



Machine Learning for Molecules

Ichigaku Takigawa

takigawa@icredd.hokudai.ac.jp

Hokkaido Univ ICReDD-Faculty of Medicine Joint Symposium

15 October 2021 @ Hokkaido University

A machine learning (ML) researcher working for



ML for Stem Cell Biology



RIKEN Center for AI Project
Medical-risk Avoidance based on iPS Cells Team
(A joint lab with Kyoto Univ CiRA)

ML for Chemistry



Inst. Chemical Reaction Design & Discovery
Hokkaido Univ

A machine learning (ML) researcher working for



ML for Stem Cell Biology



RIKEN Center for AI Project

Medical-risk Avoidance based on iPS Cells Team
(A joint lab with Kyoto Univ CiRA)

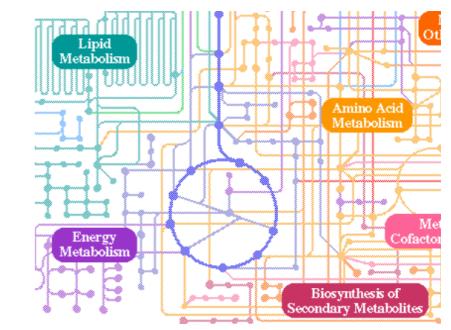
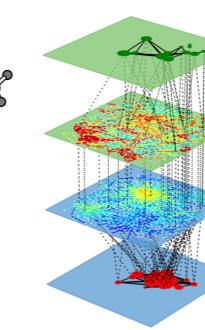
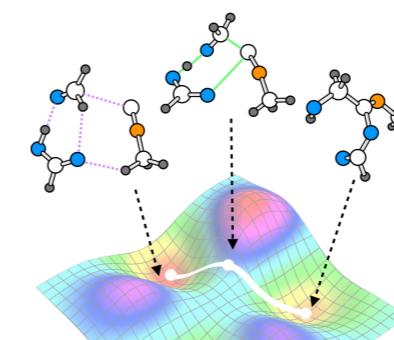
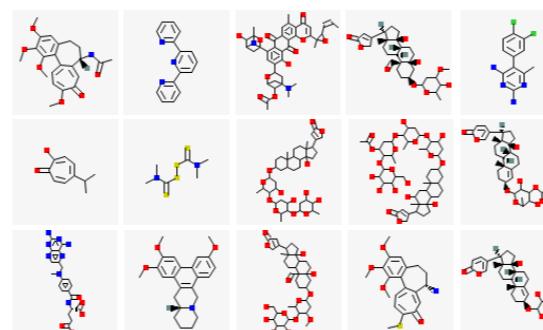
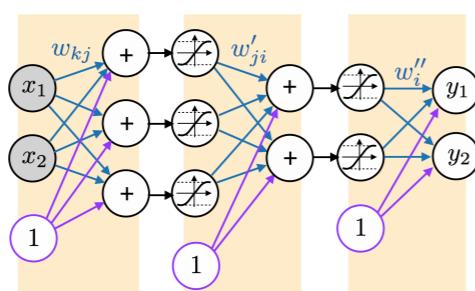
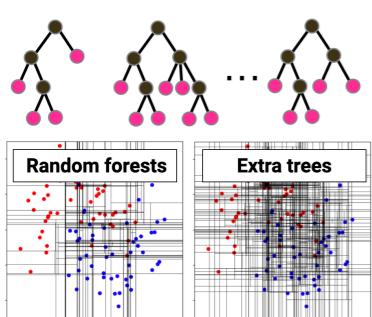
ML for Chemistry



*Inst. Chemical Reaction Design & Discovery
Hokkaido Univ*

Interests: Machine Learning and Machine Discovery

An intersection of ML with **combinatorial structures** + ML for **natural sciences**



A machine learning (ML) researcher working for



ML for Stem Cell Biology



RIKEN Center for AI Project

Medical-risk Avoidance based on iPS Cells Team
(A joint lab with Kyoto Univ CiRA)

ML for Chemistry

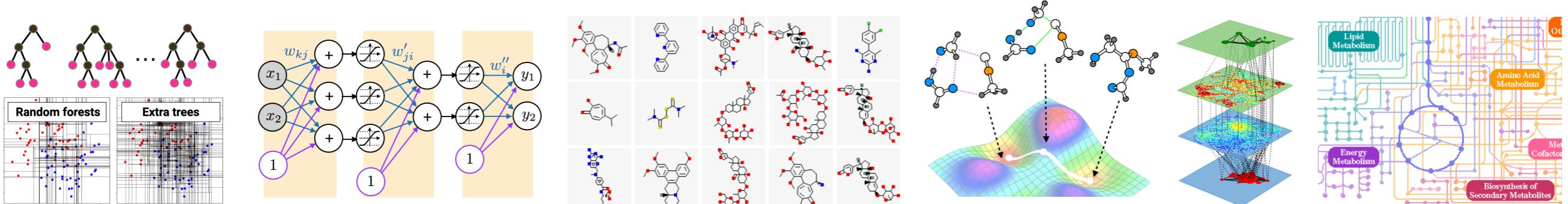


Inst. Chemical Reaction Design & Discovery

Hokkaido Univ

Interests: Machine Learning and Machine Discovery

An intersection of ML with **combinatorial structures** + ML for **natural sciences**



Joint project with HU Med Dept:

- Prof. Shinya Tanaka: Cancer diagnosis with fluorescent markers, Enzyme-catalyzed reaction design
- Prof. Shigetsugu Hatakeyama: Mediator complex of transcription regulations (*Nat Commun* 2020, 2015)
- Prof. Ichiro Yabe: Video-based predictions of motor symptom severity
- Prof. Yasuyuki Fujita: Cell competition (*Cell Reports* 2018; *Sci Rep* 2015)
- Prof. Hidenao Sasaki: Copy number variations for neurodegenerative diseases (*Mol Brain* 2017)

X-informatics: Bio- and Chemo-informatics

In addition to pure ML research, I've worked for **bio/chemo-informatics**

- **Biochemical reaction networks (Metabolic pathways)**

Bioinformatics 2007, 2008a, 2008b, 2009, 2010

Nucleic Acids Res 2011, PLoS One 2012, 2013, KDD'07

- **Drug-target interactions (Polypharmacology)**

PLoS One 2011, Drug Discov Today 2013, Brief Bioinform 2014

- **Modulatory proteolysis**

Mol Cell Proteom 2016, Genome Informatics 2009

(We also developed a database <http://calpain.org>)

- **Genomic repeats**

Discrete Appl Math 2013, 2016, AAAI 2020

- **Genetic variations in cancer cells**

Brief Bioinform 2014

- **Mediator complex and transcription regulation**

Nat Commun 2015, 2020 (w/ Prof. Hatakeyama)

- **Genomic copy number variations for neurodegenerative diseases**

Mol Brain 2017 (w/ Prof. Sasaki)

- **Cell competitions and cancer cells**

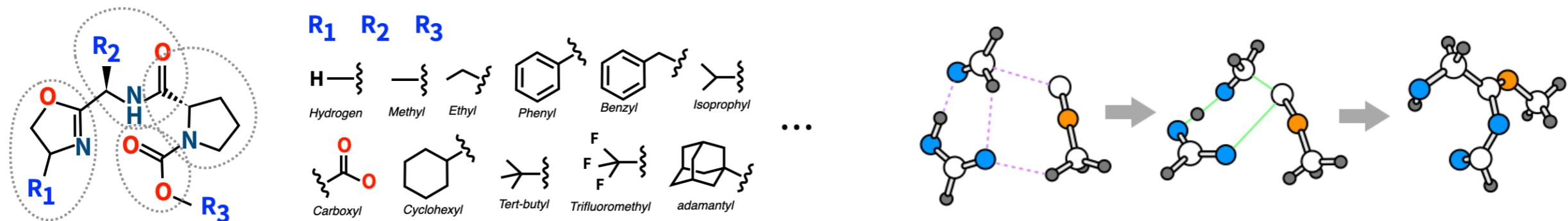
Cell Reports 2018, Sci Rep 2015 (w/ Prof. Fujita)



Pathway Commons

Access and discover data integrated from public pathway and interactions databases.

Molecules have a combinatorial aspect



c&en

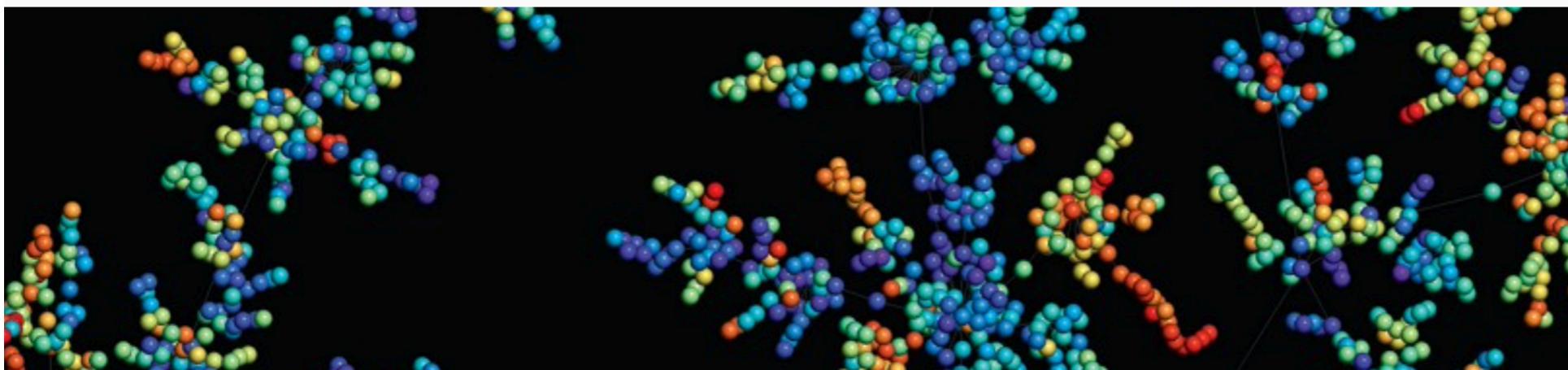
COMPUTATIONAL CHEMISTRY

Exploring chemical space: Can AI take us where no human has gone before?

Artificial intelligence is helping us find novel, useful molecules. For the field to really take off, though, these tools will need to be accessible to the wider chemistry community

by Sam Lemonick

April 6, 2020 | A version of this story appeared in **Volume 98, Issue 13**



BY THE NUMBERS

10¹⁸⁰

An upper estimate of the number of possible molecules

10⁸⁰

Estimated number of atoms in the universe

10⁶⁰

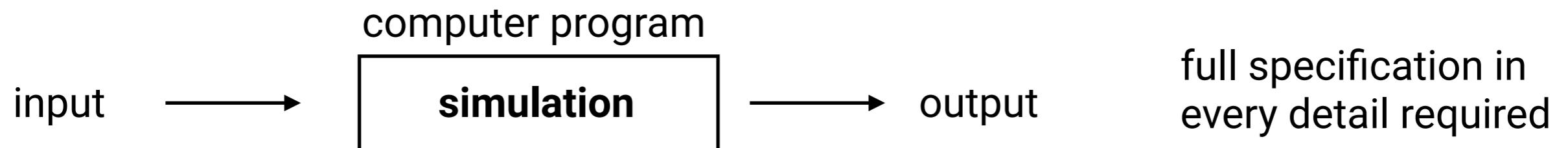
An estimate of the number of possible small organic molecules

10⁸

The number of organic and inorganic substances in the CAS database

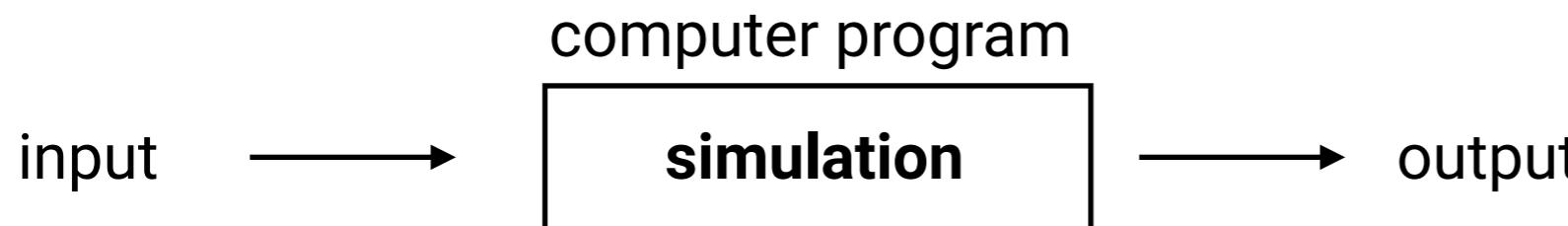
ML: A new way for (lazy) programming

deductive (rationalism)



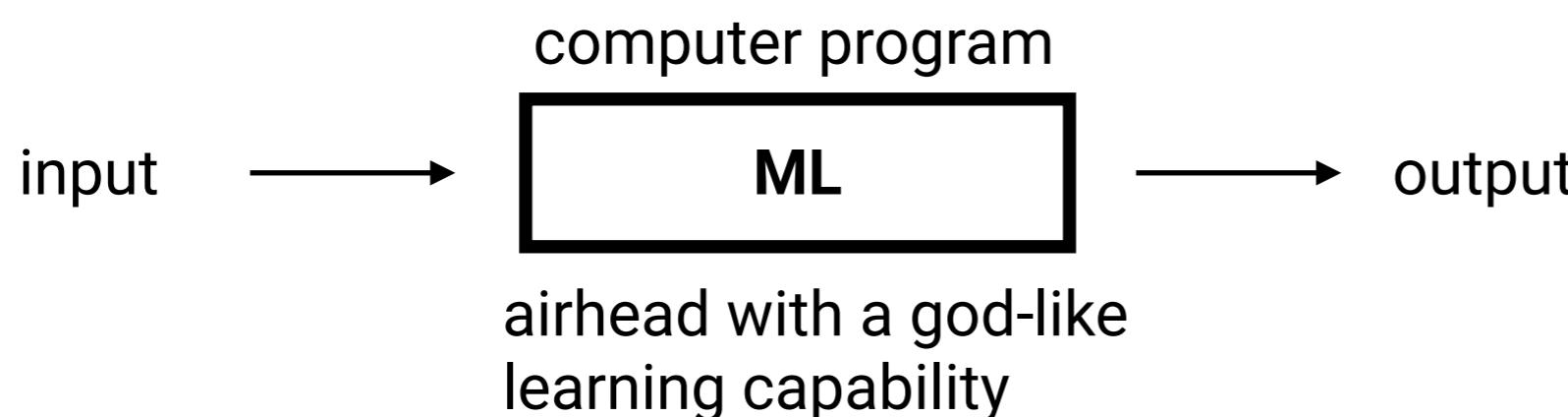
ML: A new way for (lazy) programming

deductive (rationalism)



full specification in
every detail required

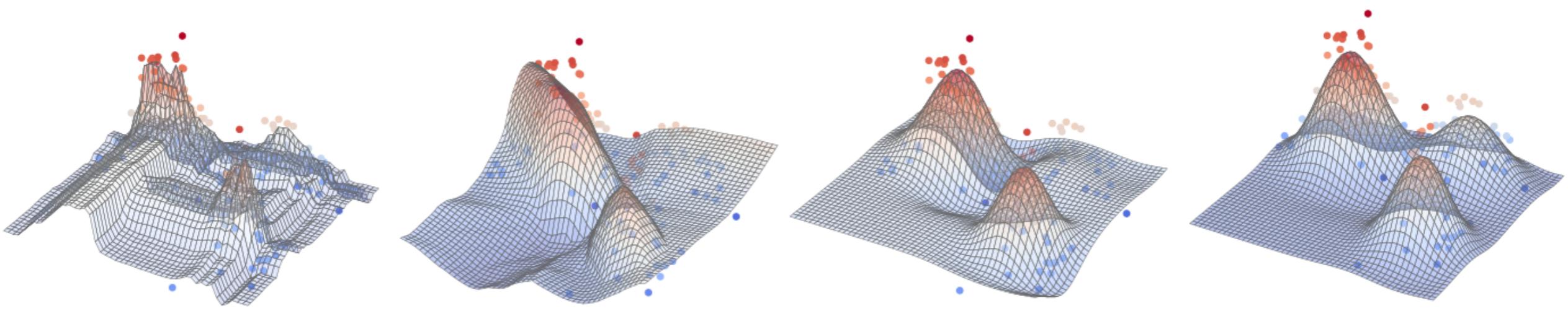
inductive (empiricism)



give up explicit model

instead, grab a tunable
model, and **show it many
input-output instances**

All about fitting a **very-flexible** function to **finite** points in **high-dimensional** space.



Random Forest

Neural Networks

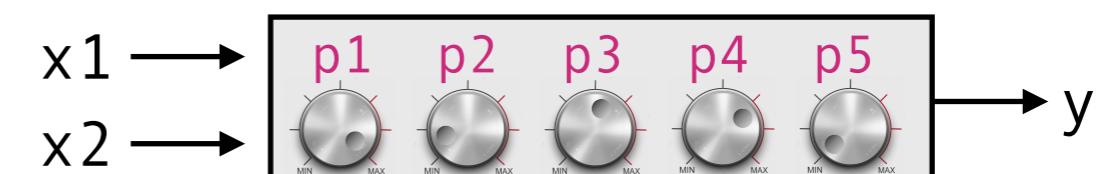
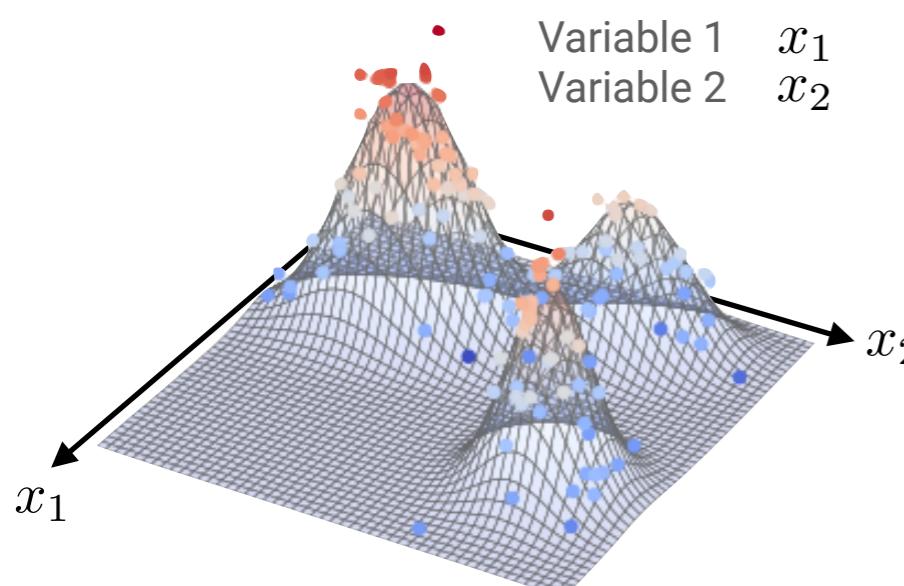
SVR

Kernel Ridge

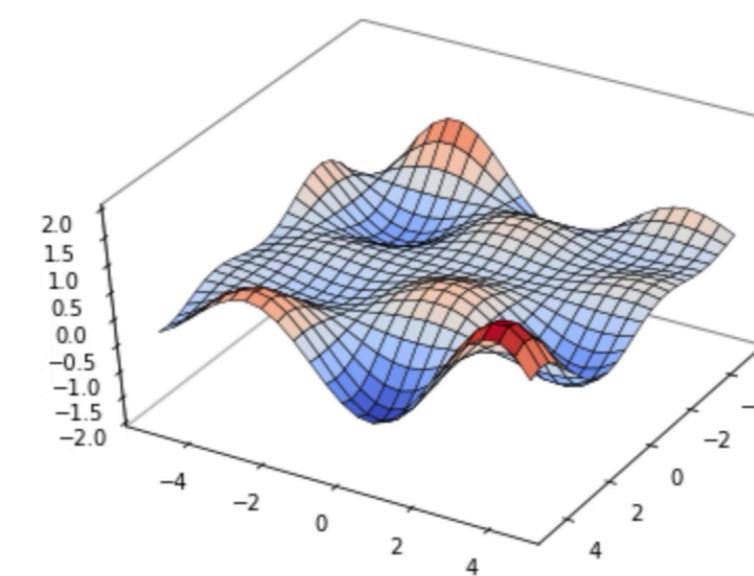
ML: A new way for (lazy) programming

All about statistical and algorithmic techniques for surface-model fitting to data points by adjusting **model parameters**.

ML = Tweak **these parameter values** to best fit a surface model to the given points.



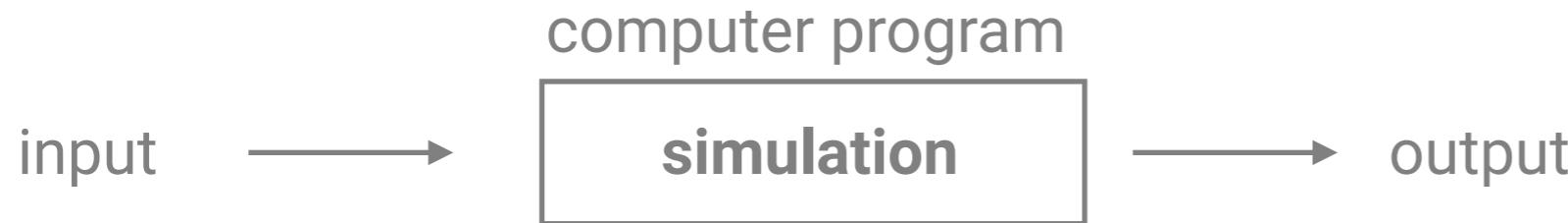
5 params



p_1	3.90
p_2	2.00
p_3	0.20
p_4	0.10
p_5	0.20

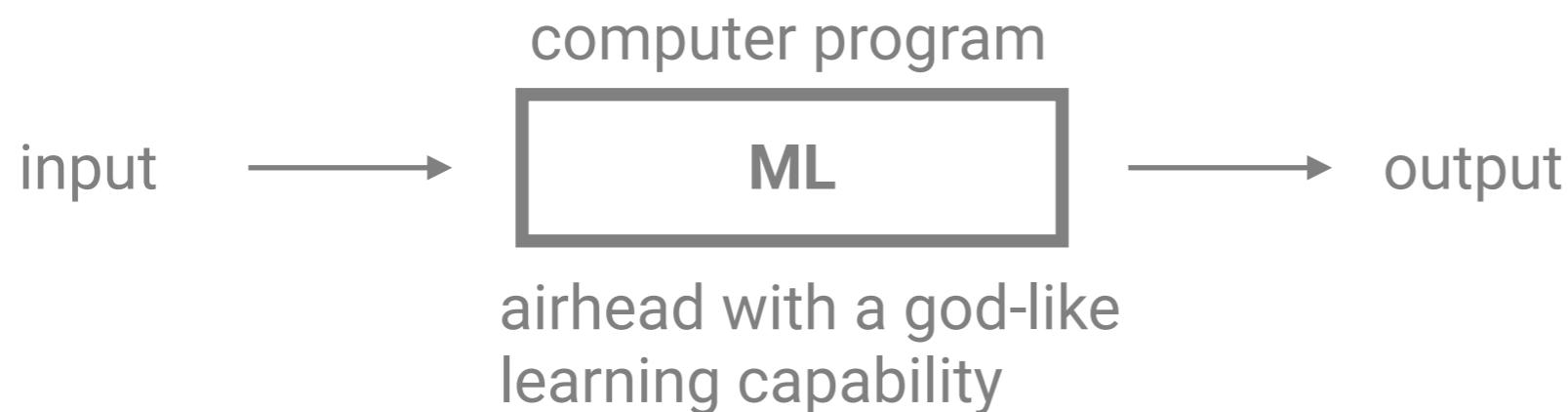
A modern aspect of ML

deductive (rationalism)



full specification in every detail required

inductive (empiricism)



give up explicit model

instead, grab a tunable model, and show it many input-output instances

All about fitting a **very-flexible** function to **finite** points in **high-dimensional** space.

ResNet50: **26 million** params

ResNet101: **45 million** params

EfficientNet-B7: **66 million** params

VGG19: **144 million** params

12-layer, 12-heads BERT: **110 million** params

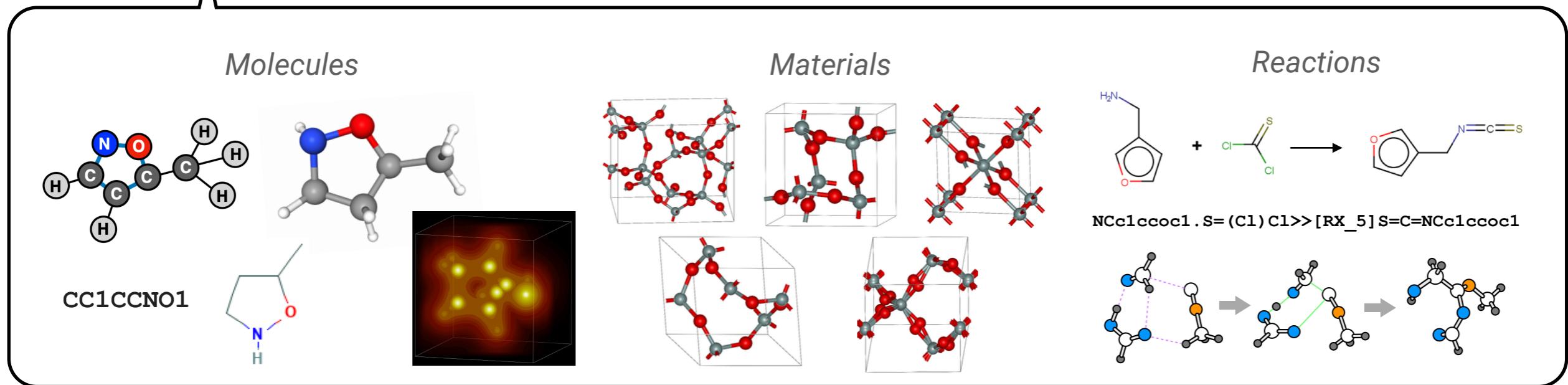
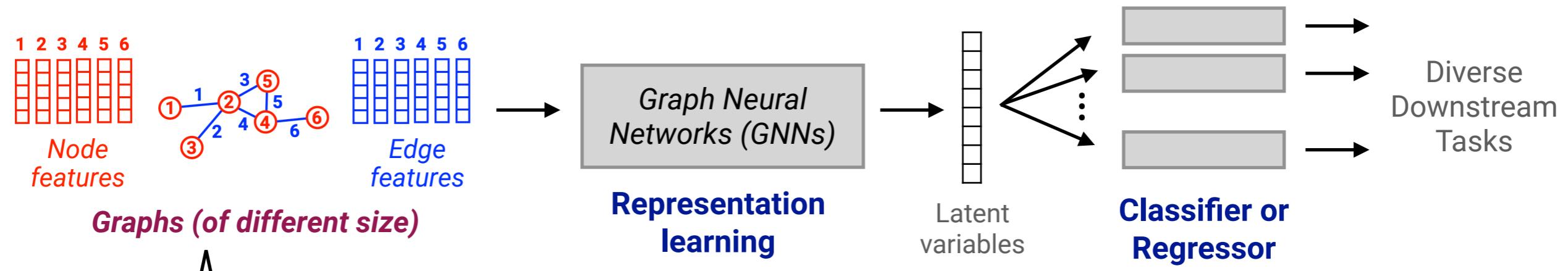
24-layer, 16-heads BERT: **336 million** params

GPT-2 XL: **1558 million** params

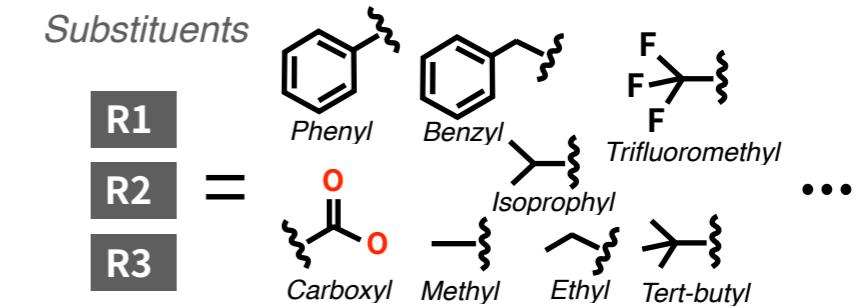
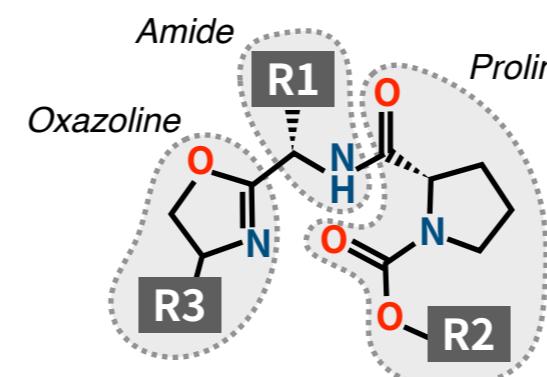
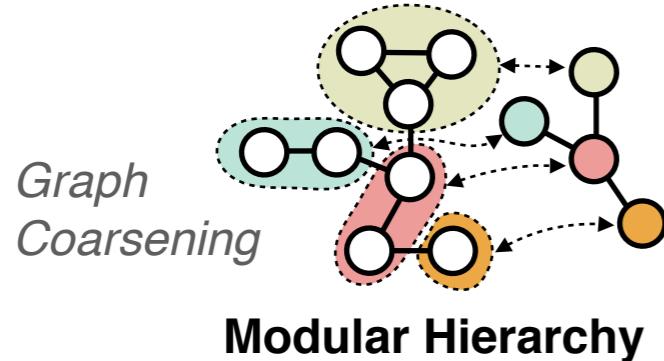
GPT-3: **175 billion** params

Modern ML: Can we imagine what would happen if we try to fit a function having **175 billion** parameters to **100 million** data points in **10 thousand** dimension??

This Talk: ML for Molecular Graphs



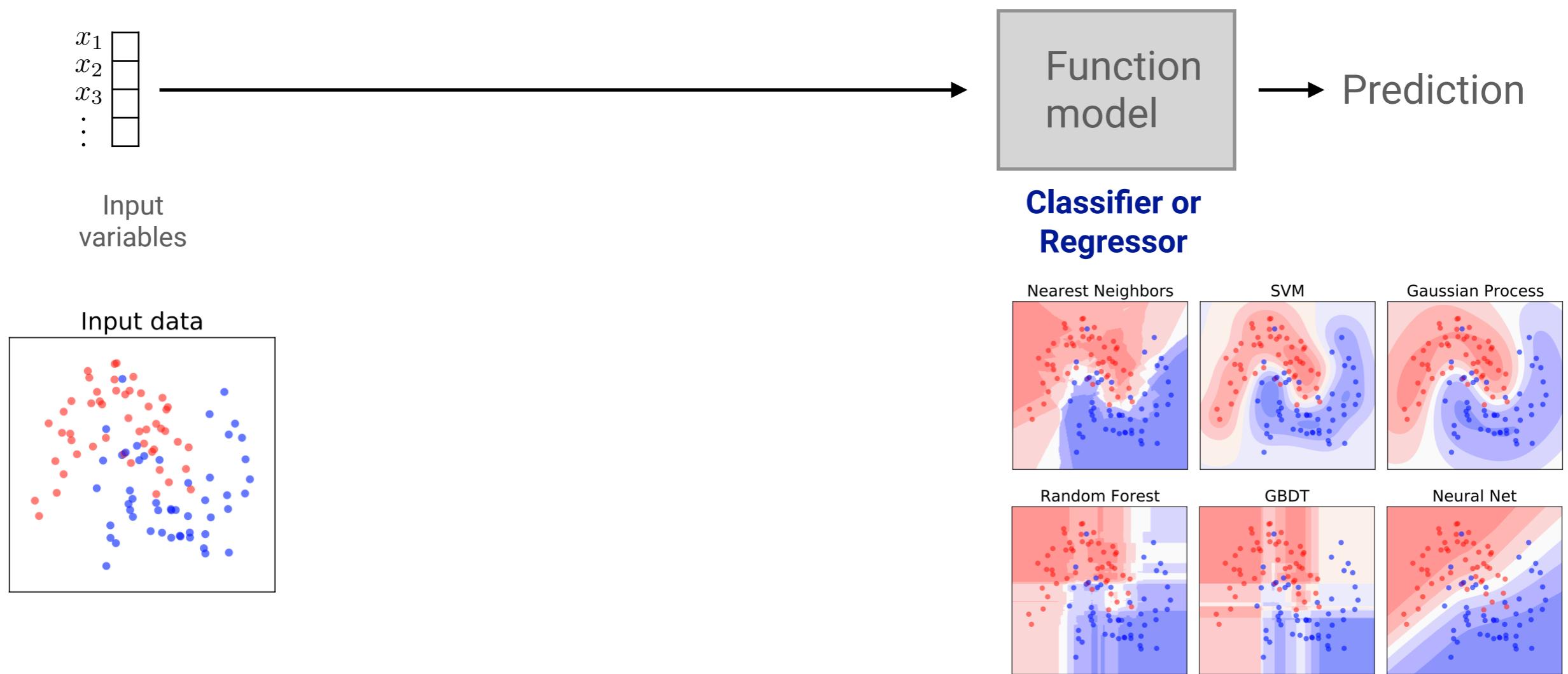
Combinatorial aspects



Compositionality

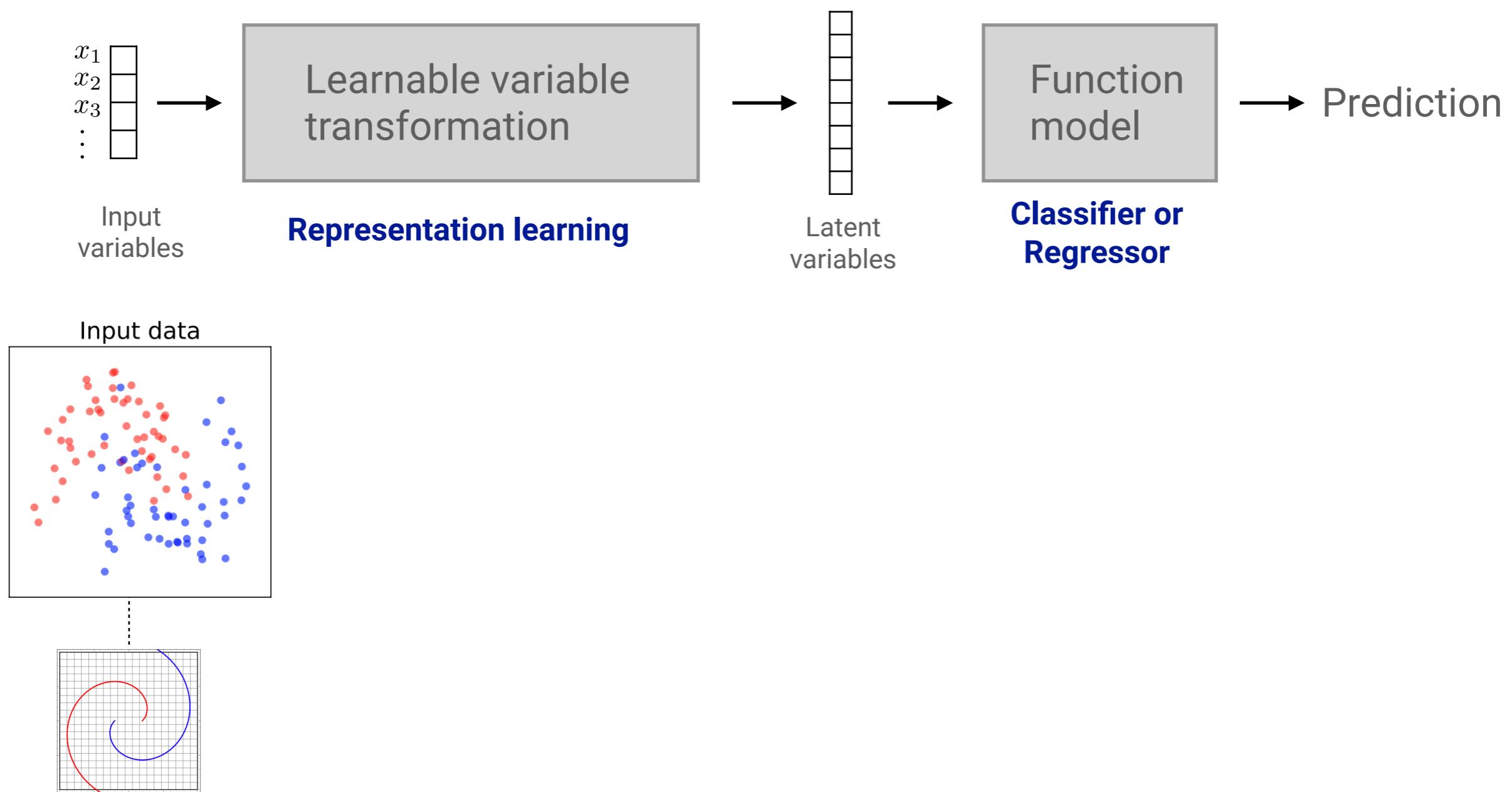
Representation Learning

Use **some inductive biases** to constrain/regularize the model space.



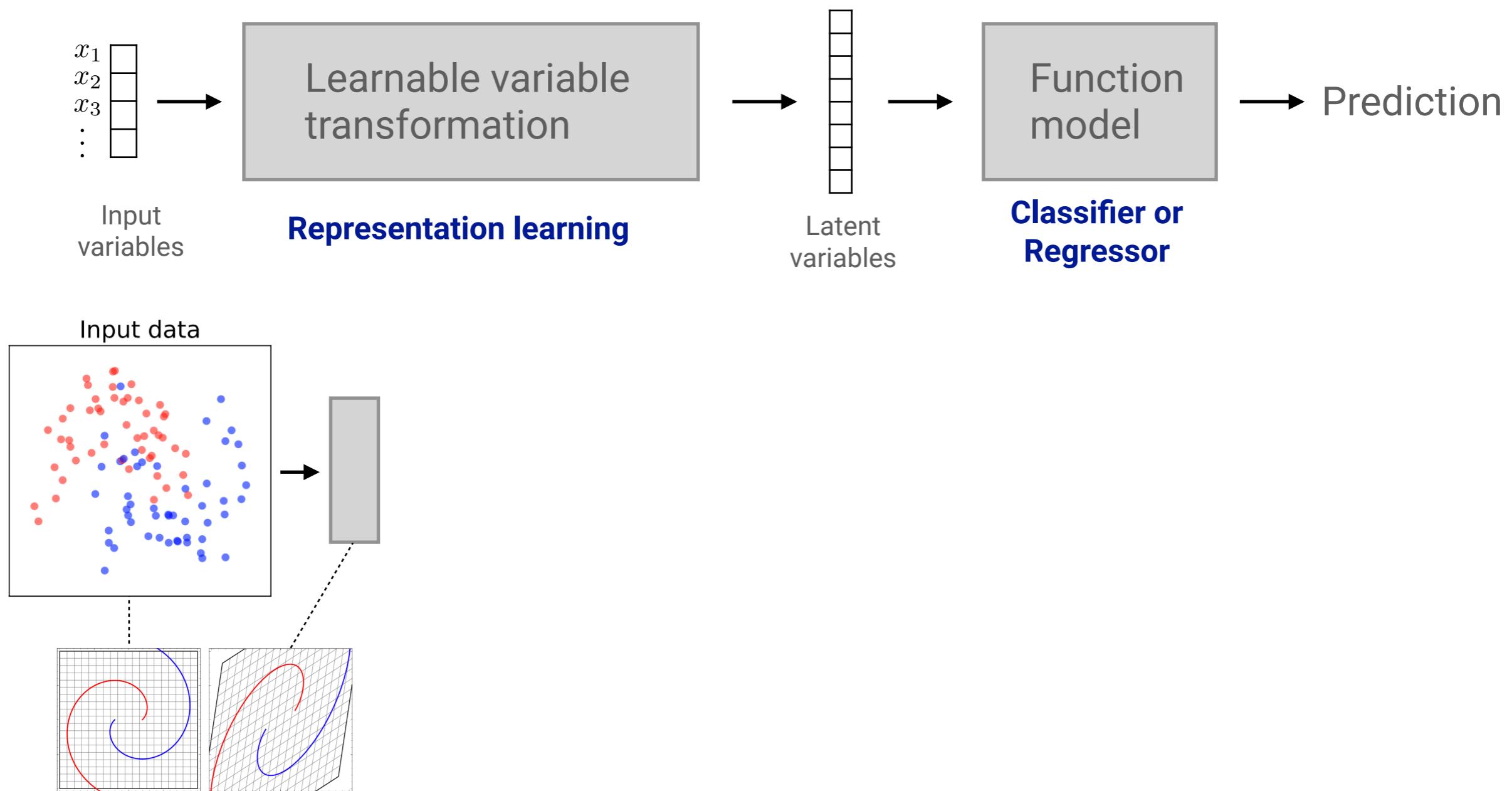
Representation Learning

Use **some inductive biases** to constrain/regularize the model space.



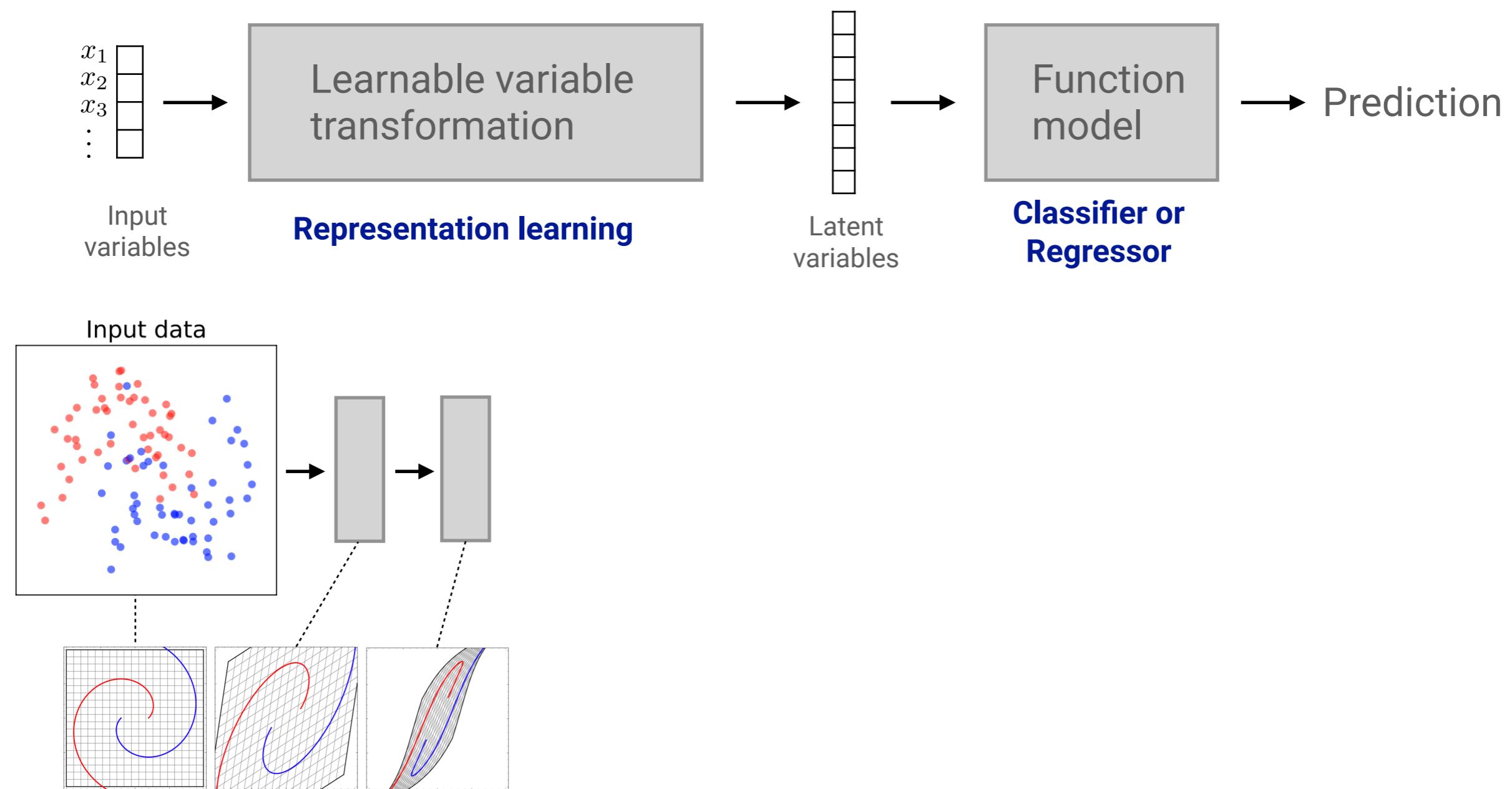
Representation Learning

Use **some inductive biases** to constrain/regularize the model space.



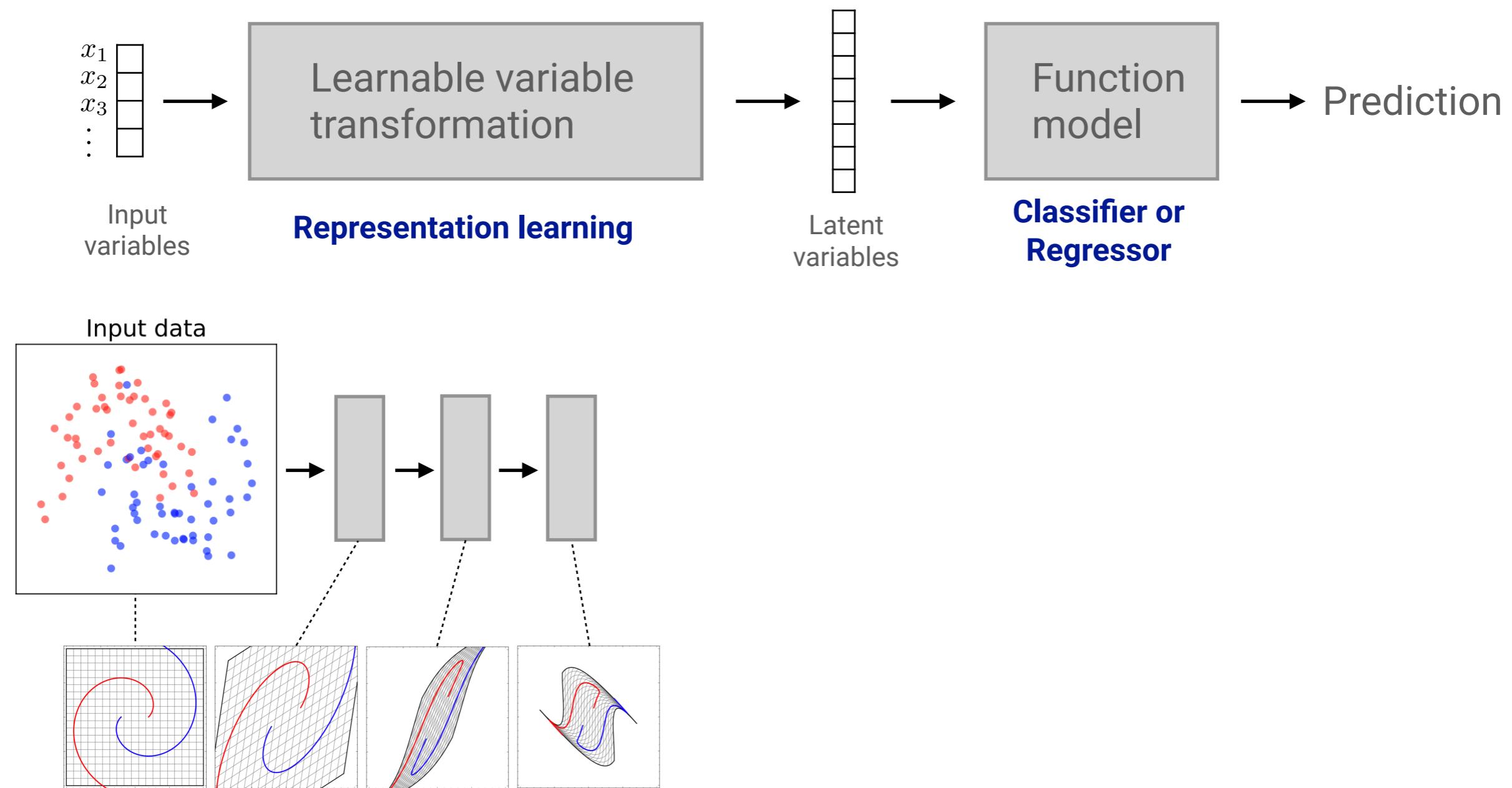
Representation Learning

Use **some inductive biases** to constrain/regularize the model space.



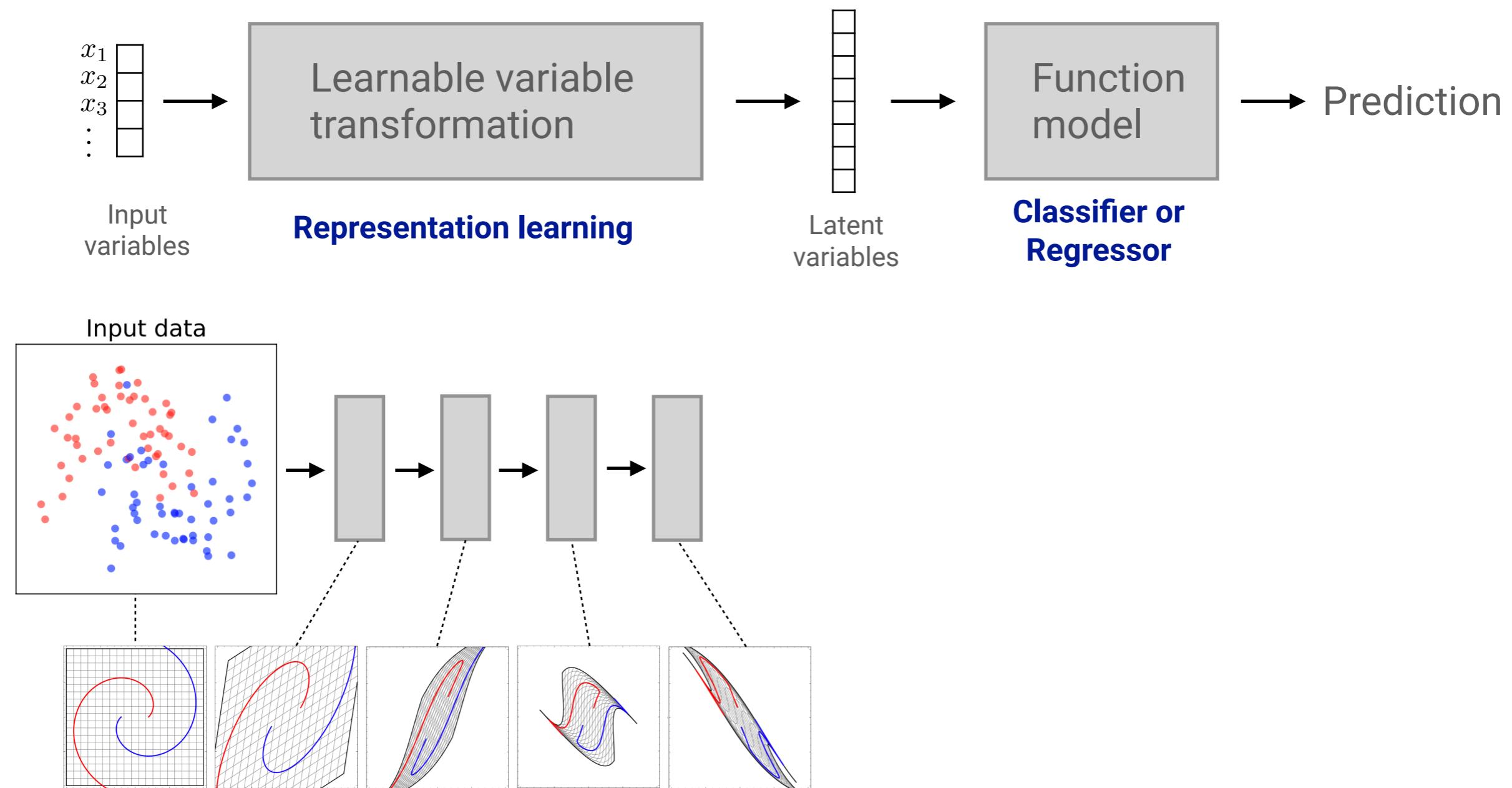
Representation Learning

Use **some inductive biases** to constrain/regularize the model space.



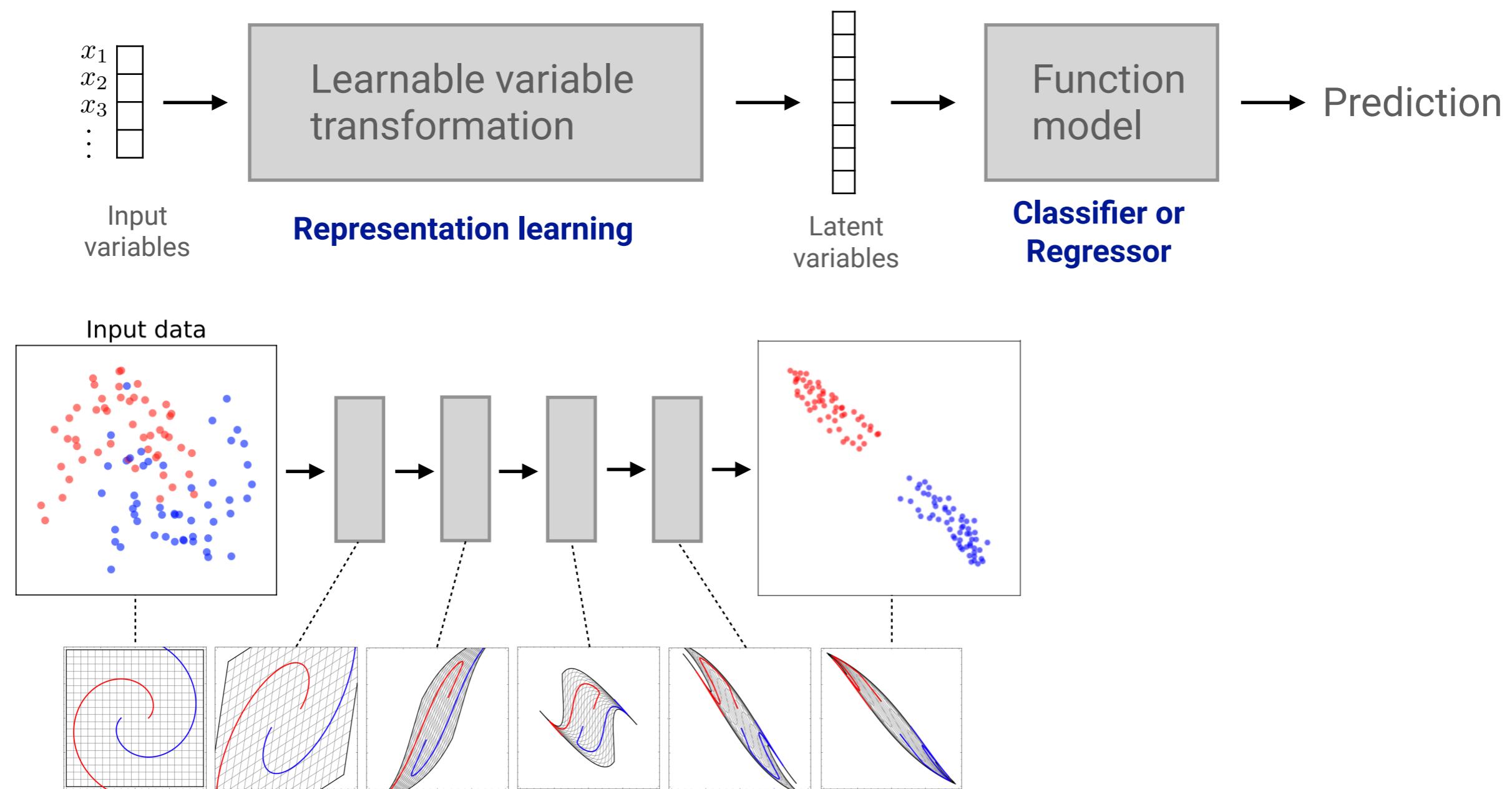
Representation Learning

Use **some inductive biases** to constrain/regularize the model space.



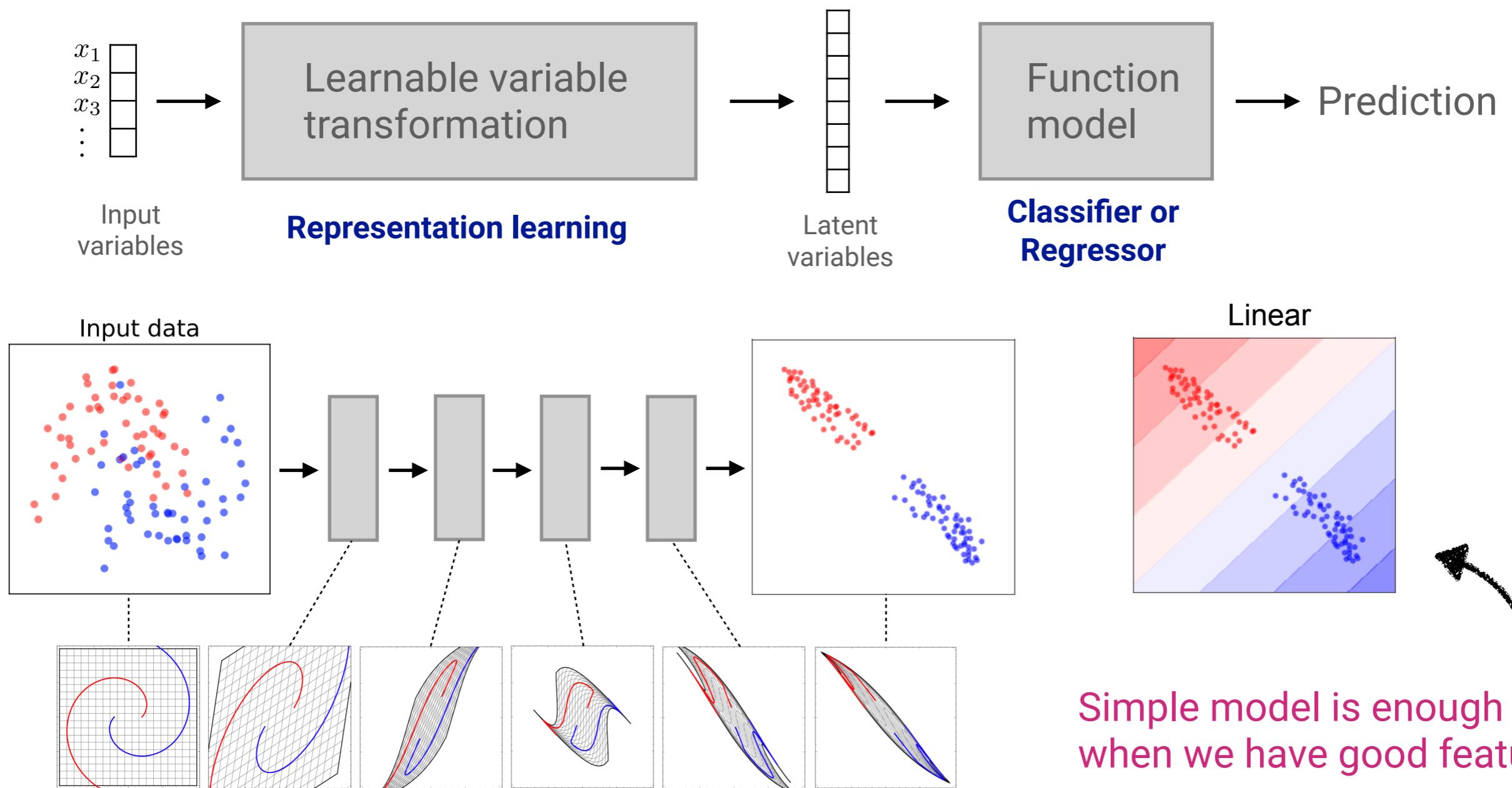
Representation Learning

Use **some inductive biases** to constrain/regularize the model space.



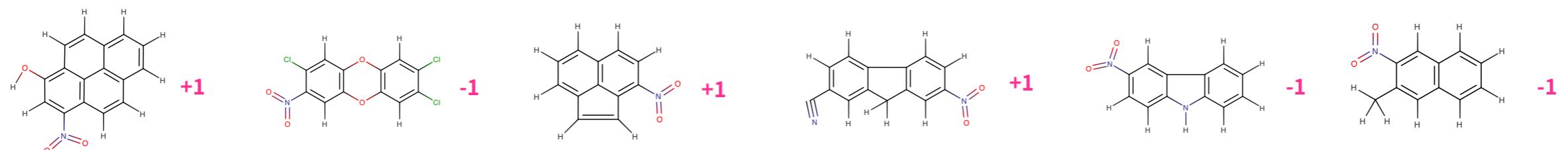
Representation Learning

Use **some inductive biases** to constrain/regularize the model space.

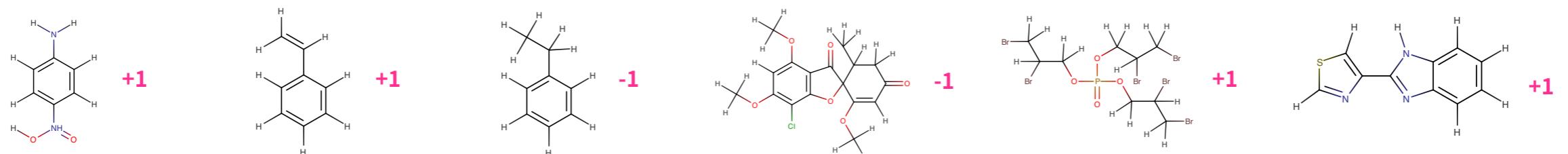


Use Case 1: Virtual Screening (QSAR/QSPR)

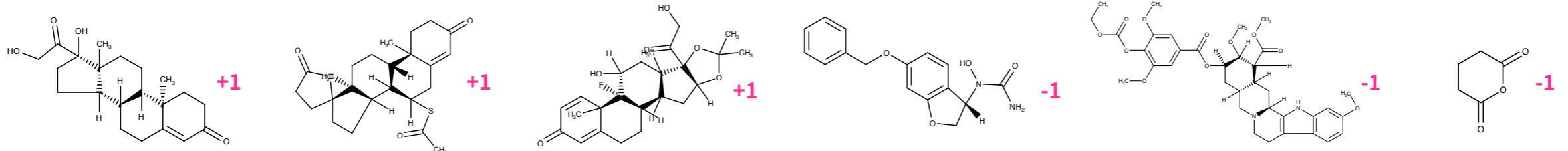
- Mutagenic potency**



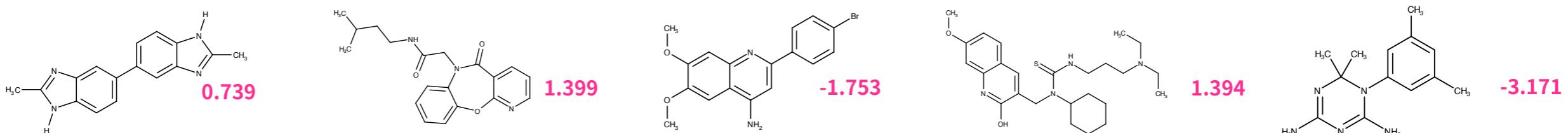
- Carcinogenic potency**



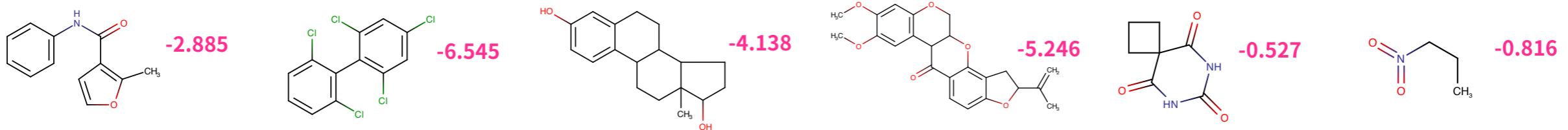
- Endocrine disruption**



- Growth inhibition**



- Aqueous solubility**



Use Case 1: Virtual Screening (QSAR/QSPR)

<https://pubchem.ncbi.nlm.nih.gov/bioassay/1>

BIOASSAY RECORD

NCI human tumor cell line growth inhibition assay. Data for the NCI-H23 Non-Small Cell Lung cell line

 About Blog Submit Contact

 Search PubChem

 Cite
 Download

CONTENTS

- Title and Summary**
- 1 Description
- 2 Comment
- 3 Result Definitions
- 4 Data Table
- 5 Entrez Crosslinks
- 6 Identity
- 7 BioAssay Annotations
- 8 Information Sources

PubChem AID 1

Source DTP/NCI

External ID [NCI human tumor cell line growth inhibition assay. Data for the NCI-H23 Non-Small Cell Lung cell line](#)

BioAssay Type Confirmatory

Tested Substances  All (53,554)  Active (3,025)  Inactive (50,655) [Data Table](#) 

Tested Compounds  All (51,583)  Active (2,814)  Inactive (48,922)

Version 2.1 [Revision History](#)

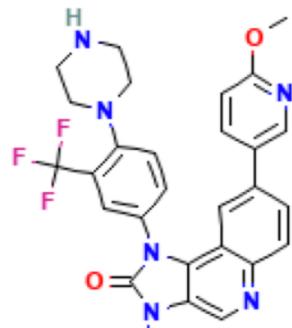
Status Live

Dates Modify 2021-07-12 Deposit 2004-08-15

Please note that the bioassay record (AID 1) is presented as provided to PubChem by the source(depositor). When possible, links to additional information have been provided by PubChem.

Use Case 1: Virtual Screening (QSAR/QSPR)

input



CID 11978790

ML

output

activity: "Active"
LogGI50: -7.8811

GI50: concentration required
for 50% inhibition of growth

Tested Compounds

All (51,583)

Active (2,814)

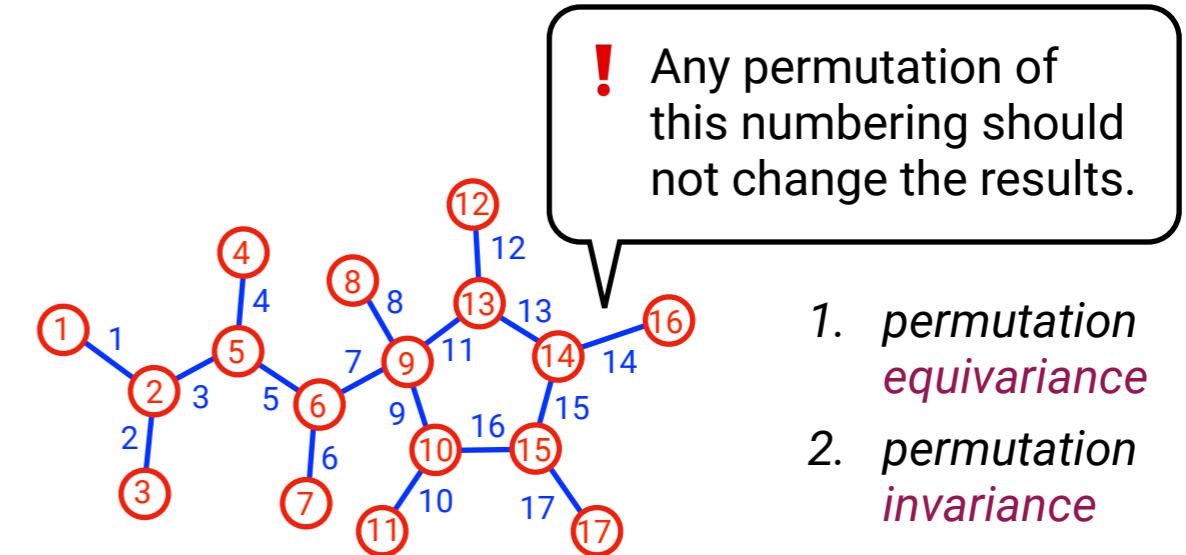
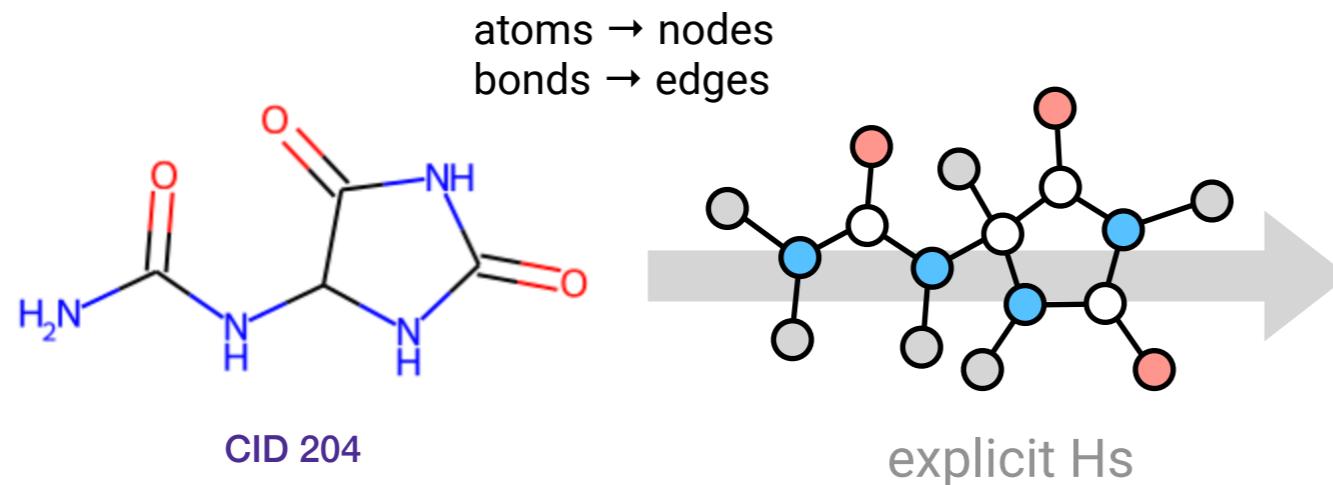
Inactive (48,922)

Tested Substance			Activity	Score	LogGI50_M ⓘ	LogGI50_u ⓘ	LogGI50_V ⓘ
Structure	CID	SID					
	5298	121832	Active	67	-8		
	363173	493713	Active	43	-6.5871		
	399631	530868	Active	51	-7.0678		
	399630	530867	Active	60	-7.617		

Tested Substance			Activity	Score	LogGI50_M ⓘ	LogGI50_u ⓘ	LogGI50_V ⓘ
Structure	CID	SID					
	390324	521601	Inactive	0	-4		
	390311	521588	Inactive	0	-4		
	390312	521589	Inactive	4	-4.214		
	135489876	521590	Inactive	13	-4.7552		

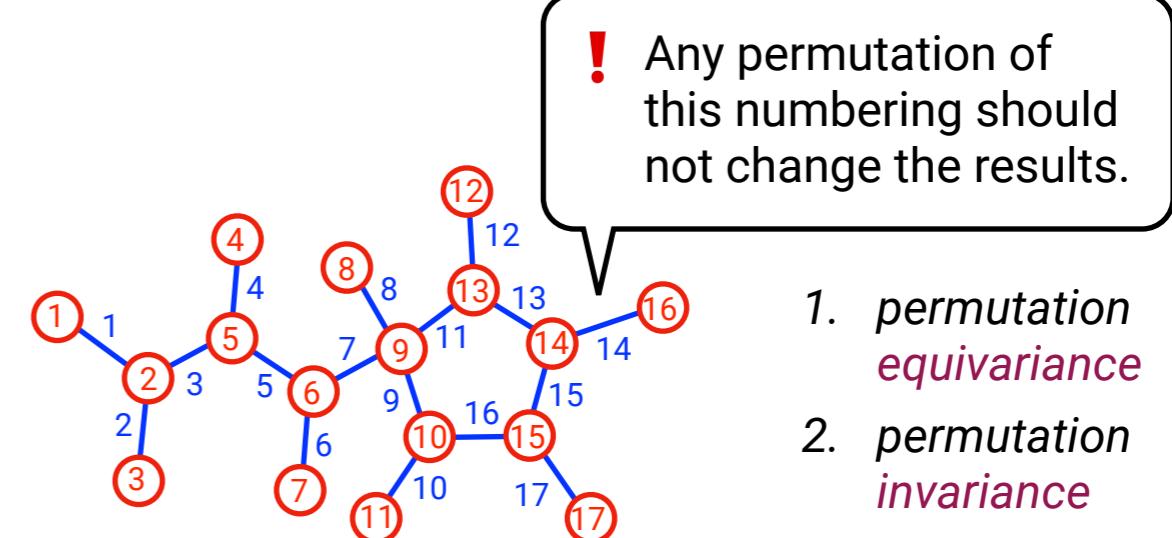
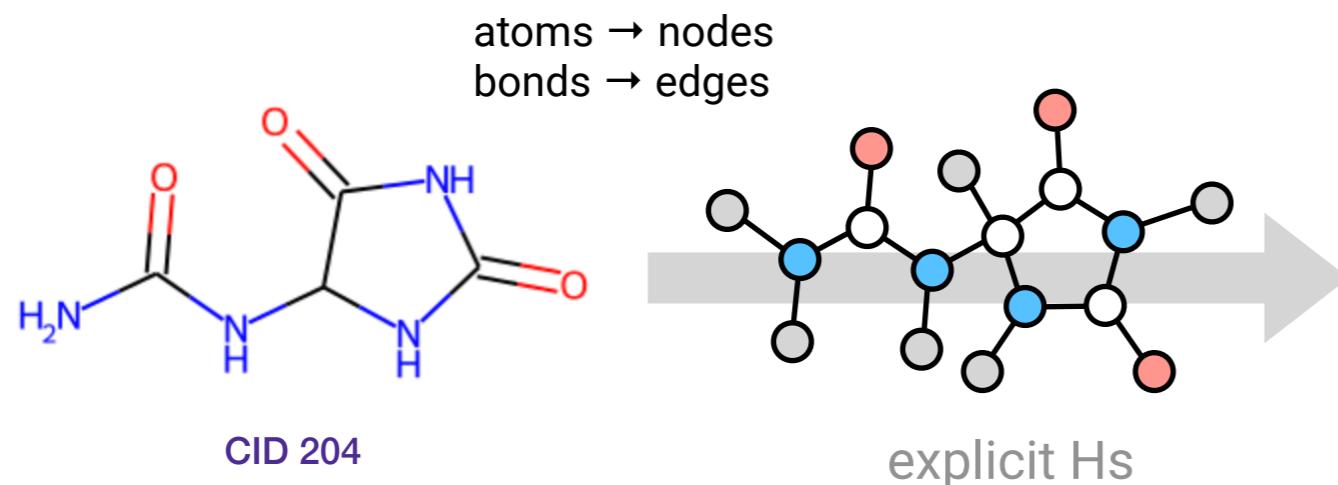
Use Case 1: Virtual Screening (QSAR/QSPR)

Input representation (molecular graph)



Use Case 1: Virtual Screening (QSAR/QSPR)

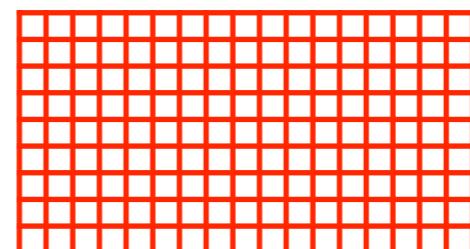
Input representation (molecular graph)



e.g. Features for ChemProp (Yang et al, 2019)

node(atom) features

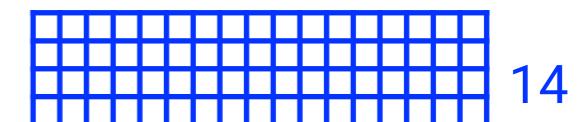
17



133

edge(bond) features

17



14

133 features

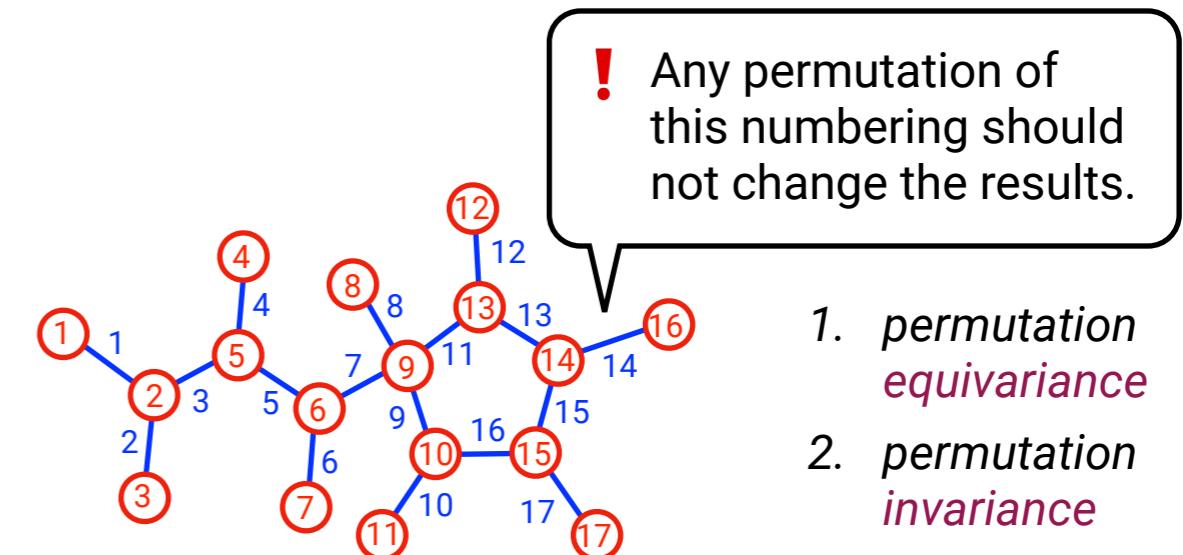
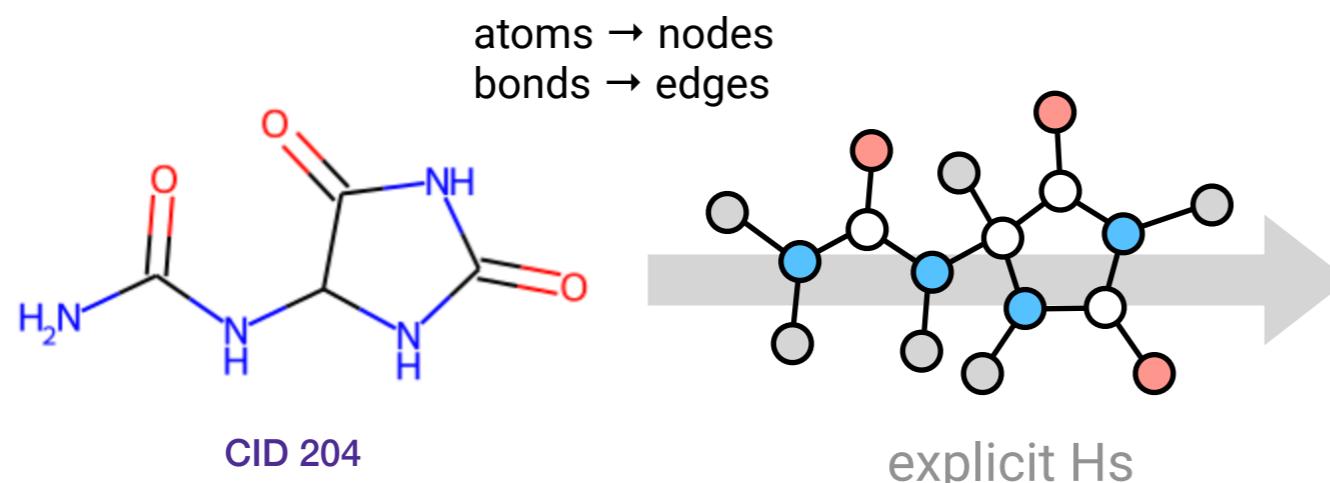
- atomic_num (one-hot, 101)
- total_degree (one-hot, 7)
- formal_charge (one-hot, 6)
- chiral_tag (one-hot, 5)
- num_Hs (one-hot, 6)
- hybridization (one-hot, 6)
- is_aromatic (binary, 1)
- atomic_mass (real, 1)

14 features

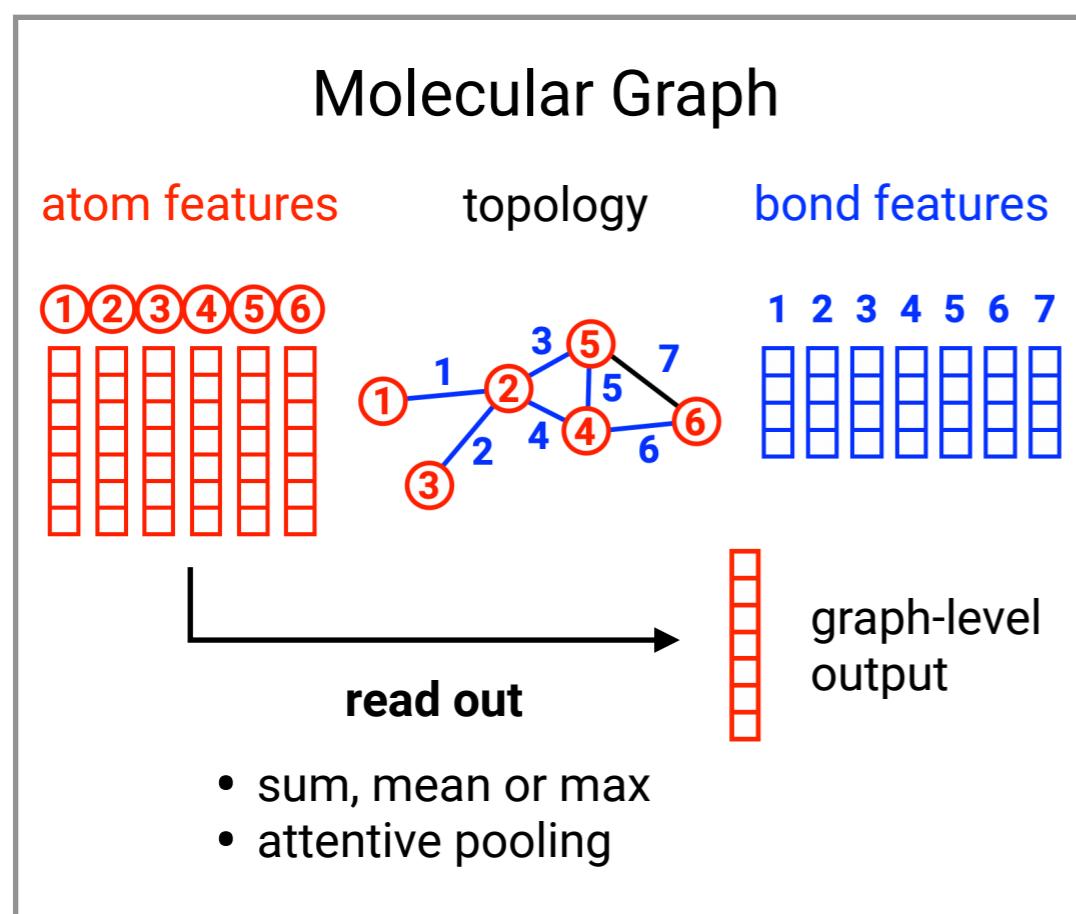
- no_bond (binary, 1)
- is_single (binary, 1)
- is_double (binary, 1)
- is_triple (binary, 1)
- is_aromatic (binary, 1)
- is_connjugated (binary, 1)
- is_in_ring (binary, 1)
- stereo (one-hot, 7)

Use Case 1: Virtual Screening (QSAR/QSPR)

Input representation (molecular graph)

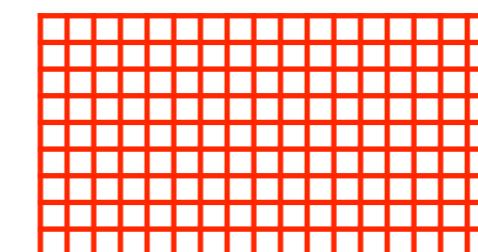


e.g. Features for ChemProp (Yang et al, 2019)

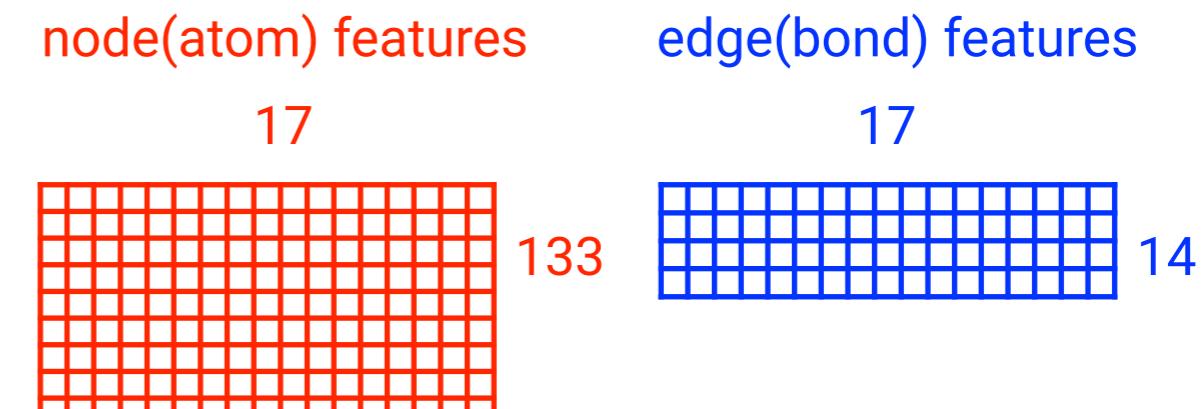


node(atom) features

17



133



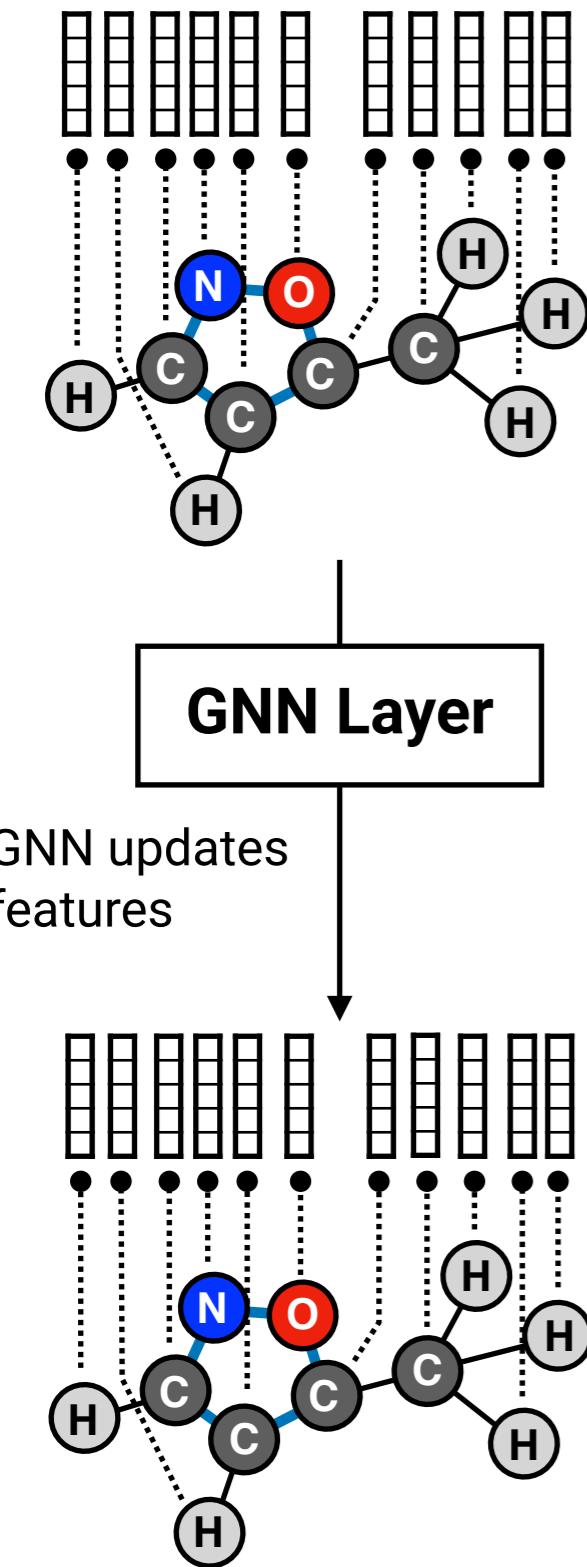
133 features

- atomic_num (one-hot, 101)
- total_degree (one-hot, 7)
- formal_charge (one-hot, 6)
- chiral_tag (one-hot, 5)
- num_Hs (one-hot, 6)
- hybridization (one-hot, 6)
- is_aromatic (binary, 1)
- is_connjugated (binary, 1)
- is_in_ring (binary, 1)
- stereo (one-hot, 7)

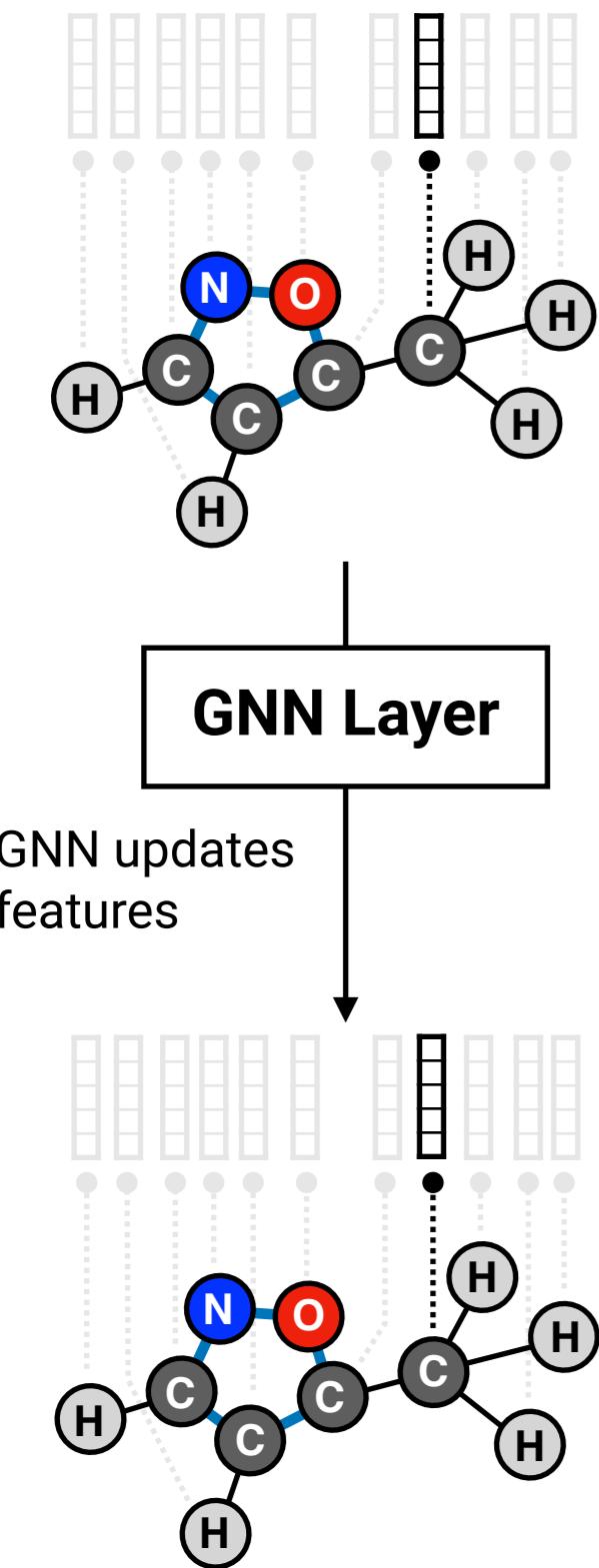
14 features

- no_bond (binary, 1)
- is_single (binary, 1)
- is_double (binary, 1)
- is_triple (binary, 1)
- is_aromatic (binary, 1)
- is_connjugated (binary, 1)
- is_in_ring (binary, 1)
- stereo (one-hot, 7)

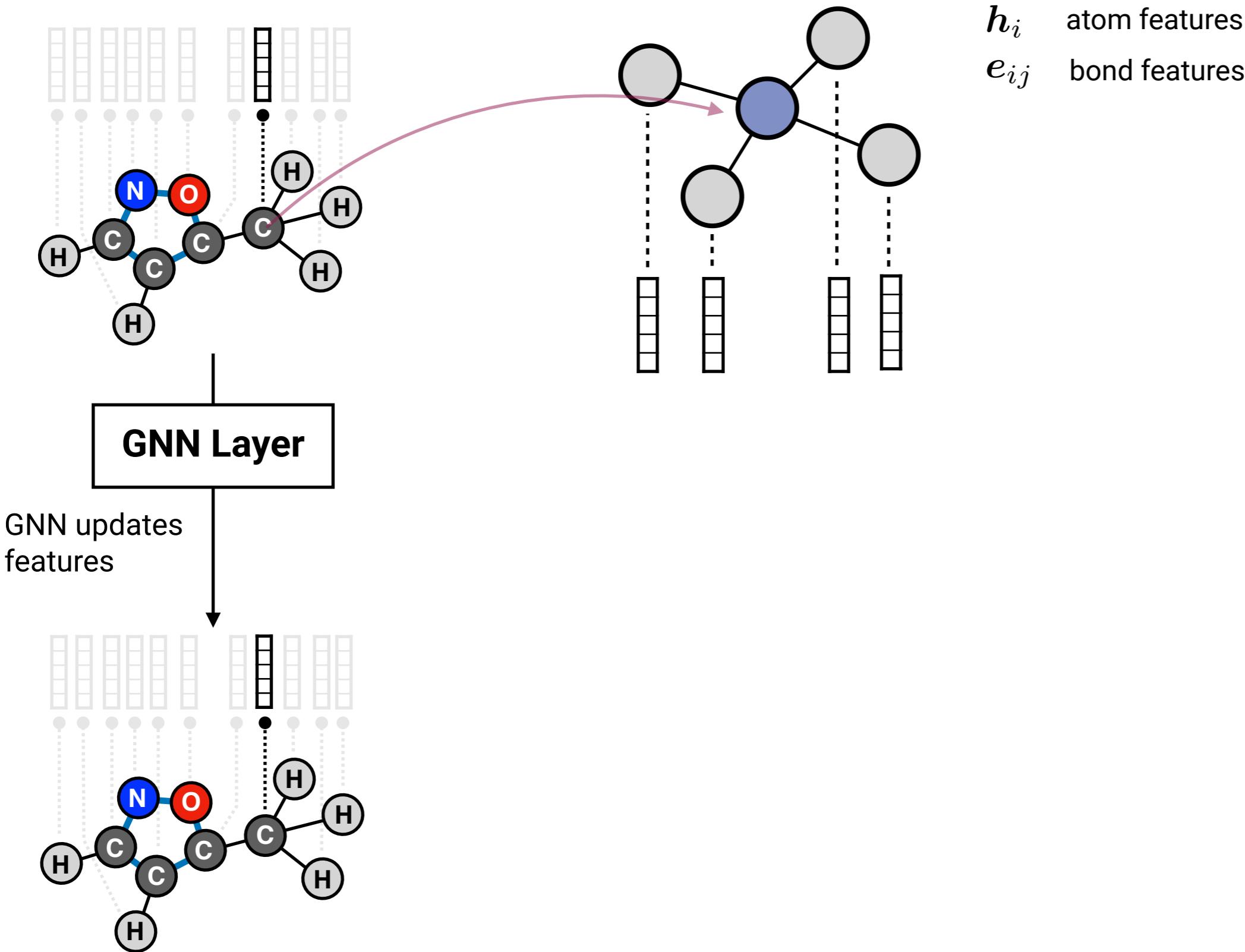
Graph Neural Networks (GNNs)



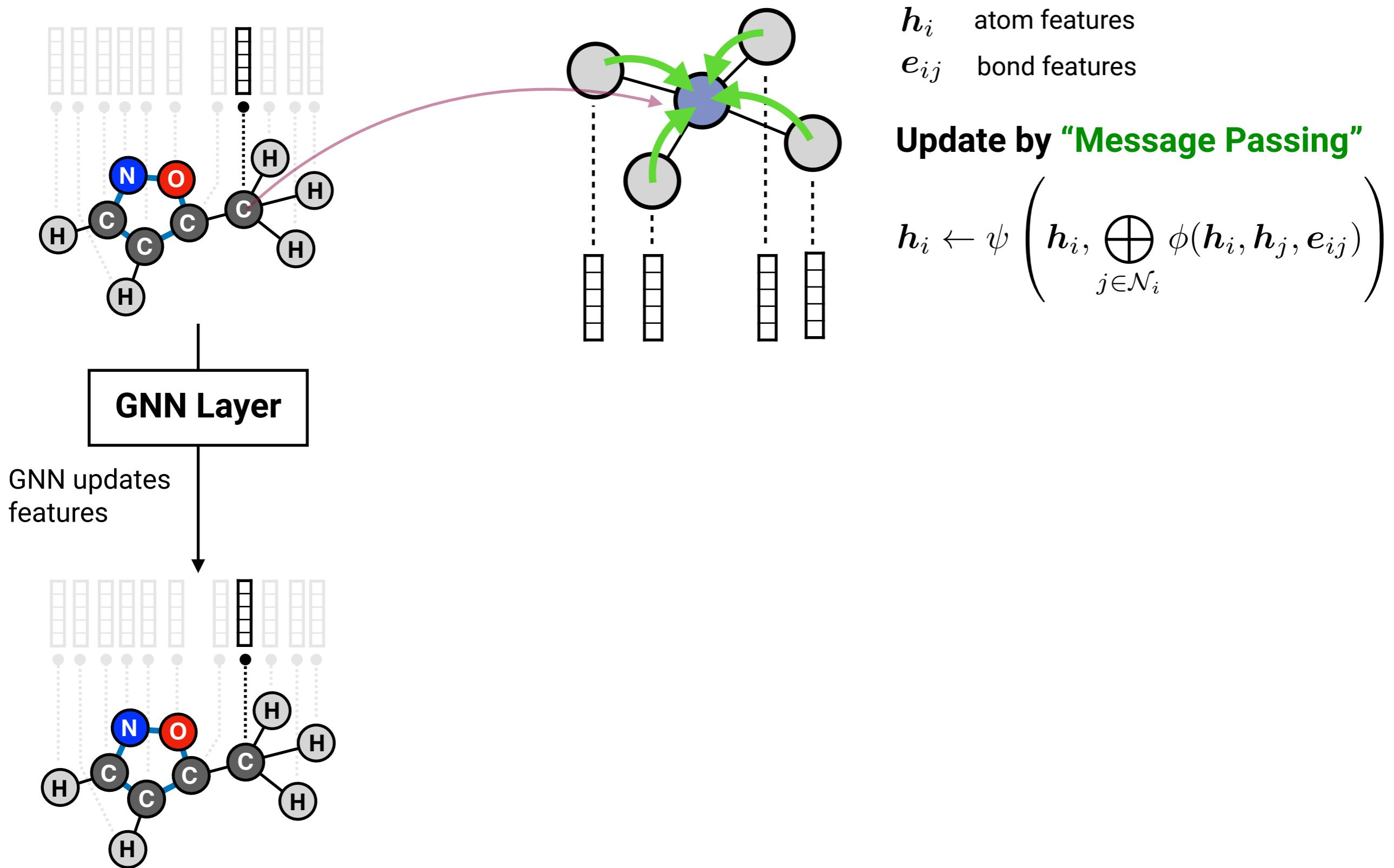
Graph Neural Networks (GNNs)



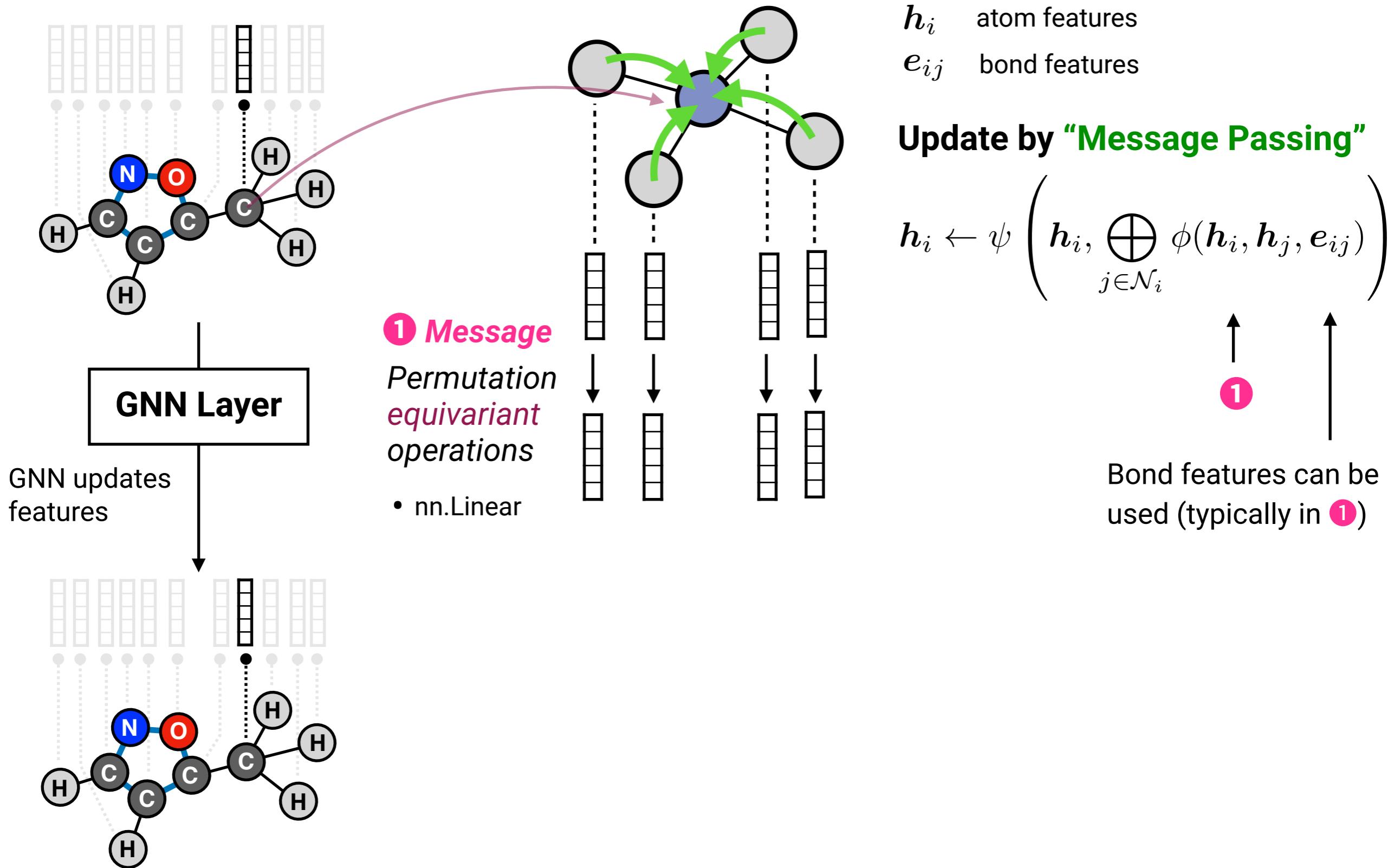
Graph Neural Networks (GNNs)



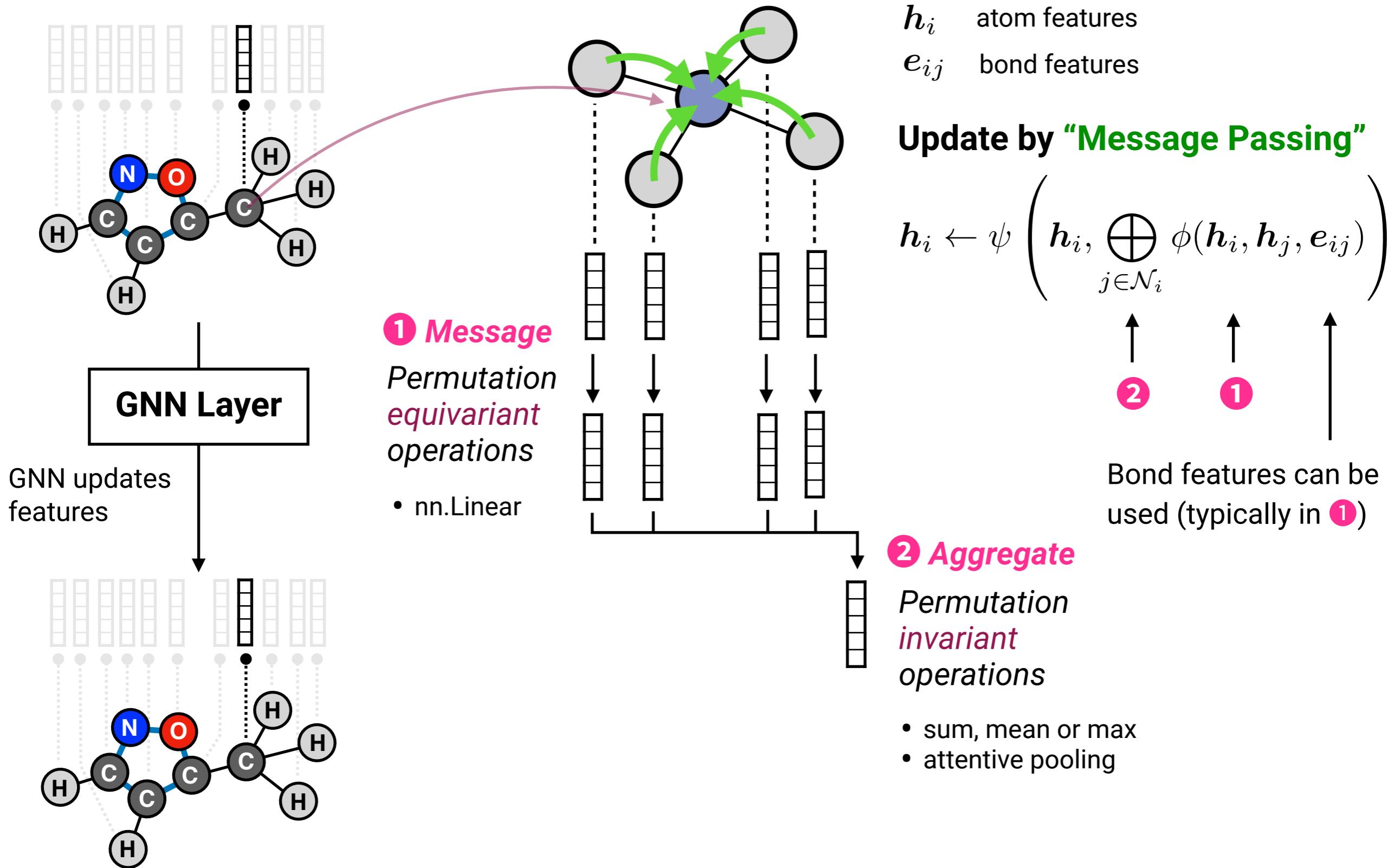
Graph Neural Networks (GNNs)



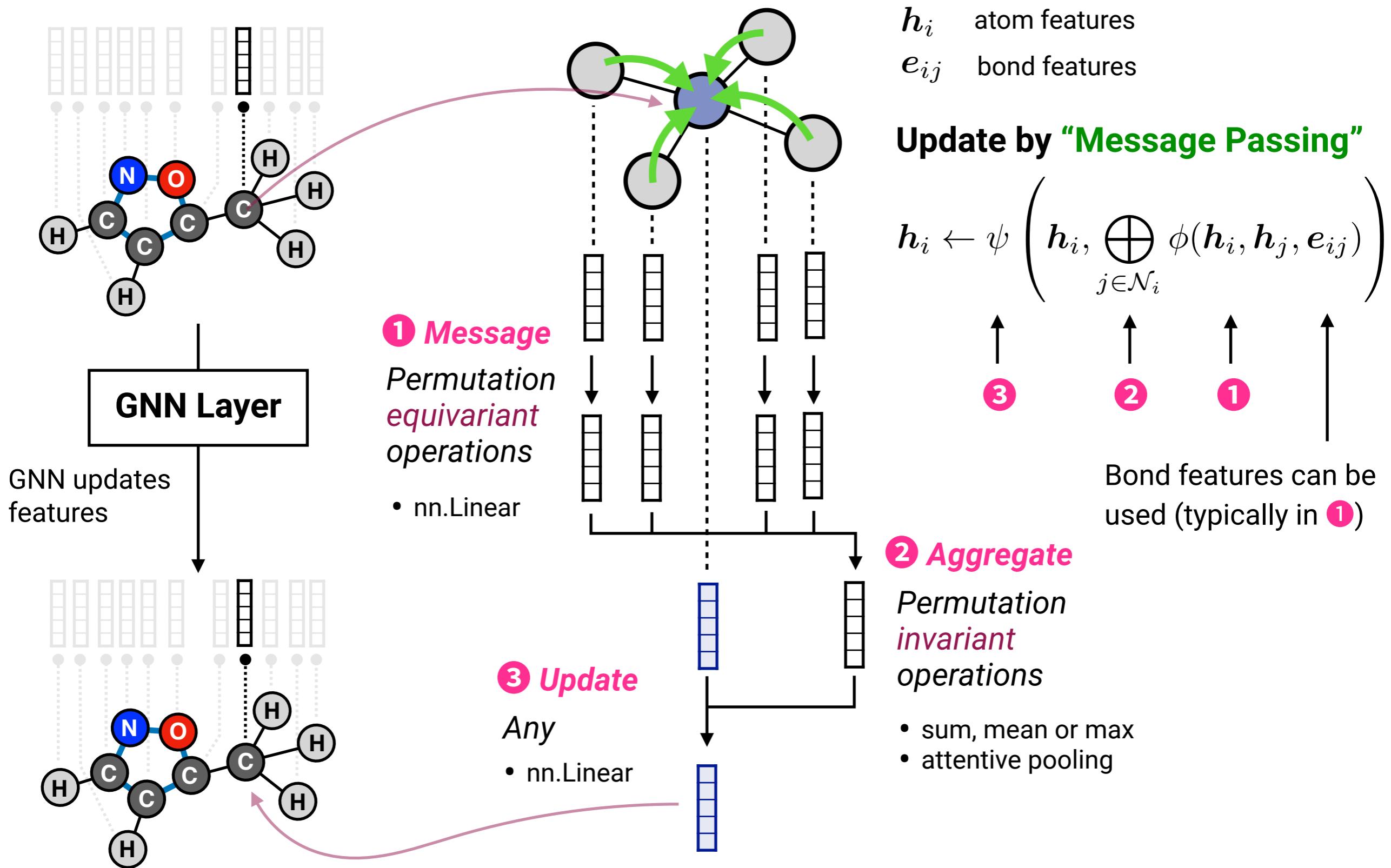
Graph Neural Networks (GNNs)



Graph Neural Networks (GNNs)



Graph Neural Networks (GNNs)



Use Case 1: Virtual Screening (QSAR/QSPR)

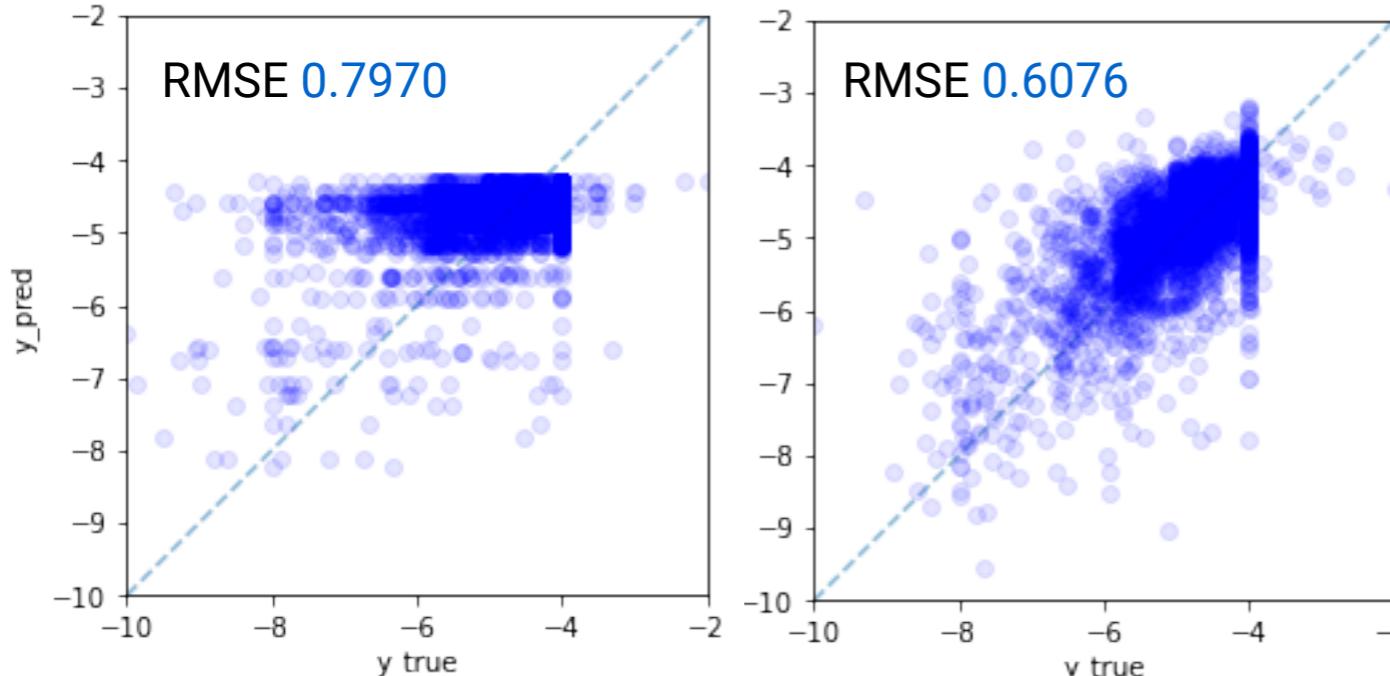
Performance for **unseen (test) data:**

Active/Inactive (Classification), LogGI50 (Regression)

Standard ML

ExtraTrees
w/ ECFP6(1024)

- Classification accuracy
95.079% (Active/Inactive)
- Regression for LogGI50



Disclaimer: This is just for a toy demo. This should be taken as classification for ACTIVITY_OUTCOME (Active or Inactive)

GNN

ChemProp
(Directed MPNN)

- Classification accuracy
95.604% (Active/Inactive)
- Regression for LogGI50

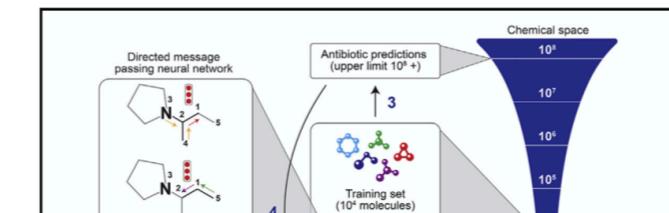
ChemProp (Yang et al, 2019)

from MIT MLPDS (Machine Learning for Pharmaceutical Discovery and Synthesis) Consortium

Cell

A Deep Learning Approach to Antibiotic Discovery

Graphical Abstract



Authors

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, ..., Tommi S. Jaakkola, Regina Barzilay, James J. Collins

Correspondence

regina@csail.mit.edu (R.B.), jimjc@mit.edu (J.J.C.)

Stokes et al, *Cell* (2020) <https://doi.org/10.1016/j.cell.2020.01.021>

nature

NEWS | 20 February 2020

Powerful antibiotics discovered using AI

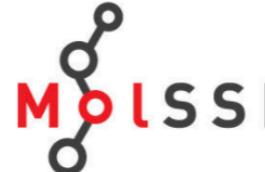
Machine learning spots molecules that work even against ‘untreatable’ strains of bacteria.

Jo Marchant

Marchant, *Nature* (2020) <https://doi.org/10.1038/d41586-020-00018-3>

Use Case 2: Quantum chemistry

https://qcarchive.molssi.org/apps/ml_datasets/


Machine Learning Datasets Repository

Add your Dataset
License

Name	↑↓	Quality	↑↓	Data Points	↑↓	Elements	↑↓	Sampling	↑↓	Download
+ ANI-1	↑↓	DFT	↑↓	22,057,374	↑↓	C H N O	↑↓	NMS	↑↓	Download HDF5 Download TEXT
+ ANI-1x	↑↓	DFT	↑↓	4,956,005	↑↓	C H N O	↑↓	MD,NMS,DS,TS	↑↓	Download HDF5
- QM9	↑↓	DFT	↑↓	133,885	↑↓	C H F N O	↑↓	Minima	↑↓	Download HDF5 Download TEXT

Description

Small organic molecules with up to 9 heavy atoms sampled from GDB-17, optimized at the B3LYP/6-31G(2df,p) level of theory. Ground state, orbital, and thermodynamic properties are available (at the B3LYP/6-31G(2df,p) level). All molecules are neutral singlets. This dataset was sourced from [quantum-machine.org](#) and [qmml.org](#).

Elements: C H F N O

Labels

energy homo lumo polarizability dipole frequency zpve
enthalpy free energy heat capacity rotational constant

Tags

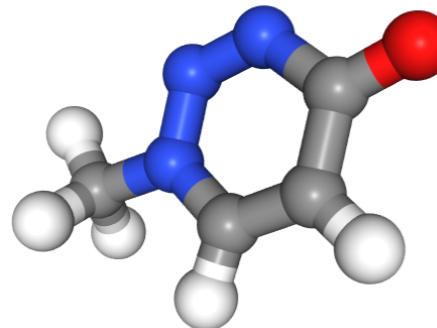
organic thermodynamics GDB

Citations

- Blum, L. C. & Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13.

Use Case 2: Quantum chemistry

input



gdb_21014

	x	y	z
O	0.314096	-0.129589	-0.389150
C	0.111219	2.102676	-0.051749
C	2.331344	3.941075	0.212303
O	4.667017	2.677399	0.437948
C	6.152491	3.062553	-1.780599
C	4.732264	5.009654	-3.282819
C	2.562527	5.549427	-2.143825
H	-1.771427	3.048695	0.071772
H	1.977918	5.086871	1.919865
H	8.050245	3.696867	-1.222422
H	6.372399	1.276980	-2.825015
H	5.428656	5.805758	-5.033531
H	1.118529	6.857080	-2.763050

~ 1000 sec

Quantum chemical calculations

by solving a one-electron Schrödinger equation (Kohn–Sham equation)

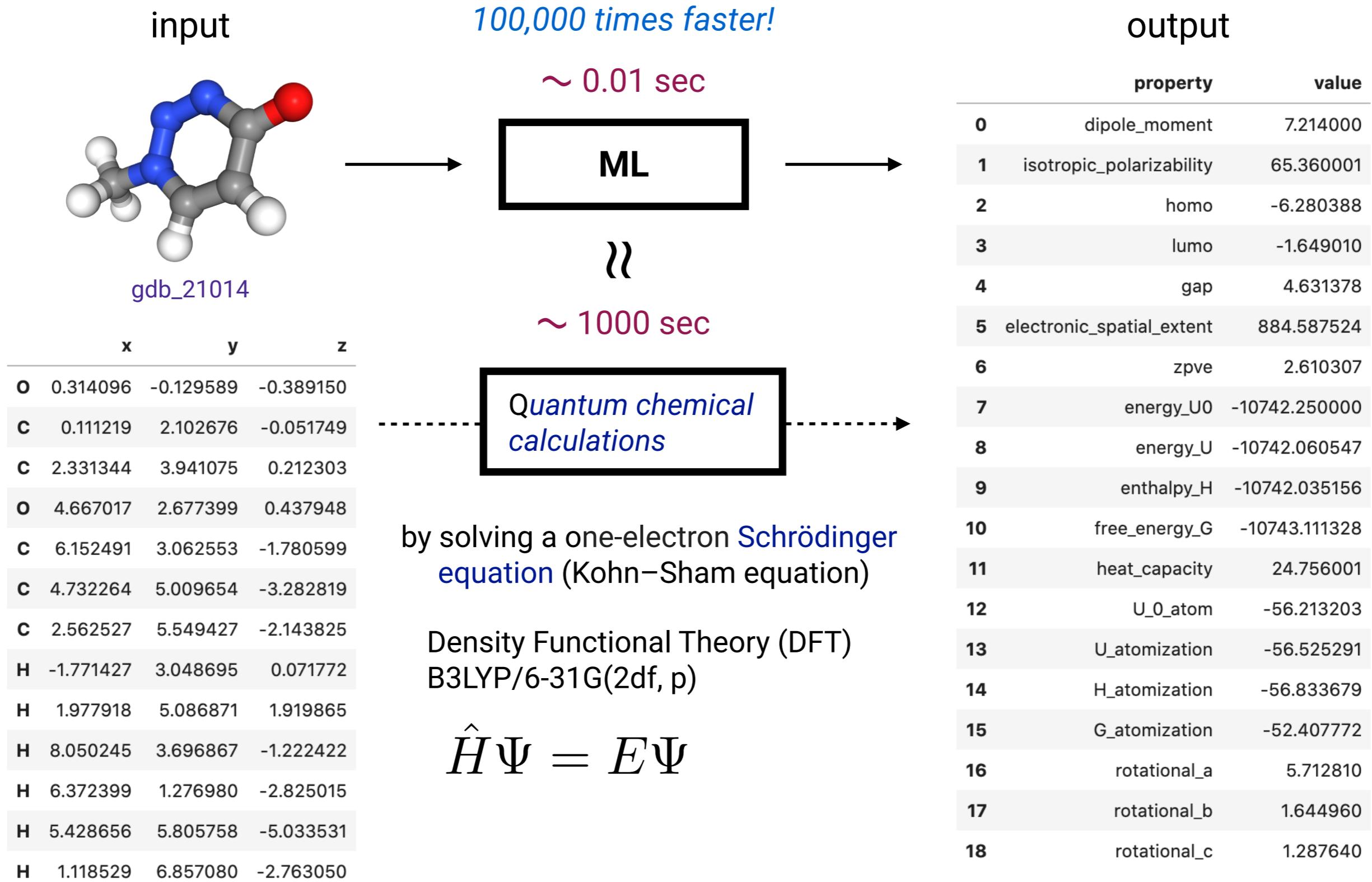
Density Functional Theory (DFT)
B3LYP/6-31G(2df, p)

$$\hat{H}\Psi = E\Psi$$

output

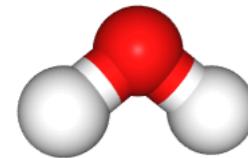
	property	value
0	dipole_moment	7.214000
1	isotropic_polarizability	65.360001
2	homo	-6.280388
3	lumo	-1.649010
4	gap	4.631378
5	electronic_spatial_extent	884.587524
6	zpve	2.610307
7	energy_U0	-10742.250000
8	energy_U	-10742.060547
9	enthalpy_H	-10742.035156
10	free_energy_G	-10743.111328
11	heat_capacity	24.756001
12	U_0_atom	-56.213203
13	U_atomization	-56.525291
14	H_atomization	-56.833679
15	G_atomization	-52.407772
16	rotational_a	5.712810
17	rotational_b	1.644960
18	rotational_c	1.287640

Use Case 2: Quantum chemistry



Use Case 2: Quantum chemistry

input molecule H₂O



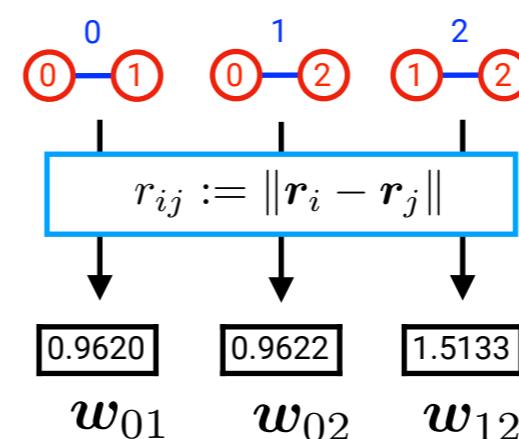
gdb_3

	x	y	z	AN
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1

atom features

$$\begin{array}{ccc} \textcircled{0} & \textcircled{1} & \textcircled{2} \\ \boxed{8} & \boxed{1} & \boxed{1} \\ \boldsymbol{x}_0 & \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{array}$$

bond features

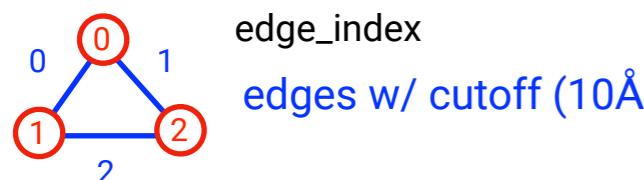


SchNet (Schütt et al, 2017)

Message Passing with residual connections

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \psi \left(\sum_{j \in \mathcal{N}_i} \phi(\boldsymbol{x}_j) \odot \omega_{ij} \right)$$

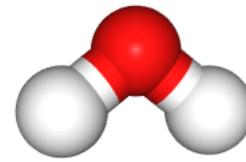
graph (SchNet)



$$\begin{aligned} \mathbf{r}_0 &= [-0.0344, 0.9775, 0.0076] & Z_0 &= 8 \\ \mathbf{r}_1 &= [0.0648, 0.0206, 0.0015] & Z_1 &= 1 \\ \mathbf{r}_2 &= [0.8718, 1.3008, 0.0007] & Z_2 &= 1 \end{aligned}$$

Use Case 2: Quantum chemistry

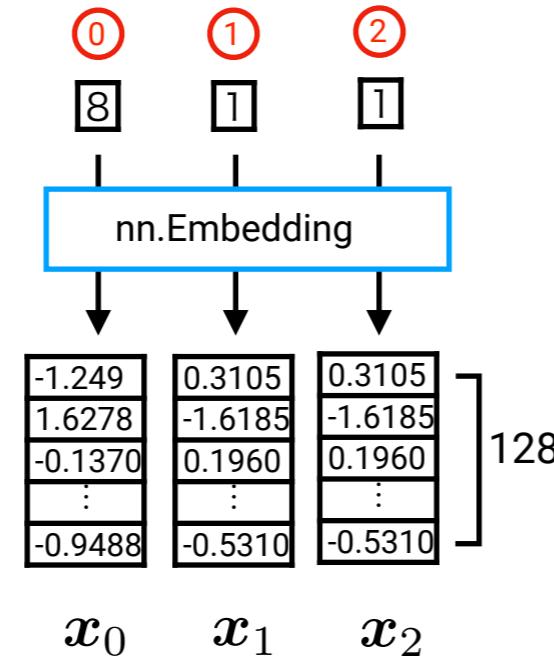
input molecule H₂O



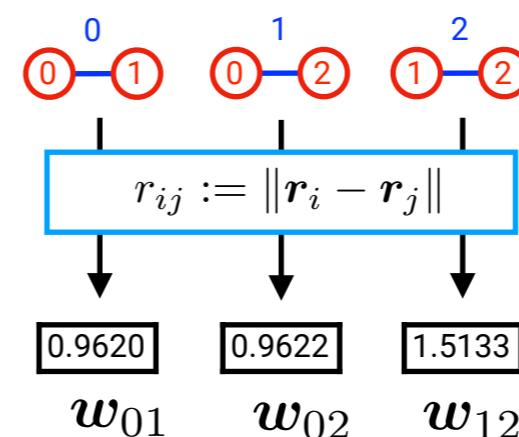
gdb_3

	x	y	z	AN
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1

atom features



bond features

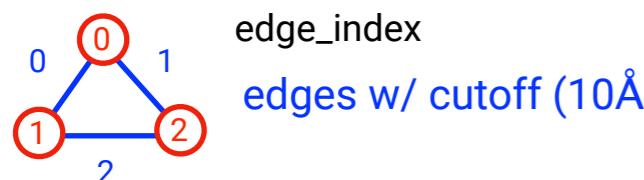


SchNet (Schütt et al, 2017)

Message Passing with residual connections

$$\mathbf{x}_i \leftarrow \mathbf{x}_i + \psi \left(\sum_{j \in \mathcal{N}_i} \phi(\mathbf{x}_j) \odot \omega_{ij} \right)$$

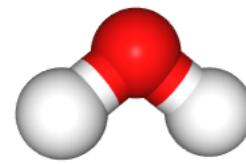
graph (SchNet)



$$\begin{aligned} \mathbf{r}_0 &= [-0.0344, 0.9775, 0.0076] & Z_0 &= 8 \\ \mathbf{r}_1 &= [0.0648, 0.0206, 0.0015] & Z_1 &= 1 \\ \mathbf{r}_2 &= [0.8718, 1.3008, 0.0007] & Z_2 &= 1 \end{aligned}$$

Use Case 2: Quantum chemistry

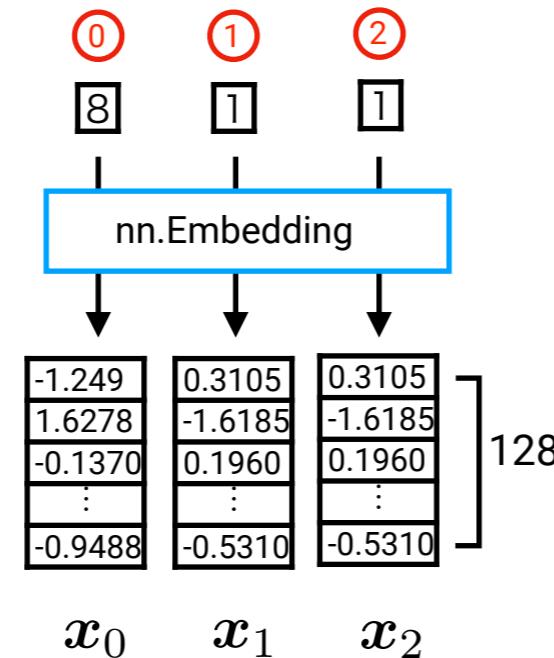
input molecule H₂O



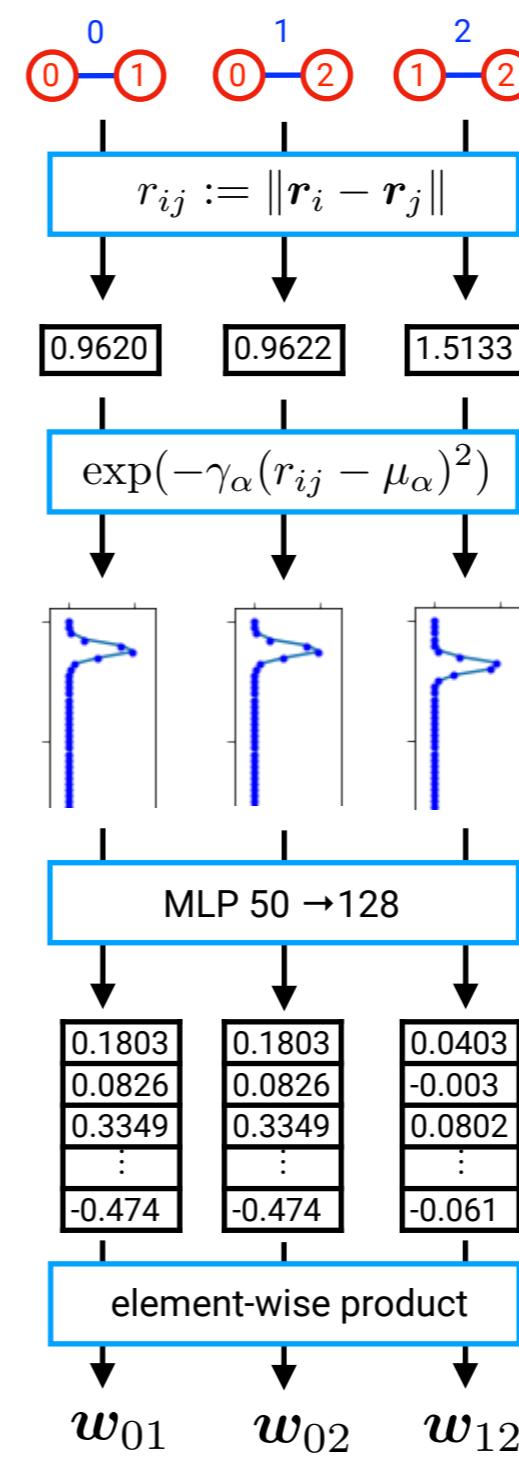
gdb_3

	x	y	z	AN
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1

atom features



bond features

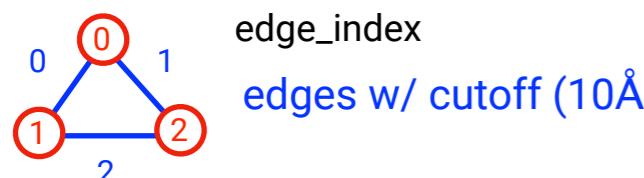


SchNet (Schütt et al, 2017)

Message Passing with residual connections

$$\mathbf{x}_i \leftarrow \mathbf{x}_i + \psi \left(\sum_{j \in \mathcal{N}_i} \phi(\mathbf{x}_j) \odot \omega_{ij} \right)$$

graph (SchNet)



edge_index
edges w/ cutoff (10 Å)

$$\begin{aligned} \mathbf{r}_0 &= [-0.0344, 0.9775, 0.0076] & Z_0 &= 8 \\ \mathbf{r}_1 &= [0.0648, 0.0206, 0.0015] & Z_1 &= 1 \\ \mathbf{r}_2 &= [0.8718, 1.3008, 0.0007] & Z_2 &= 1 \end{aligned}$$

cutoff function

$$\frac{1}{2} \left(\cos \left(\frac{\pi r_{ij}}{r_c} \right) + 1 \right)$$

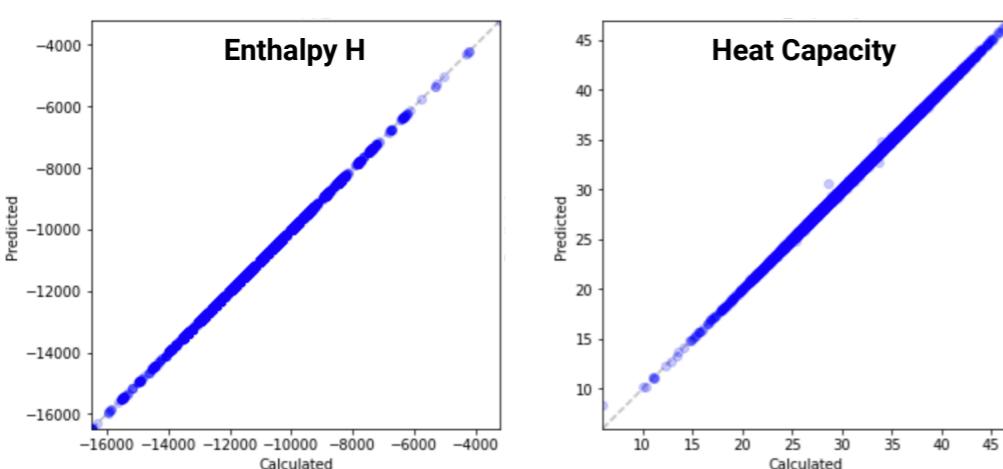
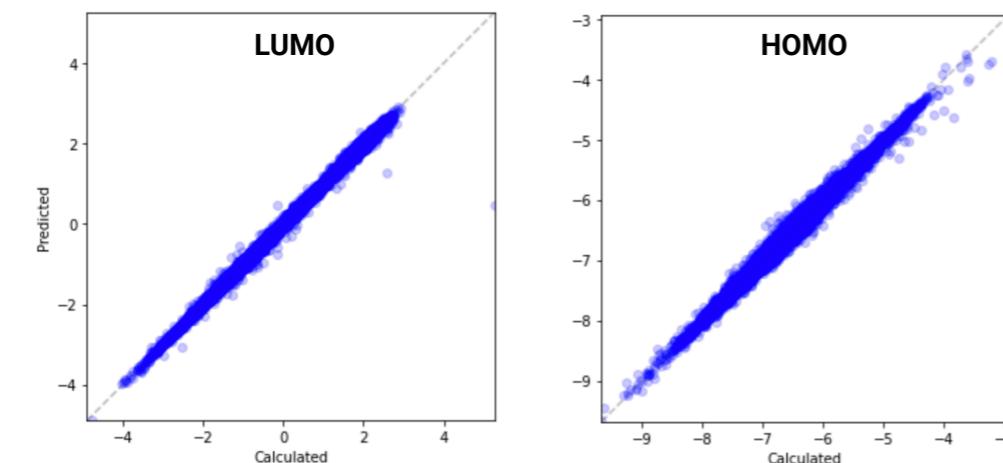
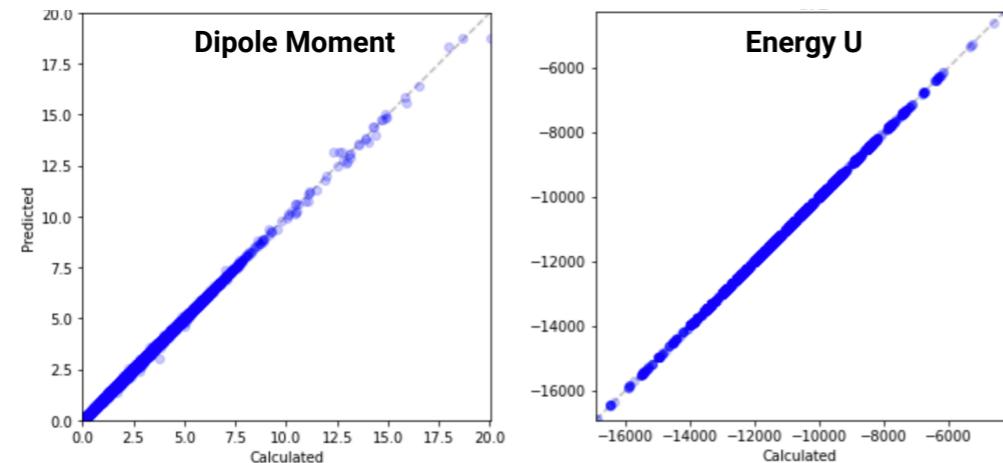
50

128

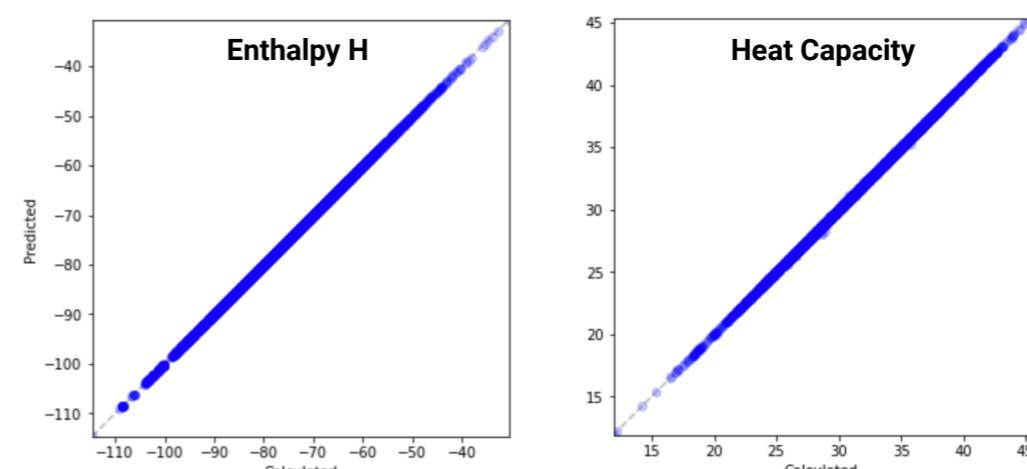
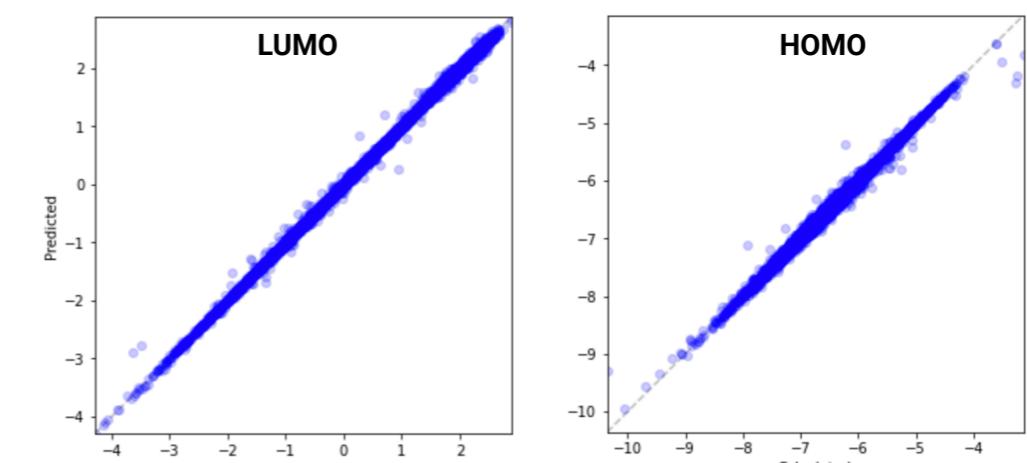
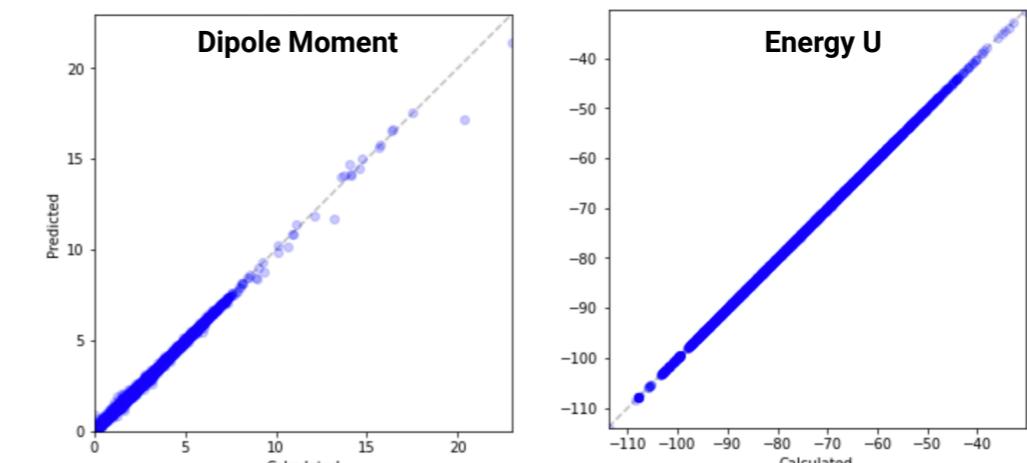
Weighted ACSFs (ACSFs = atom-centered symmetry functions)
for Behler-Parrinello potentials

Use Case 2: Quantum chemistry

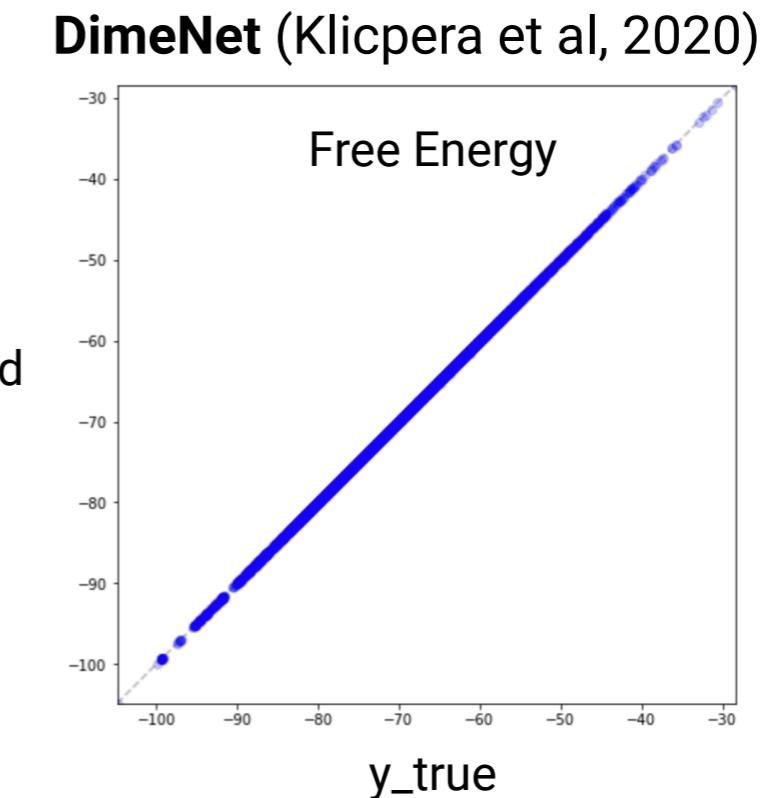
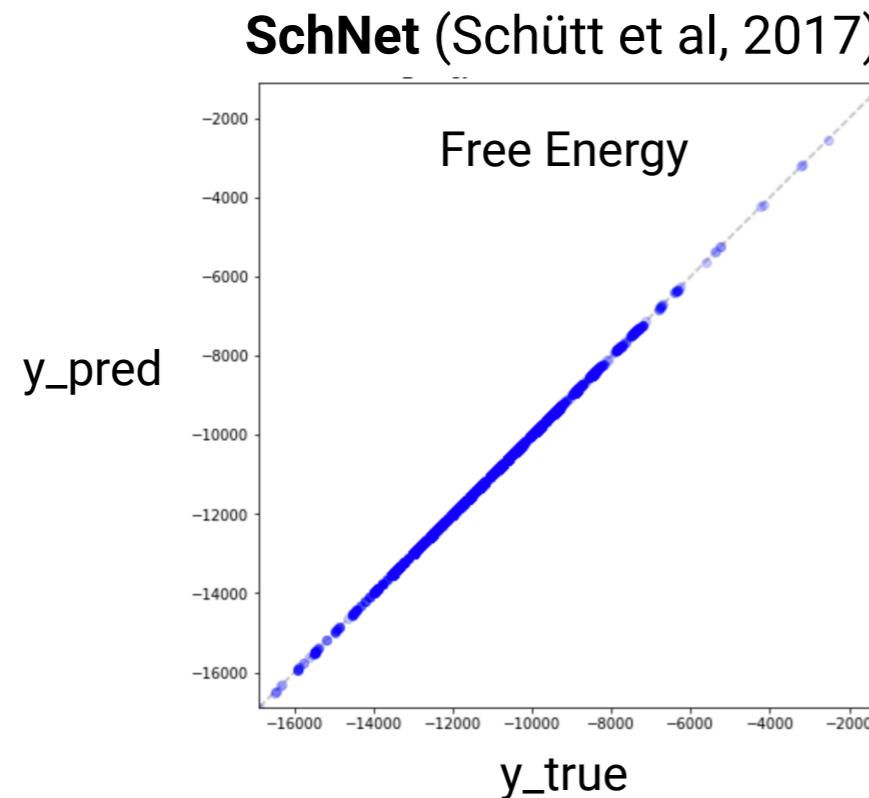
pred vs true for **SchNet** (Schütt et al, 2017)



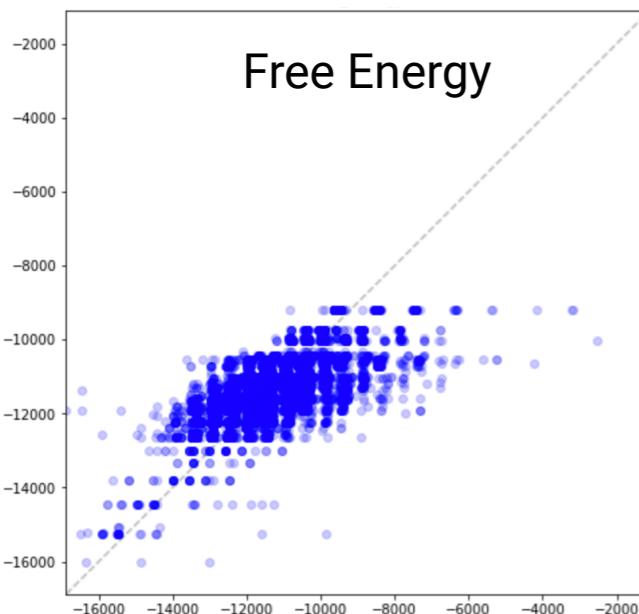
pred vs true for **DimeNet** (Klicpera et al, 2020)



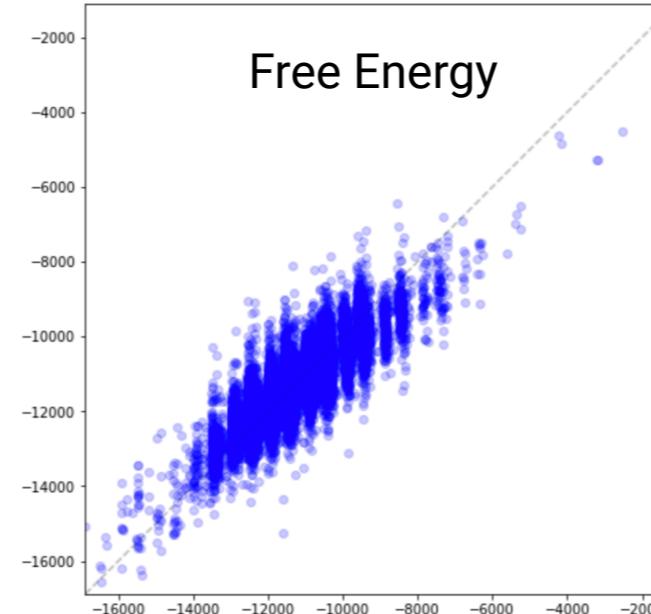
Use Case 2: Quantum chemistry



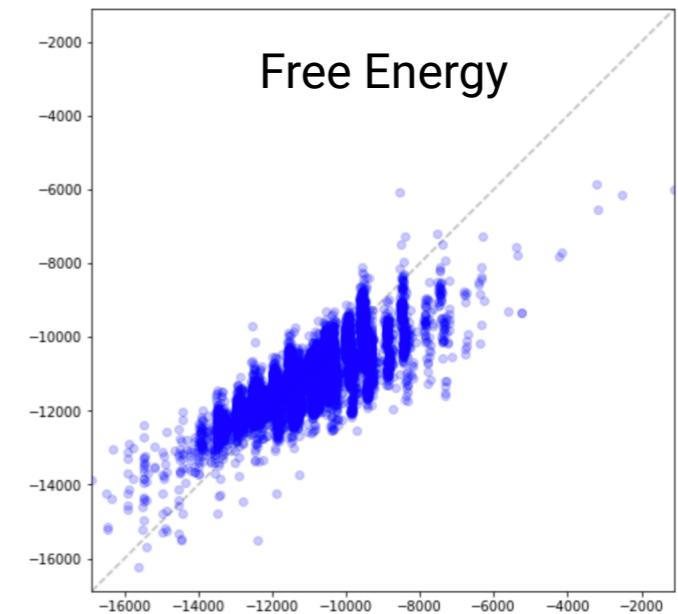
ExtraTrees w/ ECFP6
(without 3D geometry)



LightGBM w/ ECFP6
(without 3D geometry)



3-Layer MLP w/ ECFP6
(without 3D geometry)

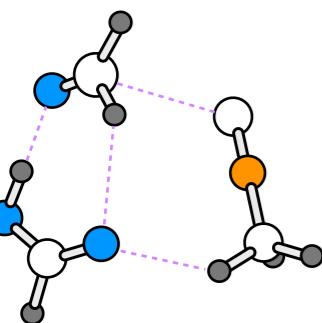


Chemical Reaction Design and Discovery



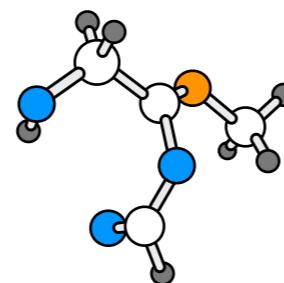
Chemical Reaction

EQ1



How we can have this?

EQ2

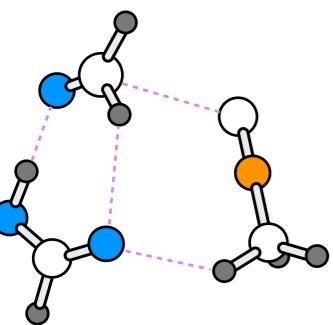


Chemical Reaction Design and Discovery

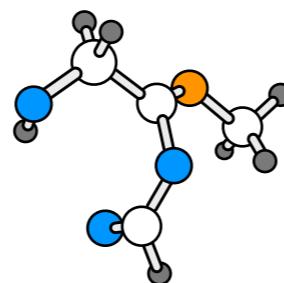


Chemical Reaction

EQ1



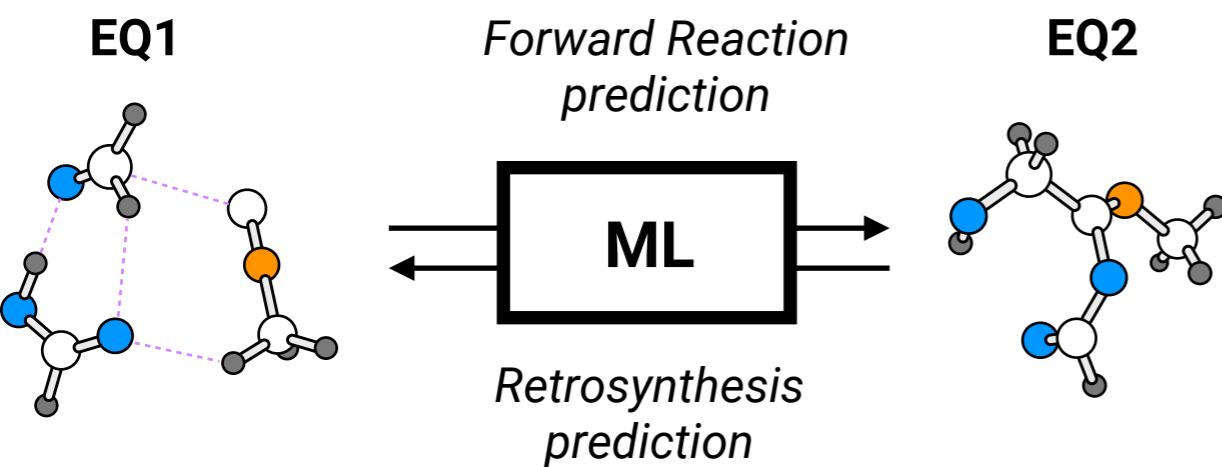
EQ2



Chemical Reaction Design and Discovery



Chemical Reaction

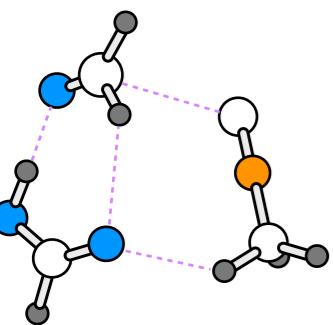


Chemical Reaction Design and Discovery

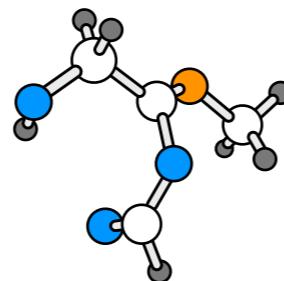


Chemical Reaction

EQ1



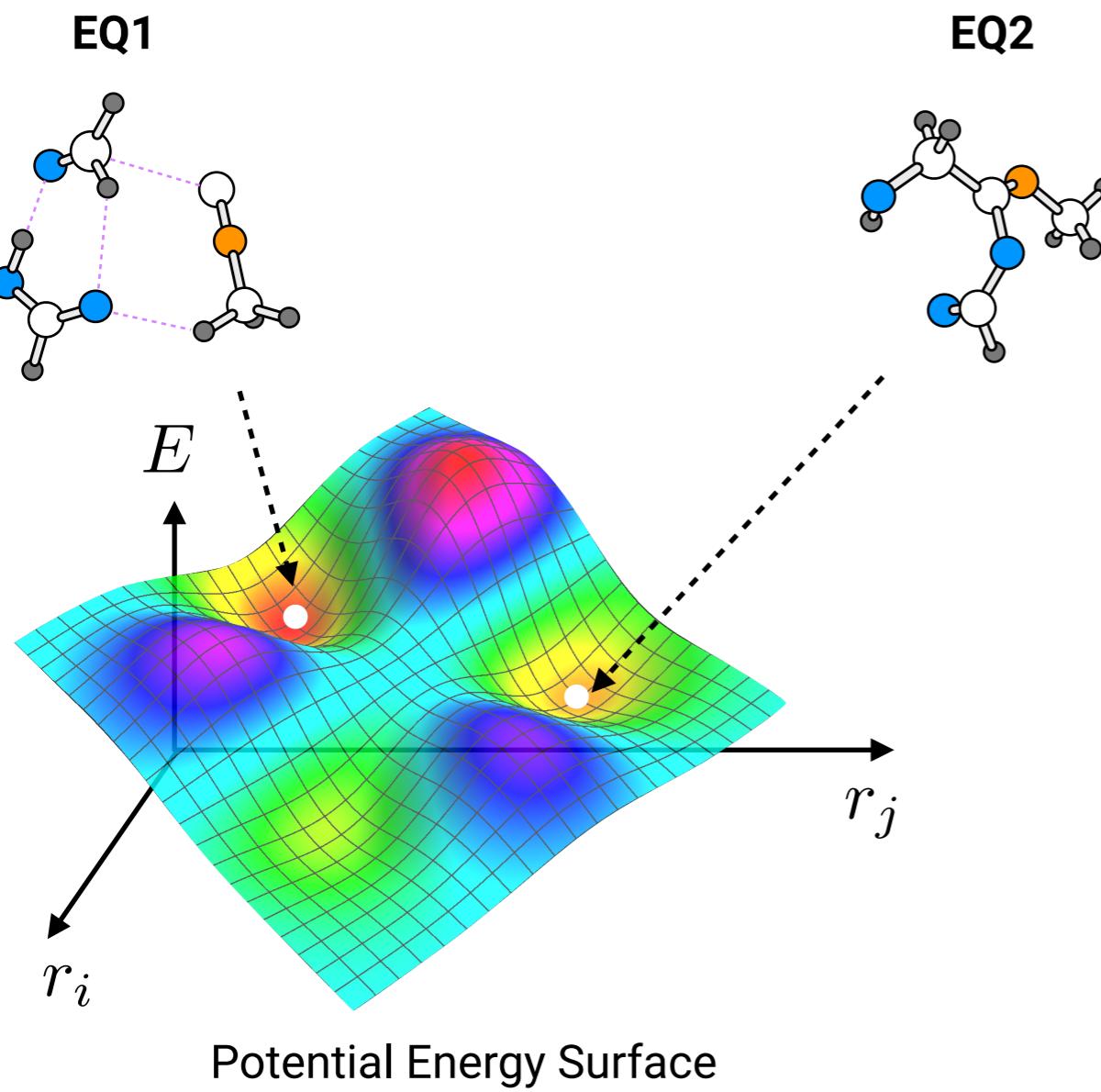
EQ2



Chemical Reaction Design and Discovery



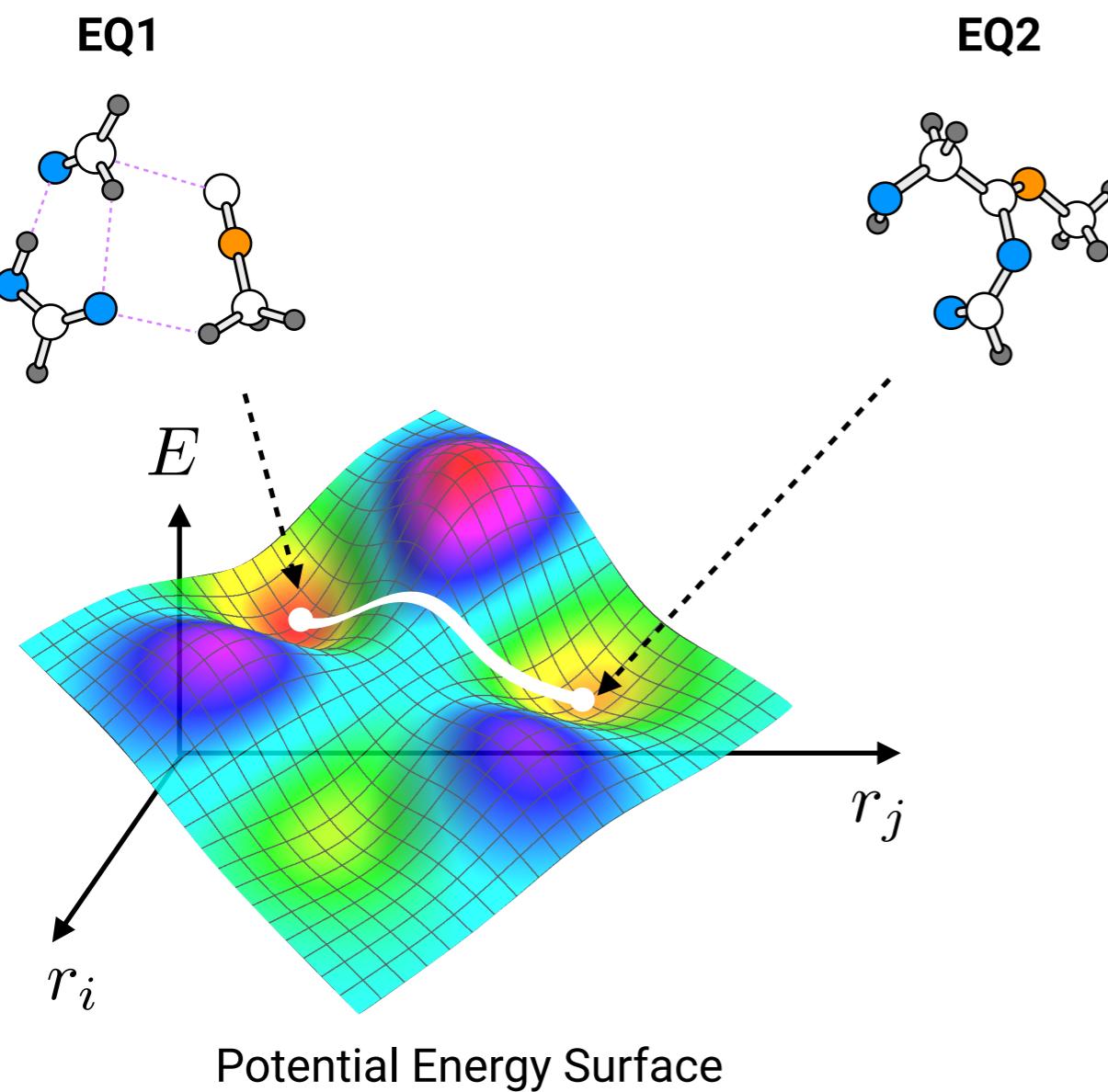
Chemical Reaction



Chemical Reaction Design and Discovery



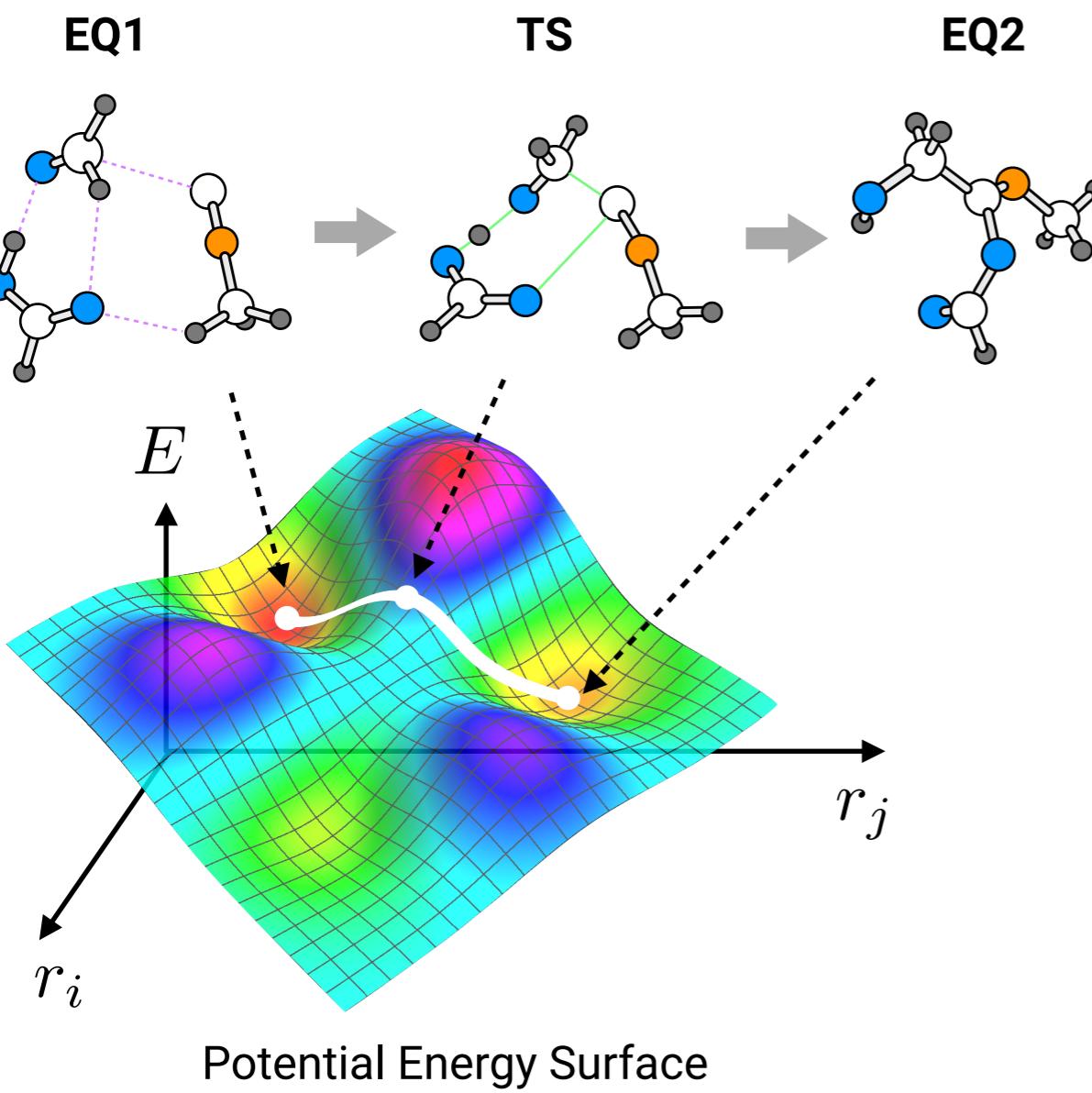
Chemical Reaction



Chemical Reaction Design and Discovery



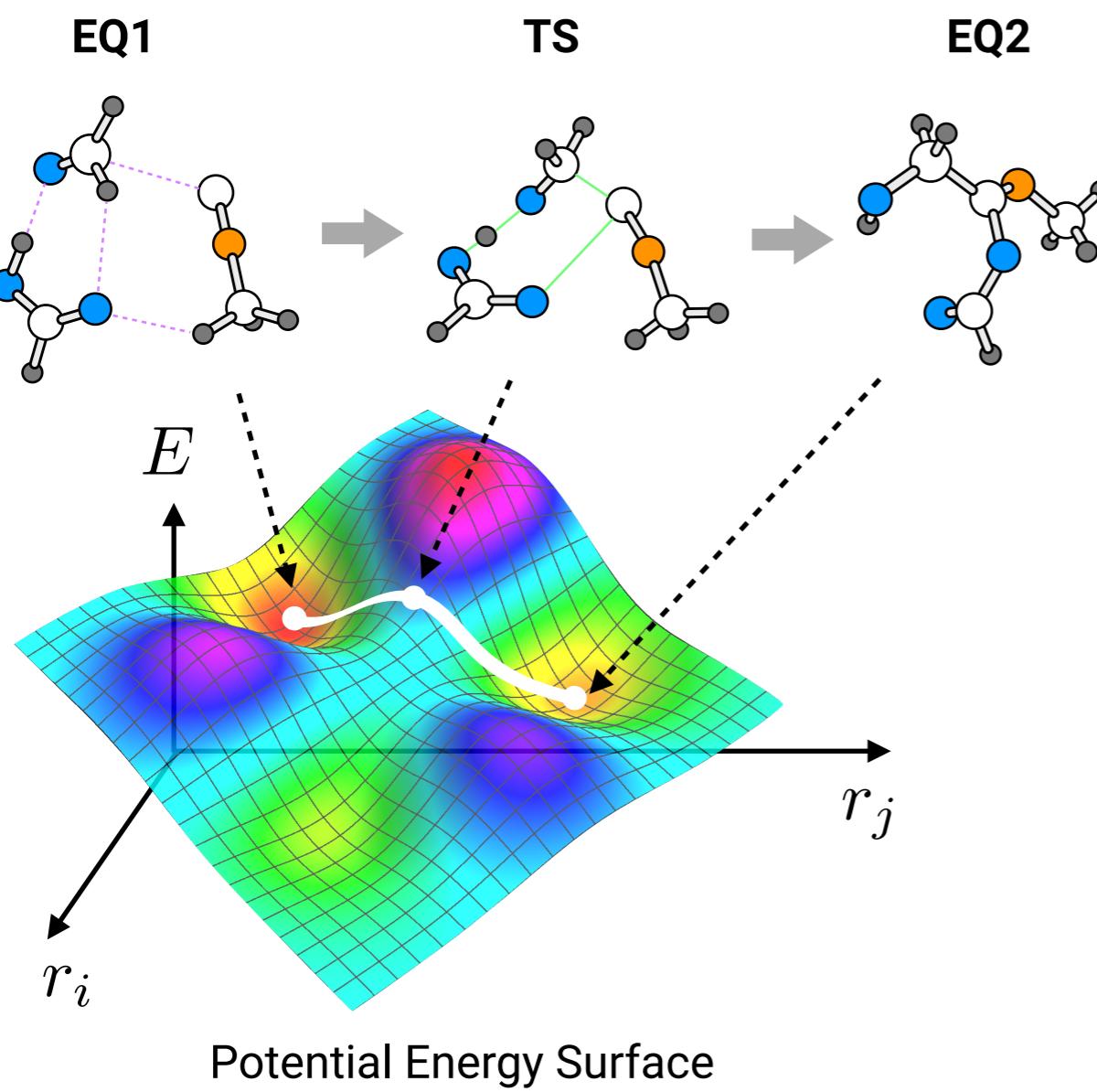
Chemical Reaction



Chemical Reaction Design and Discovery

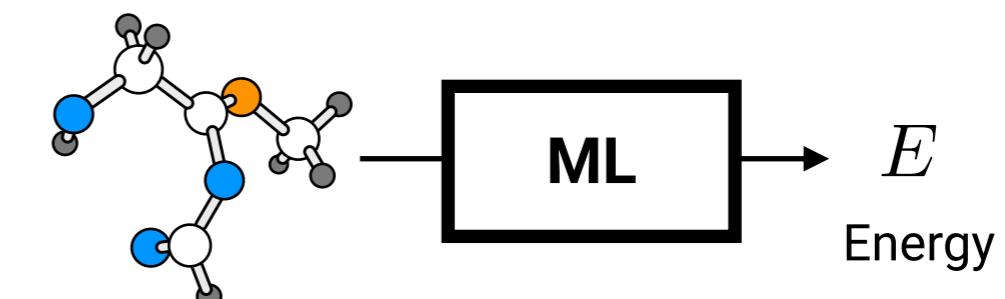


Chemical Reaction

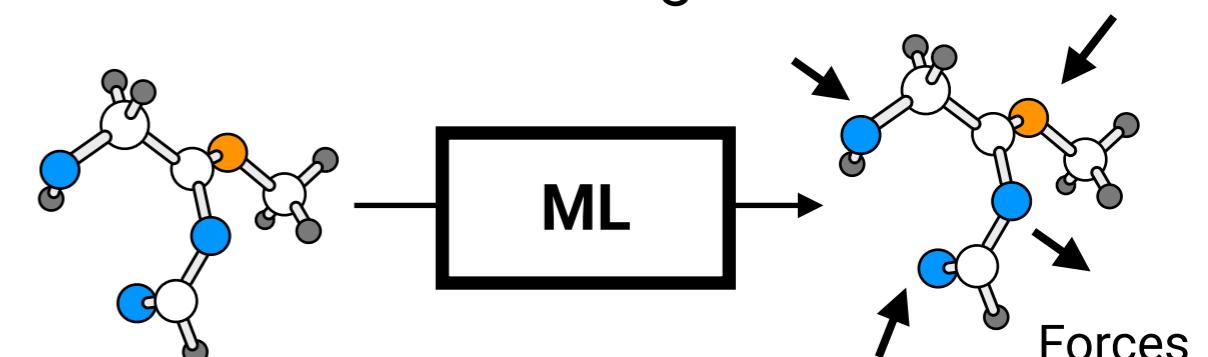


- Fill any gap between theory and experiments (reality) by data?

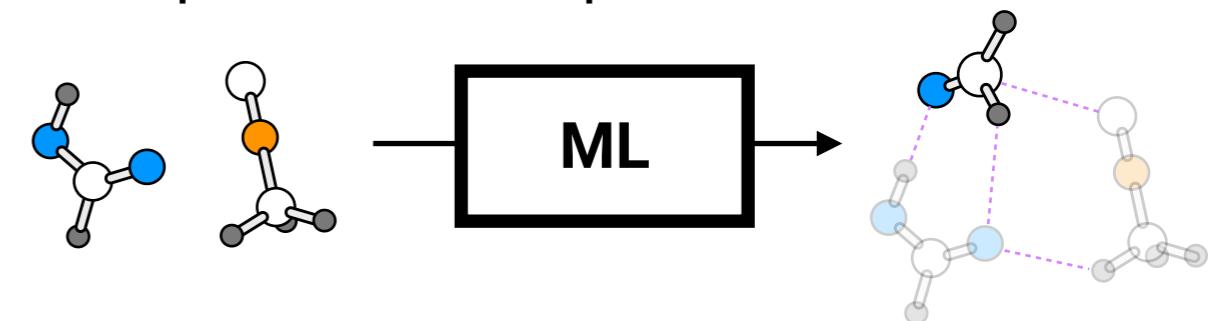
- Acceleration by ML potential?



- Artificial force learning?



- Scope/network expansion?



Machine Learning and Machine Discovery



An exciting “real-world” test bench for ML researchers:

- **Machine Learning**: many fascinating technical topics of my long-standing interests on “ML with combinatorial structures” (such as GNNs)
- **Machine Discovery**: many long-standing important open problems towards “AI for automating discovery”

