is often very useful.

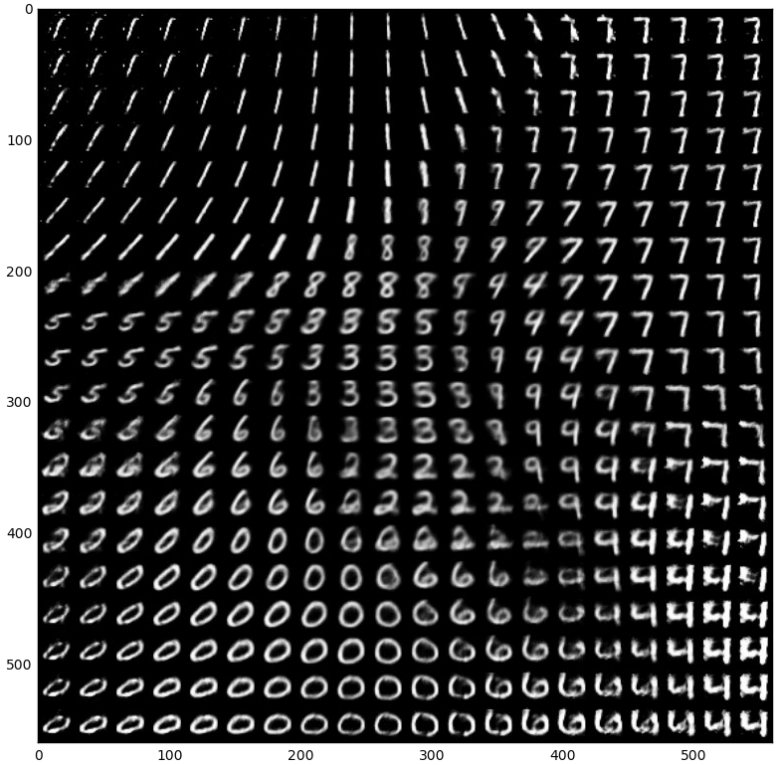


Latent space is either small or has a penalty to have a specified distribution.

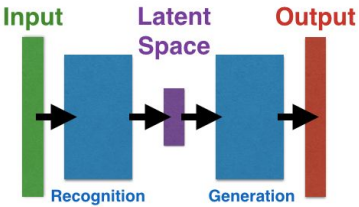
Deep learning shows promise for learning abstractions from data…

Example MNIST digits:

2 dimensional latent space for autoencoder trained on MNIST handwritten digit images

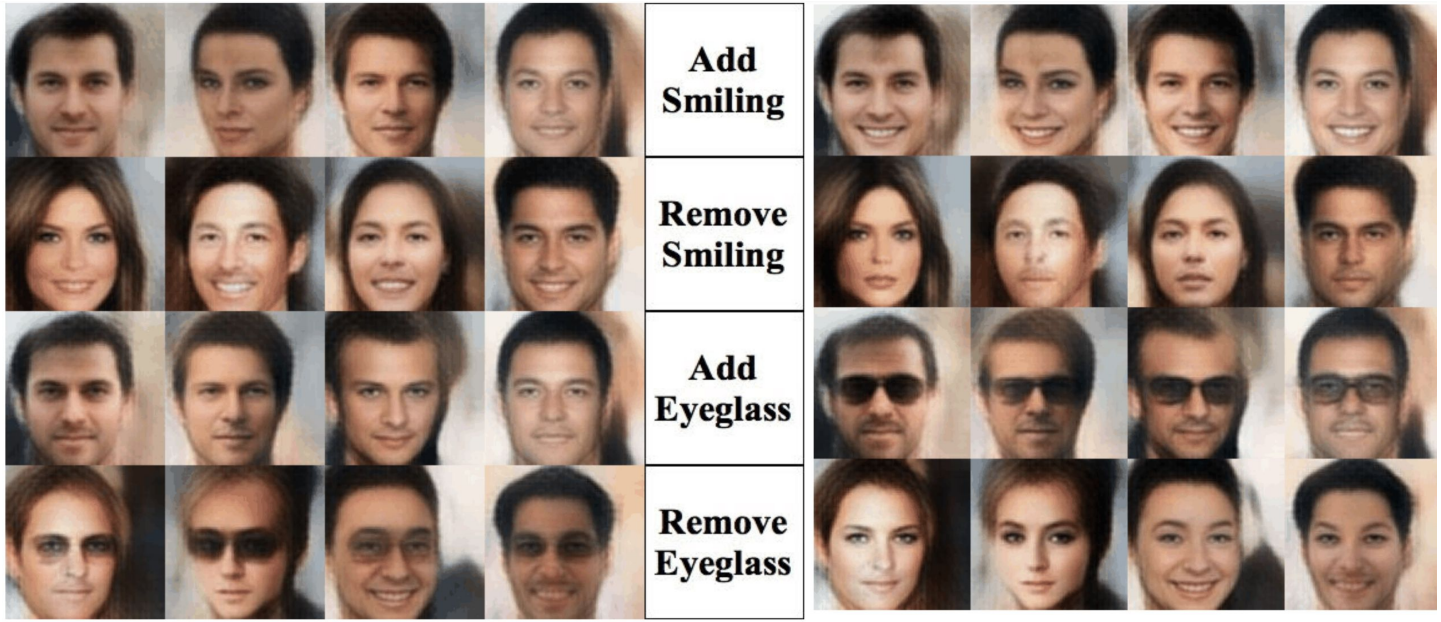


VAE Tutorial: https://jmetzen.github.io/2015-11-27/vae.html



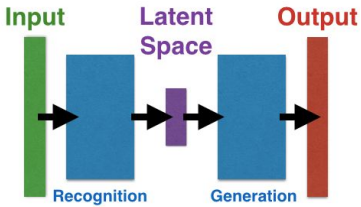
Want **Input** = **Output**

Deep learning shows promise for learning abstractions from data…



https://houxianxu.github.io/assets/project/dfcvae

https://twitter.com/smilevector



Want **Input** = **Output**

Deep learning shows promise for learning abstractions from data…
but it comes with significant challenges.

How do we represent atomic structures to neural networks?

Can we make neural networks that can understand symmetry and encode rich data types?

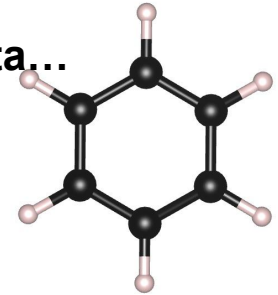
...

Deep learning shows promise for learning abstractions from data…
but it comes with significant challenges.

How do we represent atomic structures to neural networks?

Can we make neural networks that can understand symmetry and encode rich data types?

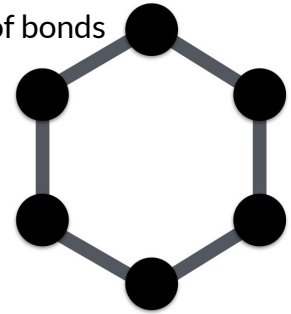
...



Take for example, benzene.

Vector (Fingerprint)

Graph of bonds



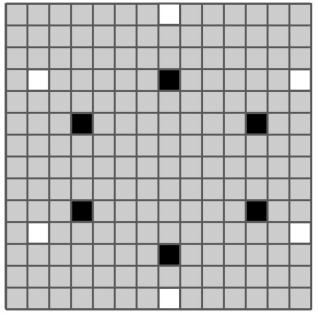
SMILES string

C1=CC=CC=C1

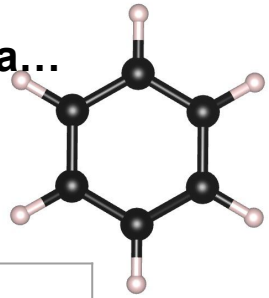
3D Coordinates

```
H   -0.21463    0.97837    0.33136
C   -0.38325    0.66317   -0.70334
C   -1.57552    0.03829   -1.05450
H   -2.34514   -0.13834   -0.29630
C   -1.78983   -0.36233   -2.36935
H   -2.72799   -0.85413   -2.64566
C   -0.81200   -0.13809   -3.33310
H   -0.98066   -0.45335   -4.36774
C    0.38026    0.48673   -2.98192
H    1.14976    0.66307   -3.74025
C    0.59460    0.88737   -1.66708
H    1.53276    1.37906   -1.39070
```

Image



Deep learning shows promise for learning abstractions from data…
but it comes with significant challenges.



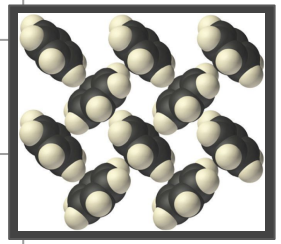
Take for example, benzene.

How do we represent atomic structures to neural networks?

Can we make neural networks that can understand symmetry and encode rich data types?

...

	Bonding	Geometry	Memory Efficient	Universality
Fingerprints	?	?	✓	?
SMILES	✓	X	✓	X
Graphs	✓	?	?	?
Images	X	✓	X	✓
Coordinates	X	✓	✓	✓



The most expressive data types require special treatment (custom networks)!

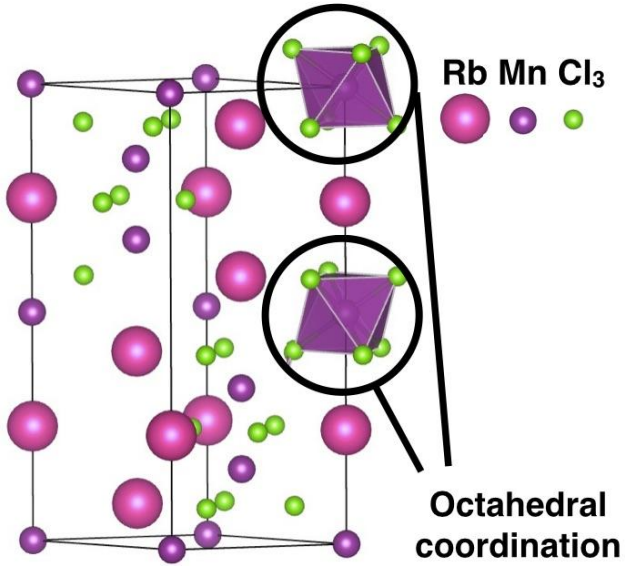
Graphs and coordinates have variable sizes.

Deep learning shows promise for learning abstractions from data…
but it comes with significant challenges.

How do we represent atomic structures to neural networks?

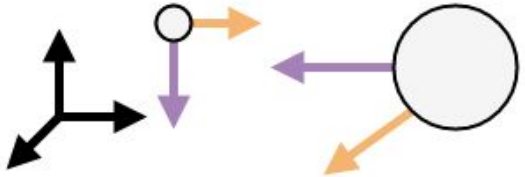
Can we make neural networks that can understand symmetry and encode rich data types?

...

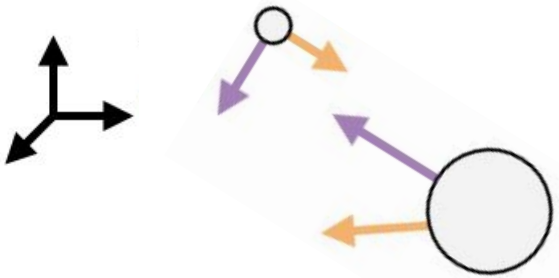


Same motif, different orientation.

Two point **masses** with **velocity** and **acceleration.**



Same system, with rotated coordinates.



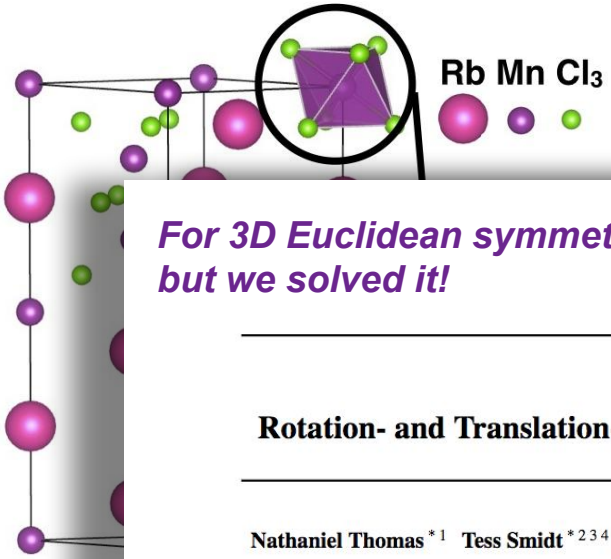
Geometric tensors transform predictably under rotation.

Deep learning shows promise for learning abstractions from data…
but it comes with significant challenges.

How do we represent atomic structures to neural networks?

Can we make neural networks that can understand symmetry and encode rich data types?

...



Two point **masses** with **velocity** and **acceleration.**

For 3D Euclidean symmetry this was an open question, but we solved it!

Tensor Field Networks:
Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds

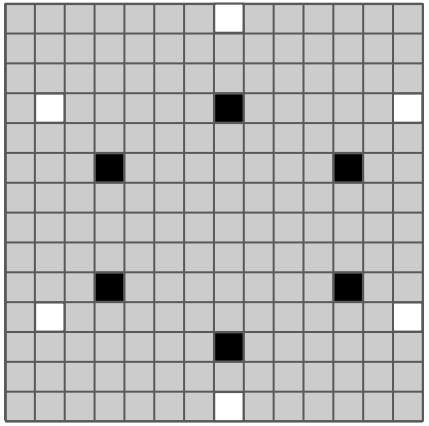
Nathaniel Thomas [*1] Tess Smidt [*2 3 4] Steven Kearnes [4] Lusann Yang [4] Li Li [4] Kai Kohlhoff [4] Patrick Riley [4]

Abstract

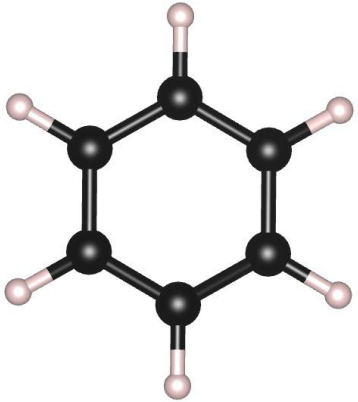
We introduce tensor field networks, which are locally equivariant to 3D rotations and translations (and invariant to permutations of points) at every layer. 3D rotation equivariance removes the need for data augmentation to identify features in arbitrary orientations. Our network uses filters built from spherical harmonics; due to the mathematical consequences of this filter choice, each

significantly more important in 3D than 2D. Without equivariant filters like those in our design, achieving an angular resolution of δ would require a factor of $\mathcal{O}(\delta^{-1})$ more filters in 2D but $\mathcal{O}(\delta^{-3})$ more filters in 3D.[1] Second, a 3D rotation- and translation-equivariant network can identify local features in different orientations and locations with the same filters, which is helpful for interpretability. Finally, the network naturally encodes geometric tensors (such as scalars, vectors, and higher-rank geometric objects), mathe-

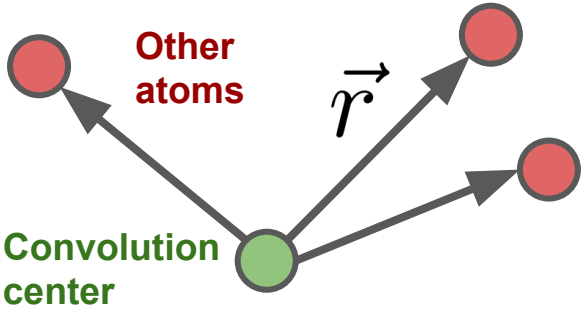
We use points. Images of atomic systems are sparse and imprecise.



vs.

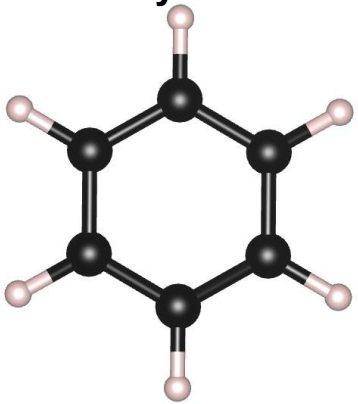


We use continuous convolutions with atoms as convolution centers.

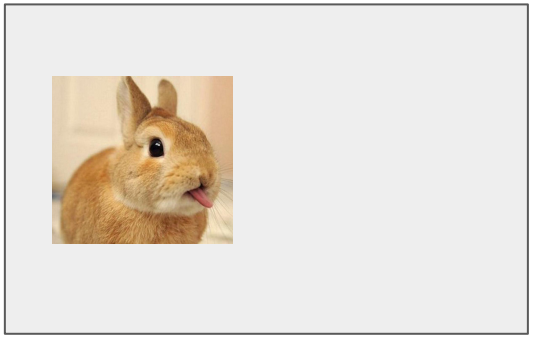


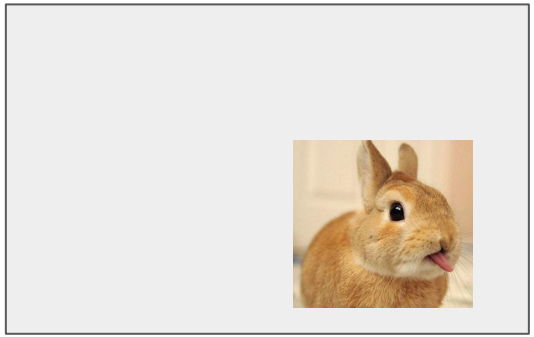
K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, Adv. in Neural Information Processing Systems 30 (2017). (arXiv: 1706.08566)

We encode the symmetries of 3D Euclidean space (3D translation- and 3D rotation-equivariance).

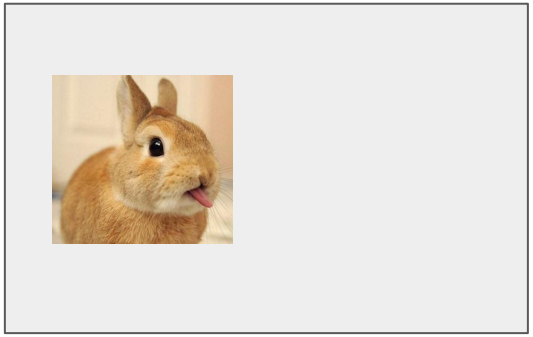


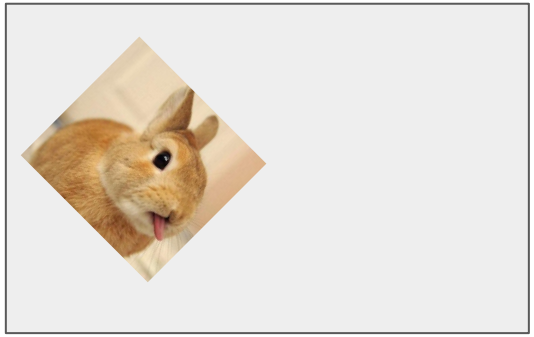
$$g \in SE(3)$$



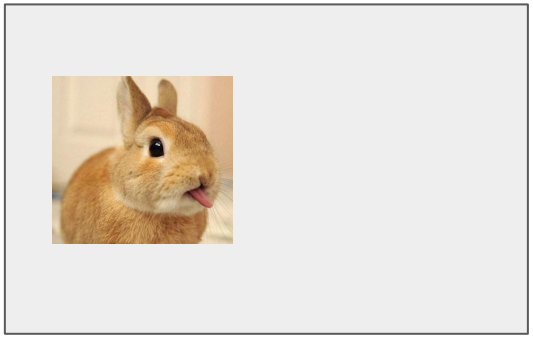


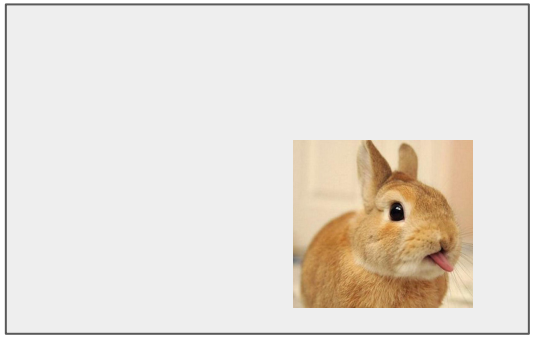
Translation equivariance





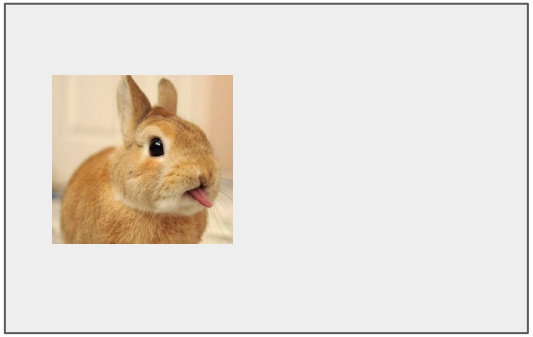
Rotation equivariance

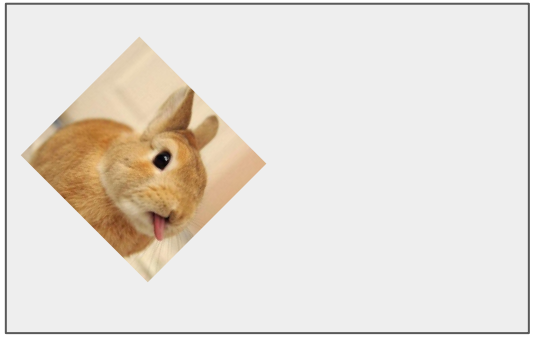




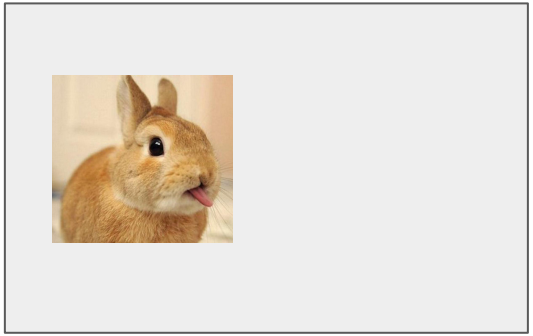
Translation equivariance

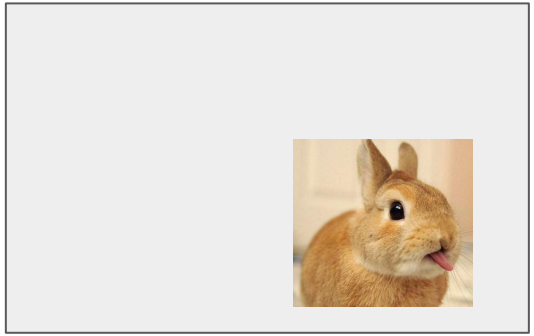
Convolutional neural network ✓





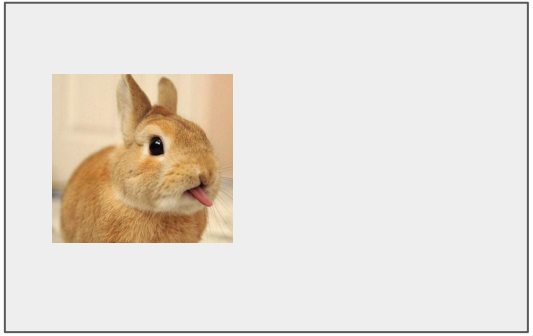
Rotation equivariance?

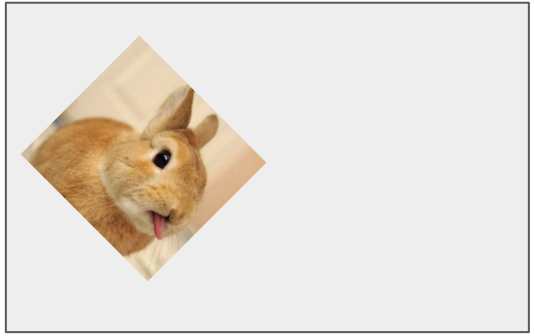




Translation equivariance

Convolutional neural network ✓





Rotation equivariance

~~Data augmentation~~

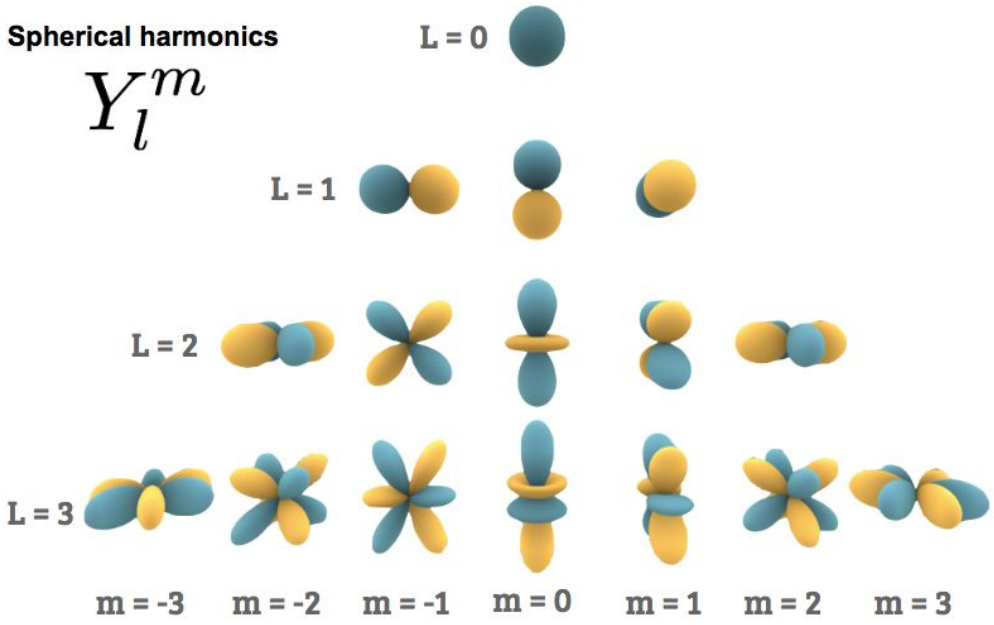
~~Radial functions~~

Want a network that both preserves geometry and exploits symmetry.

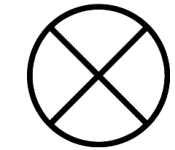
Tensor Field Networks naturally handle 3D geometry and features of physical systems.

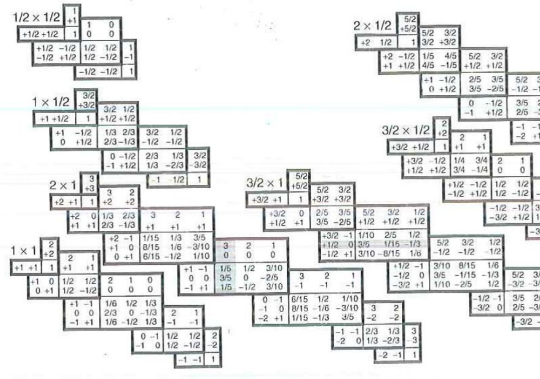
Convolutional filters based on spherical harmonics and learned radial functions.

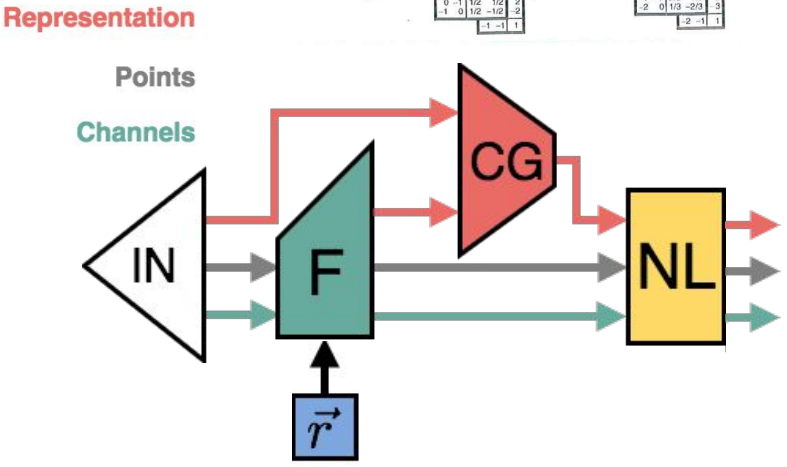
$$W(\vec{r}) = R(r)Y_l^m(\hat{r})$$



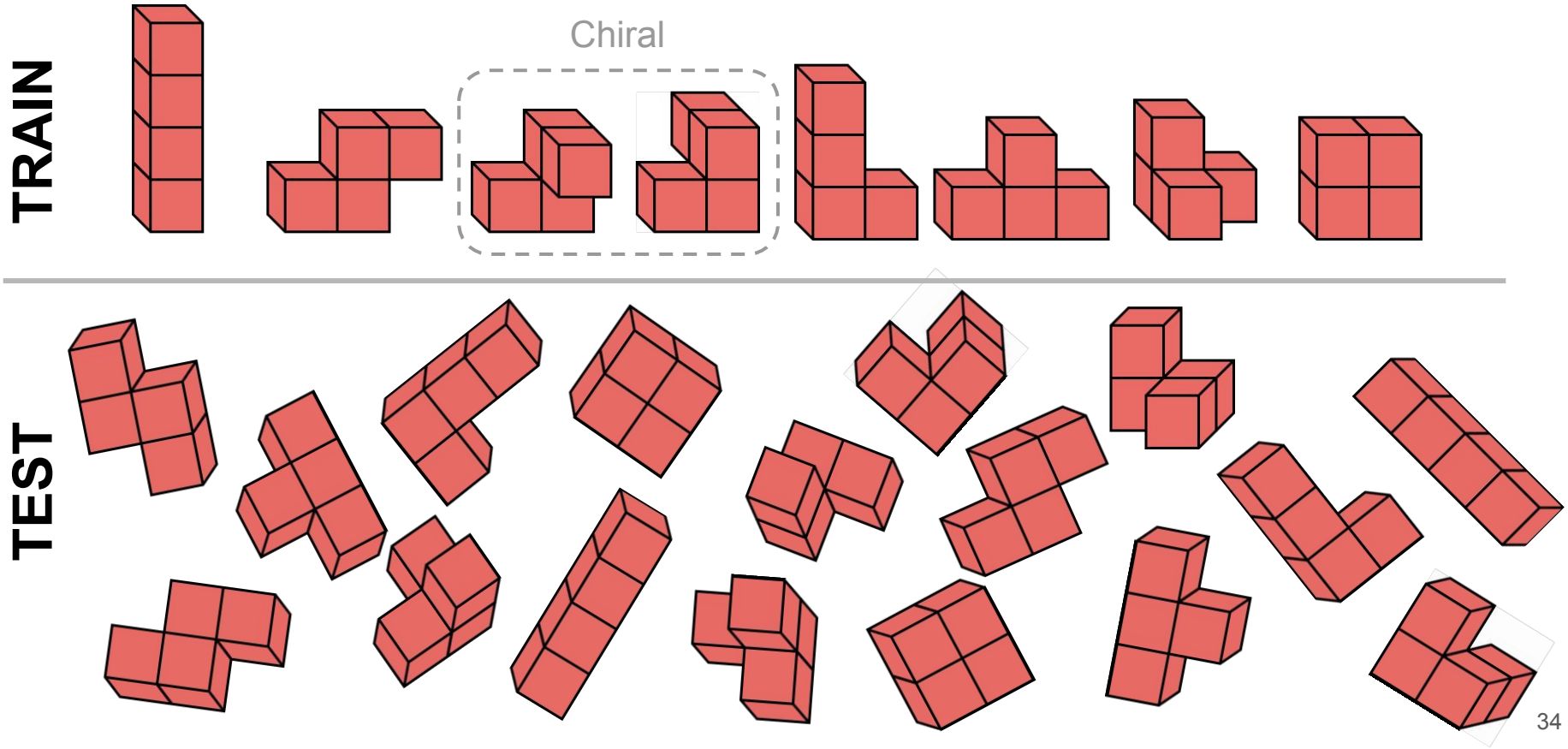
Everything in our network is a **geometric tensor**, so our network connectivity has to obey **tensor algebra**.







Test of 3D rotation equivariance: Trained on 3D Tetris shapes in one orientation, our network can perfectly identify these shapes in any orientation.

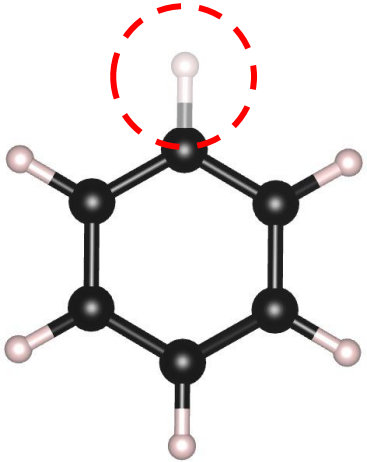


Given a small organic molecule with an atom removed, replace the correct element at the correct location in space.

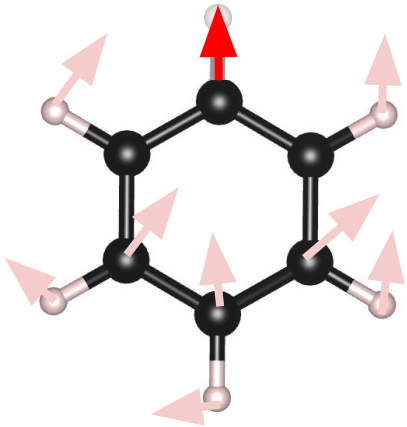
DATASET
QM9: *http://www.quantum-machine.org/datasets/*
134k molecules with 9 or less heavy atoms (non-hydrogen) and elements H, C, N, O, F.

TRAIN
1,000 molecules with 5-18 atoms

TEST
1,000 molecules with 19 atoms
1,000 molecules with 23 atoms
1,000 molecules with 25-29 atoms



Input coordinates with missing atom.

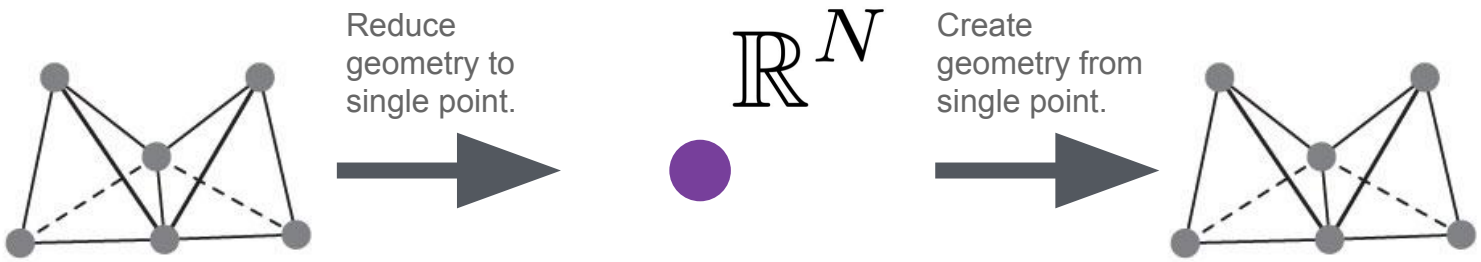


Network outputs
(N-1) atom type features (scalars),
(N-1) displacement vectors, and
(N-1) scalars indicating confidence probability used for "voting".

Atoms	Number of predictions	Accuracy (%) ($\leq$ 0.5 Å and atom type)	Distance MAE in Å
5-18 (train)	15 947	92.6	0.16
19	19 000	94.7	0.15
23	23 000	96.9	0.14
25-29	25 404	97.8	0.17

Learns to replace atoms with over 90% accuracy across train and test by seeing the same 1,000 molecules 200 times.

Creating an autoencoder for discrete geometry

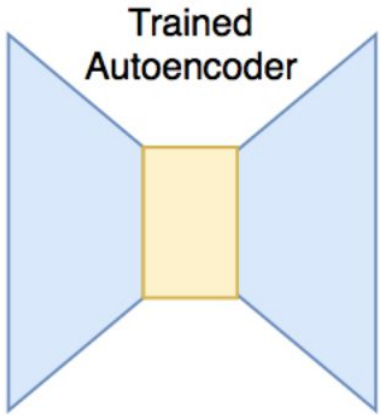


Discrete geometry

Continuous
Latent Representation
(N dimensional vector)

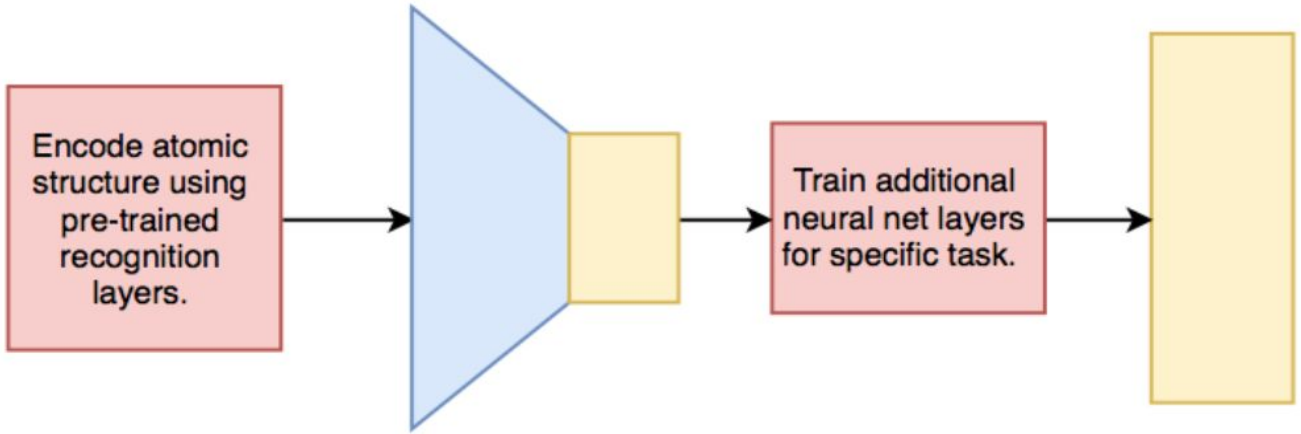
Discrete geometry

An autoencoder trained on atomic systems would solve multiple problems at once.

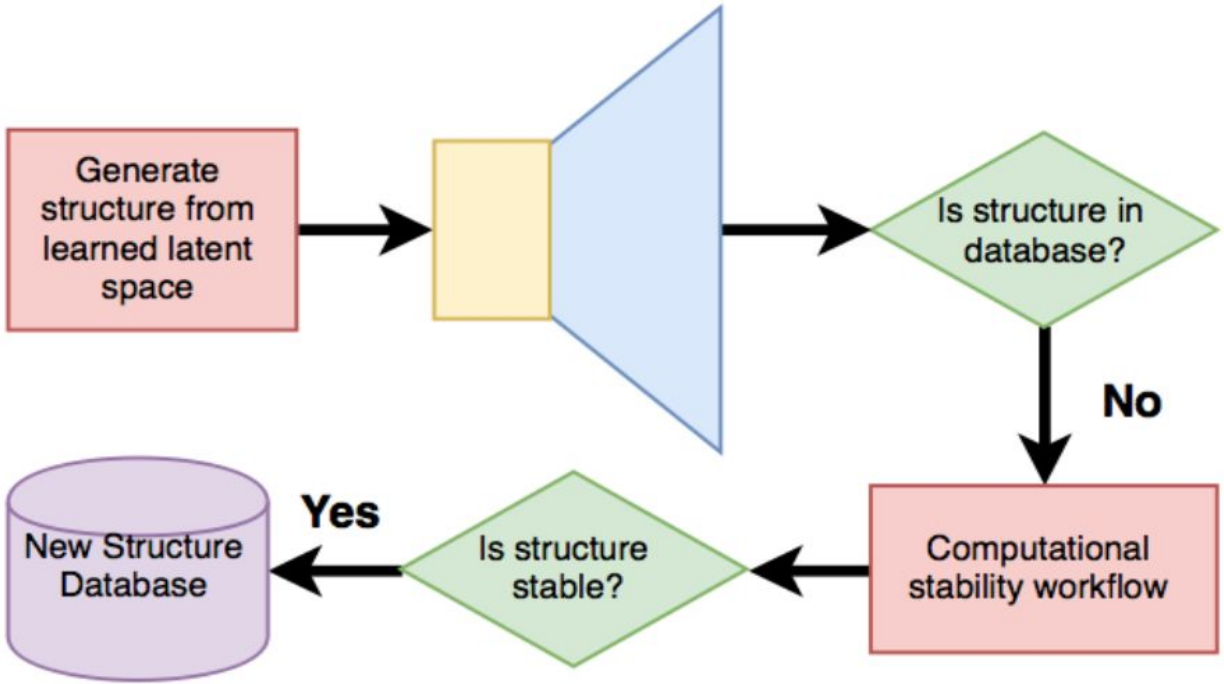


You get two models for the effort of one!

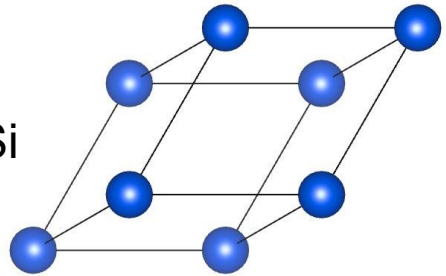
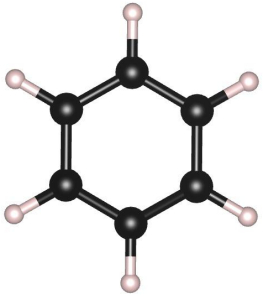
Encoding layers can be used in combination with new layers for specific tasks (predict energy, forces, etc for input structure).

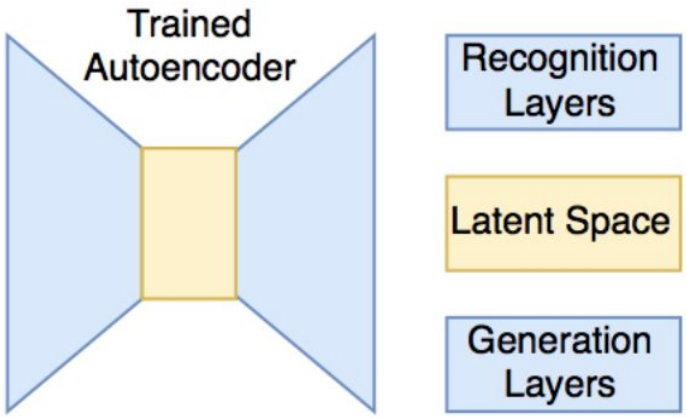


Decoding layers can be used to generate hypothetical atomic structures.

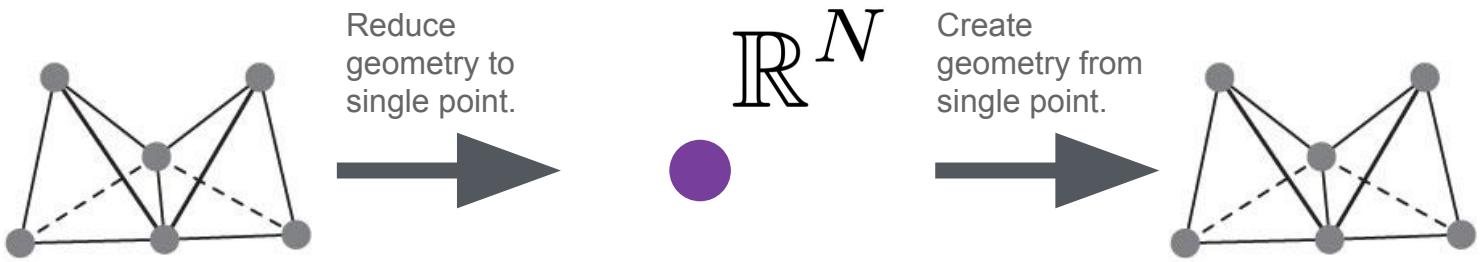


The latent space would provide a “map” for atomic systems.

Be able to relate... Si  **...to...**  benzene



Creating an autoencoder for discrete geometry

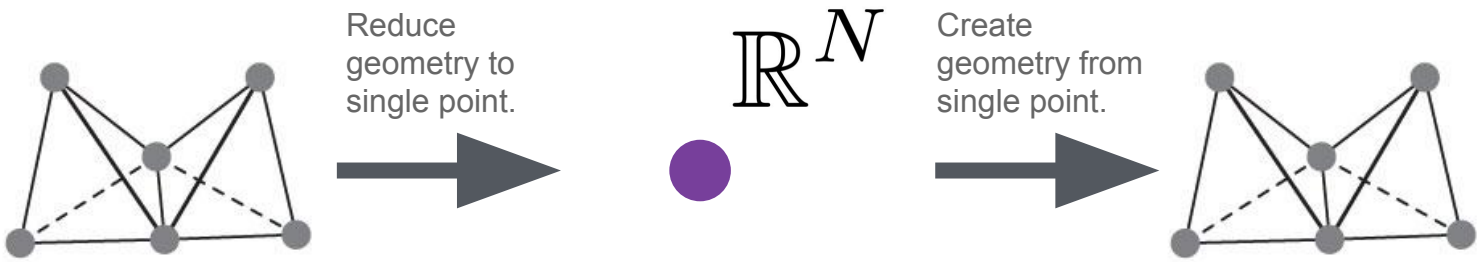


Discrete geometry

Continuous
Latent Representation
(N dimensional vector)

Discrete geometry

Creating an autoencoder for discrete geometry



Atomic structures are hierarchical and can be constructed from geometric motifs.

+ **Encode geometry** ✓
+ Encode hierarchy ?
+ Decode geometry ?
+ Decode hierarchy ?

(Need to do this in a recursive manner)

Okay, so how did I get here?

My Thesis: Toward the systematic design of complex materials from structural motifs (The TLDR; version)

Ch 1: Introduction

Ch 2: Methods (DFT)

Ch 3: Realization of a three-dimensional spin-anisotropic harmonic honeycomb iridate

- **K. Modic, T. Smidt et al, Nature Communications 5 (2014).**

Ch 4: Ab initio Studies of Structural and Energetic Trends in the Harmonic Honeycomb Iridates

- **T. Smidt et al, To be submitted to Physical Review B (2019)**

Ch 5: Silver Benzeneselenolate is a Self-Assembling Direct-Gap Metal-Organic Chalcogenide Assembly

- **M. Collins, T. Smidt, J. N. Hohman**
- **2 x publications in prep**

Ch 6: An Automatically Curated First-Principles Database of Ferroelectrics

- **T. Smidt et al, Under revision for Nature Scientific Data (2019)**

Ch 7: Tensor field networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds

- **T. Smidt*, N. Thomas* et al, arXiv:1802.08219**

Ch 8: Outlook

My Thesis: Toward the systematic design of complex materials from structural motifs (The TLDR; version)

Ch 1: Introduction

Ch 2: Methods (DFT)

Ch 3: Realization of a three-dimensional spin-anisotropic harmonic honeycomb iridate

- **K. Modic, T. Smidt et al, Nature Communications 5 (2014).**

Ch 4: Ab initio Studies of Structural and Energetic Trends in the Harmonic Honeycomb Iridates

- **T. Smidt et al, To be submitted to Physical Review B (2019)**

<u>INSERT EXISTENTIAL RESEARCH CRISIS</u>

3rd year

Ch 5: Silver Benzeneselenolate is a Self-Assembling Direct-Gap Metal-Organic Chalcogenide Assembly

- **M. Collins, T. Smidt, J. N. Hohman**
- **2 x publications in prep**

Ch 6: An Automatically Curated First-Principles Database of Ferroelectrics

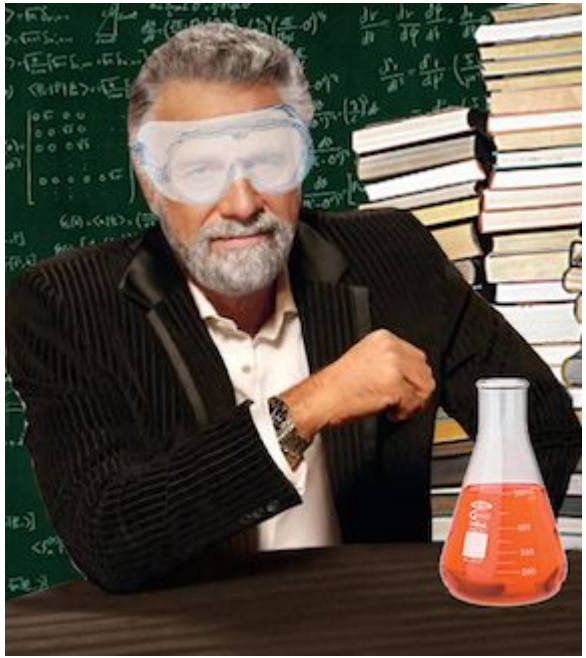
- **T. Smidt et al, Under revision for Nature Scientific Data (2019)**

Ch 7: Tensor field networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds

- **T. Smidt*, N. Thomas* et al, arXiv:1802.08219**

Ch 8: Outlook

The Lunch Experiment:
Randomized Controlled Lunches for Grad Students



I don't always eat lunch, but when I do,
I prefer The Lunch Experiment.

400+ participants
100+ lunches

Automated scheduling and invitation system maximizing for diversity of majors.

My Thesis: Toward the systematic design of complex materials from structural motifs (The TLDR; version)

Ch 1: Introduction

Ch 2: Methods (DFT)

Ch 3: Realization of a three-dimensional spin-anisotropic harmonic honeycomb iridate

- **K. Modic, T. Smidt et al, Nature Communications 5 (2014).**

Ch 4: Ab initio Studies of Structural and Energetic Trends in the Harmonic Honeycomb Iridates

- **T. Smidt et al, To be submitted to Physical Review B (2019)**

<u>INSERT EXISTENTIAL RESEARCH CRISIS</u>

3rd year

Ch 5: Silver Benzeneselenolate is a Self-Assembling Direct-Gap Metal-Organic Chalcogenide Assembly

- **M. Collins, T. Smidt, J. N. Hohman**
- **2 x publications in prep**

Ch 6: An Automatically Curated First-Principles Database of Ferroelectrics

- **T. Smidt et al, Under revision for Nature Scientific Data (2019)**

Ch 7: Tensor field networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds

- **T. Smidt*, N. Thomas* et al, arXiv:1802.08219**

Ch 8: Outlook

My Thesis: Toward the systematic design of complex materials from structural motifs (The TLDR; version)

Ch 1: Introduction

Ch 2: Methods (DFT)

Ch 3: Realization of a three-dimensional spin-anisotropic harmonic honeycomb iridate

- **K. Modic, T. Smidt et al, Nature Communications 5 (2014).**

Ch 4: Ab initio Studies of Structural and Energetic Trends in the Harmonic Honeycomb Iridates

- **T. Smidt et al, To be submitted to Physical Review B (2019)**

INSERT EXISTENTIAL RESEARCH CRISIS

3rd year

Ch 5: Silver Benzeneselenolate is a Self-Assembling Direct-Gap Metal-Organic Chalcogenide Assembly

- **M. Collins, T. Smidt, J. N. Hohman**
- **2 x publications in prep**

Ch 6: An Automatically Curated First-Principles Database of Ferroelectrics

- **T. Smidt et al, Under revision for Nature Scientific Data (2019)**

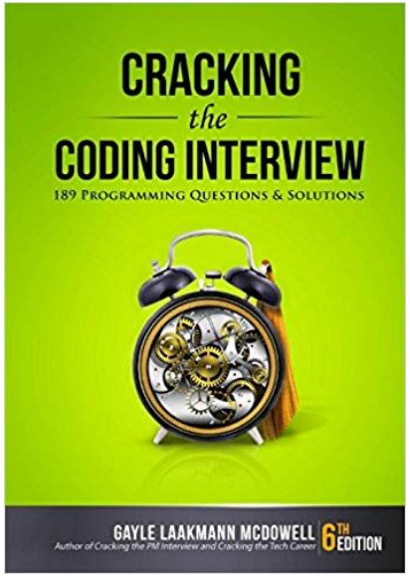
DEEP LEARNING AND GOOGLE

5-6th years

Ch 7: Tensor field networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds

- **T. Smidt*, N. Thomas* et al, arXiv:1802.08219**

Ch 8: Outlook



+

CS 182/282A

=

Spring 2019 **COMPSCI 282A 001 - LEC 001** offered through **Electrical Engineering and Computer Sciences**

Designing, Visualizing and Understanding Deep Neural Networks

John F Canny
M, W
8:00 am - 9:29 am
Dwinelle 145
Class #: 31116
Units: 4

Open Seats
30 Unreserved Seats

Deep Networks have revolutionized computer vision, language technology, robotics and control. They have growing impact in many other areas of science and engineering. They do not however, follow a closed or compact set of theoretical principles. In Yann Lecun's words they require "an interplay between intuitive insights, theoretical modeling, practical implementations, empirical studies, and...

My Thesis: Toward the systematic design of complex materials from structural motifs
(The TLDR; version)

Ch 1: Introduction

Ch 2: Methods (DFT)

Ch 3: Realization of a three-dimensional spin-anisotropic harmonic honeycomb iridate

- **K. Modic, T. Smidt et al, Nature Communications 5 (2014).**

Ch 4: Ab initio Studies of Structural and Energetic Trends in the Harmonic Honeycomb Iridates

- **T. Smidt et al, To be submitted to Physical Review B (2019)**

INSERT EXISTENTIAL RESEARCH CRISIS

3rd year

Ch 5: Silver Benzeneselenolate is a Self-Assembling Direct-Gap Metal-Organic Chalcogenide Assembly

- **M. Collins, T. Smidt, J. N. Hohman**
- **2 x publications in prep**

Ch 6: An Automatically Curated First-Principles Database of Ferroelectrics

- **T. Smidt et al, Under revision for Nature Scientific Data (2019)**

DEEP LEARNING AND GOOGLE

5-6th years

Ch 7: Tensor field networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds

- **T. Smidt*, N. Thomas* et al, arXiv:1802.08219**

Ch 8: Outlook

tensor field networks

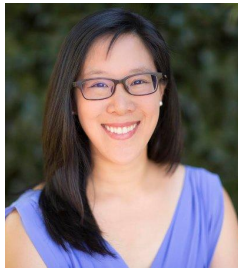
Stanford

Google Accelerated Science Team

Nate Thomas

Patrick Riley

Steve Kearnes



Lusann Yang

Li Li

Kai Kohlhoff

MOChAs

Mary Collins

Nate Hohman

Iridates

James Analytis

Sinead Griffin

Kim Modic, Itamar Kimchi, Nicholas P. Breznay, Alun Biffin, Radu Coldea, Ashvin Vishwanath, Arkady Shekhter, Ross D. McDonald...

My PhD Advisor

Jeff Neaton

Atomic Architects — Summer 2019

In summary...

There's a lot of work to do in applying deep learning methods for tasks in atomic systems. Methods may not work out of the box. For example, we made tensor field networks to naturally handle the geometry of atomic systems.

Google is an amazing place to work. I highly recommend interning during grad school if you can.

Berkeley Lab is in a great position to play a central role in how ML methods are adopted in the chemistry and materials communities.

Review on ML for molecules and materials:

Machine learning for molecular and materials science
Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev & Aron Walsh
Nature **559**, 547–555 (2018).
https://doi.org/10.1038/s41586-018-0337-2

Come visit and chat about DL for atomic systems! My office is 50F-1643.