Tutorial: Data Mining for Drug Discovery and Development

Cao Xiao, Jimeng Sun

KDD' 19

Agenda

- Motivation
- Data
- Tasks
- Future Directions

Agenda

- **Motivation**
- Data
- Tasks
- Future Directions

AI Assisted Medicine

Success in AI based diagnosis

Company	FDA Approval	Indication
Apple	September 2018	Atrial fibrillation detection
Aidoc	August 2018	CT brain bleed diagnosis
iCAD	August 2018	Breast density via mammography
Zebra Medical	July 2018	Coronary calcium scoring
Bay Labs	June 2018	Echocardiogram EF determination
Neural Analytics	May 2018	Device for paramedic stroke diagnosis
IDx	April 2018	Diabetic retinopathy diagnosis
Icometrix	April 2018	MRI brain interpretation
Imagen	March 2018	X-ray wrist fracture diagnosis
Viz.ai	February 2018	CT stroke diagnosis
Arterys	February 2018	Liver and lung cancer (MRI, CT) diagnosis
MaxQ-AI	January 2018	CT brain bleed diagnosis
Alivecor	November 2017	Atrial fibrillation detection via Apple Watch
Arterys	January 2017	MRI heart interpretation

AI based drug discovery and development

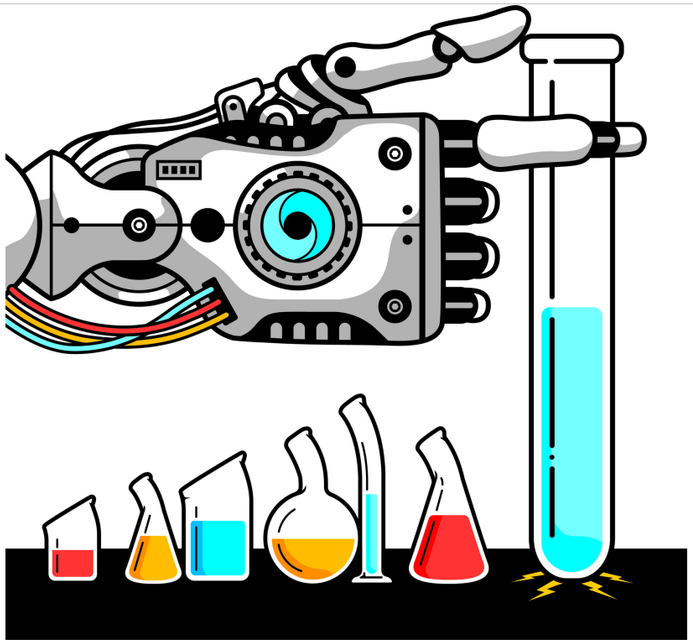
AI Accelerates Drug Discovery and Development

MERCK
Be well

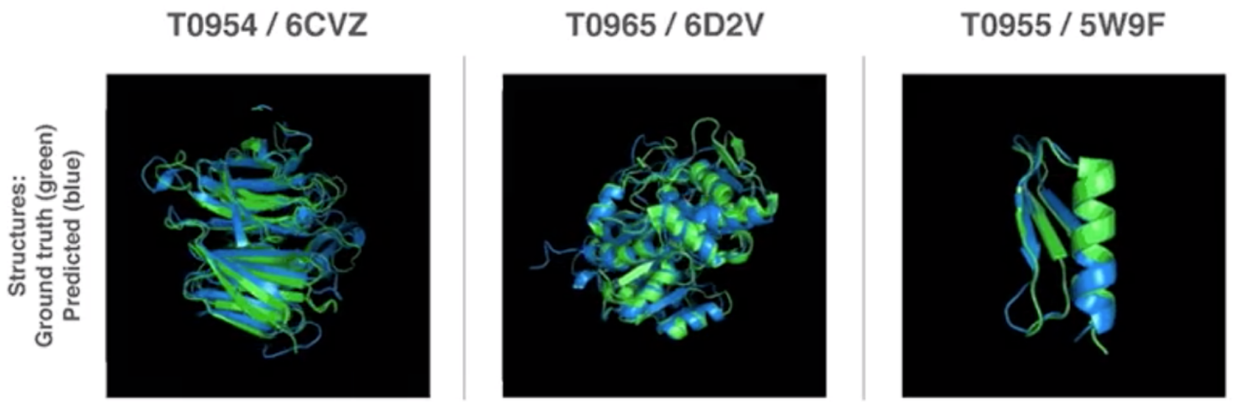
Merck Molecular Activity Challenge

Help develop safe and effective medicines by predicting molecular activity.

$40,000 · 236 teams · 7 years ago



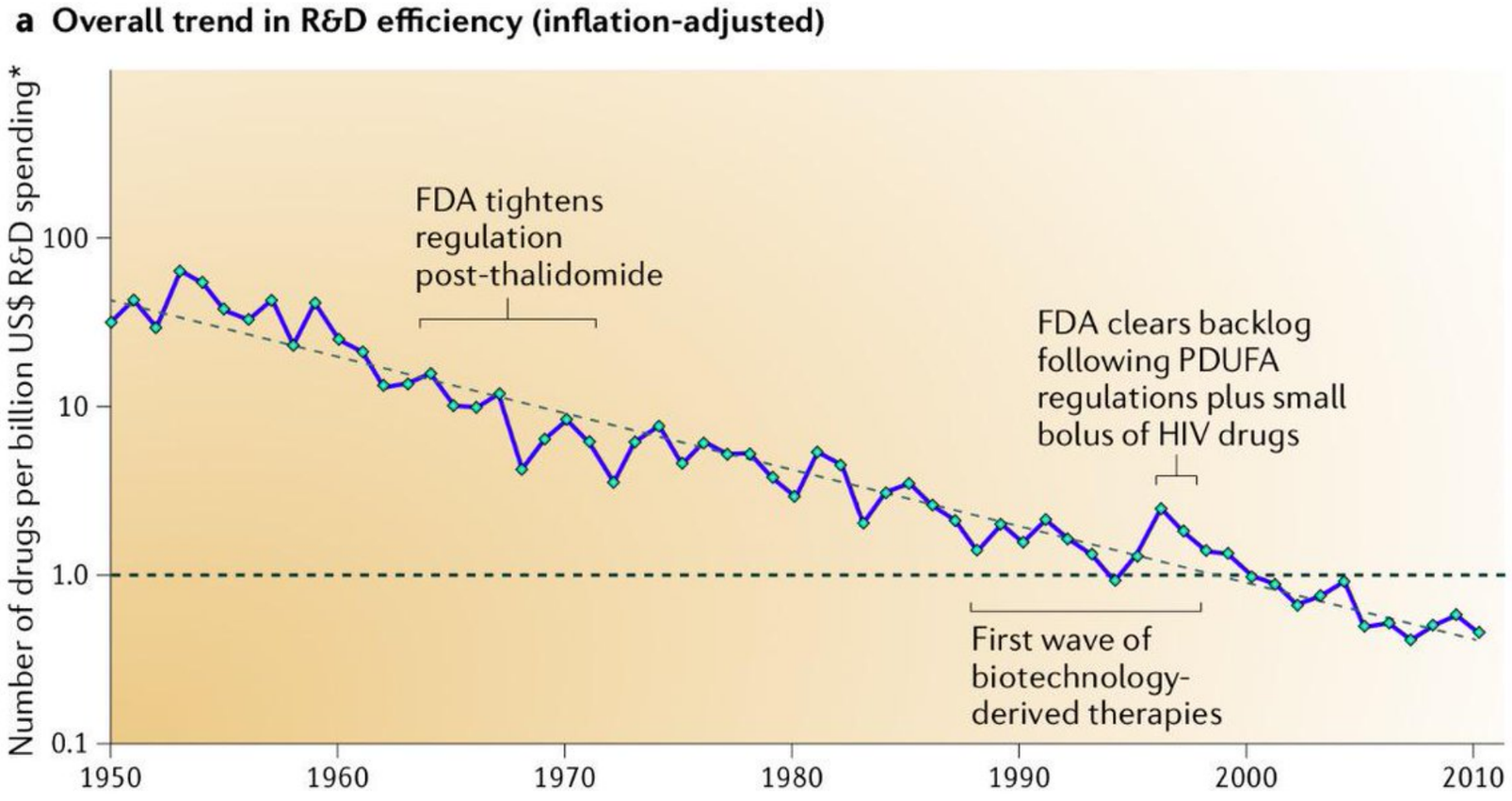
DeepMind's AI will accelerate drug discovery by predicting how proteins fold



https://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html

De novo structure prediction with deep-learning based scoring R.Evans, et al. In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts) 1-4 December 2018.

Eroom's Law in Pharmaceutical R&D

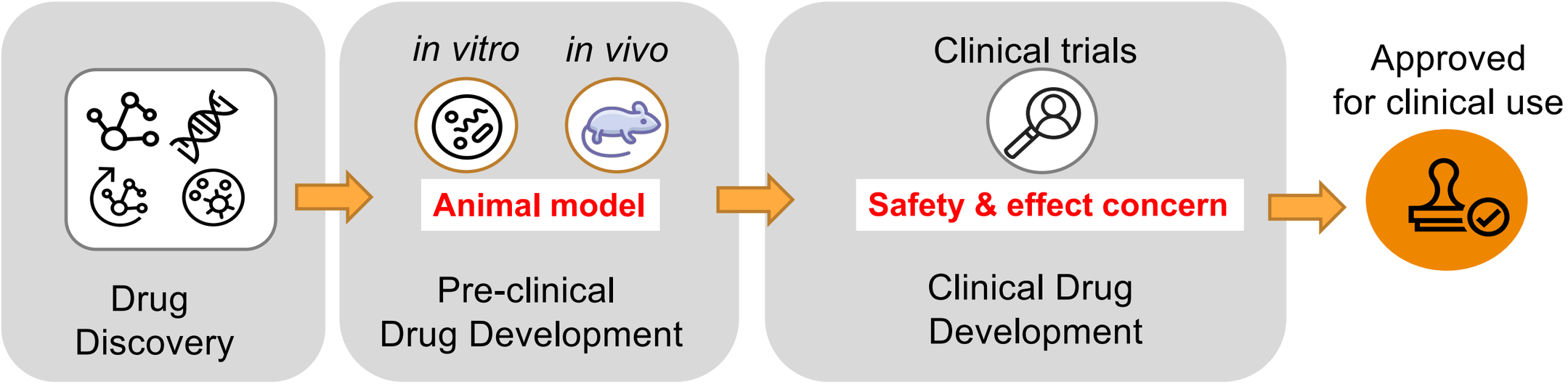


Causes of the Decline

1) Hard to improve over existing drugs
2) cautious regulator
3) "throw money at it" tendency
4) basic science-brute force

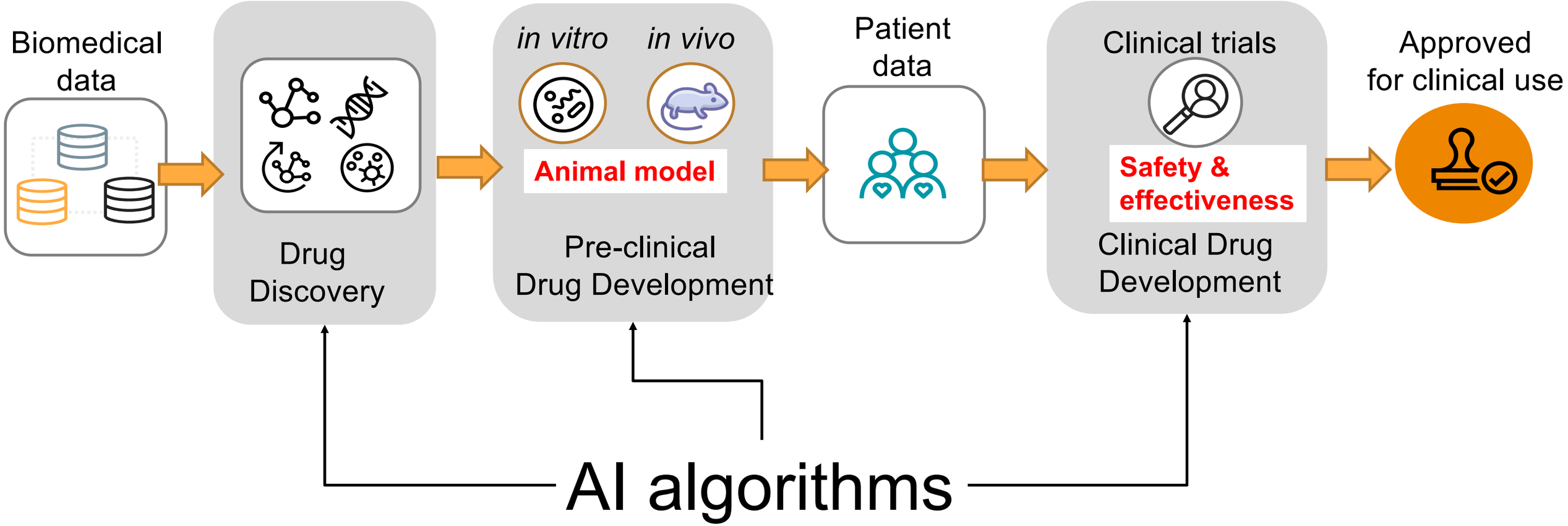
Scannel et al., Diagnosing the decline in pharmaceutical R&D efficiency, Nature Reviews Drug Discovery, 2012

Traditional Drug Discovery & Development Process

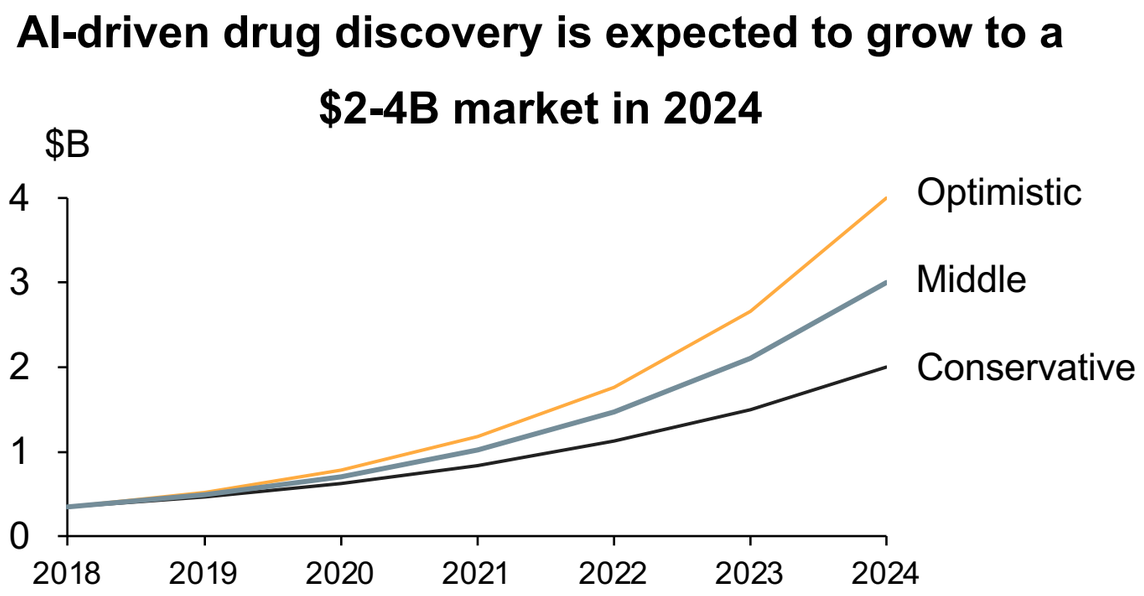


	Drug discovery	Pre-clinical	Phase 1	Phase 2	Phase 3
Time spent	4-5 years	1-2 years	1-2 years	1-2 years	2-3 years
$ spent	$550M	$125M	$225M	$250M	$250M
Output	5,000 - 10,000 compounds	10-20 candidates	5-10 candidates	2-5 candidates	1-2 candidates

AI/ML to the Rescue: Why and How?



Big pharmas show significant interests in AI



Novartis: A2A pharmaceuticals, Biovista, Watson

Merck: Synthace, Cyclica, Atomwise, Numerate, Iktos

Roche: Flatiron Health, Genialis, Exscentia, Owkin, Synapse, GNS

Sanofi: Researchably, Benevolent AI, Exscentia, Berg Health

GSK: Exscentia, Cloud Pharmaceruticals, Insilico Medicine

Biopharmatrend Report, Biopharmatrend Charts, GMInsights, Marketwatch,

Many AI start-ups in drug discovery

Competitor	Raised capital	Year funded	Employees	Approach
BenevolentAI	**$202M ($2B val)**	**2013**	**51-100**	Drug discovery, clinical trial simulation, biomarker ID, mechanism of disease with **knowledge graphs**
flatiron Roche	**($1.9B val)**	**2012**	**251-500**	External control arms (EHR), analysis through **linked EMR and genomic data**
healx	**$61.9M**	**2014**	**11-50**	Drug repositioning with **knowledge graphs**
GNS HEALTHCARE	**$54.3M**	**2000**	**101-250**	Drug discovery, clinical trial simulation, biomarker ID, mechanism of disease with **knowledge graphs**
Atomwise	**$51.3M**	**2012**	**11-50**	Drug discovery with **chemoinformatics** and CNNs
Exscientia DRIVEN BY KNOWLEDGE	**$43.7M**	**2012**	**11-50**	Drug discovery from AI to experimental, using **chemoinformatics** and phenotypic screening
OWKIN	**$18.1M**	**2016**	**11-50**	Drug discovery, clinical development optimization through **knowledge graphs**
twoXAR	**$14.3M**	**2014**	**11-50**	Drug discovery, clinical trial simulation with **linked EHR and biomolecular data**

Why drug discovery & development is interesting to data mining community

Source

- Compound databases
- Protein databases
- Disease knowledge
- Biochemical literature
- Clinical trial data

Representation

- Feature vectors
- Graphs
- Sequences
- Text

Challenges

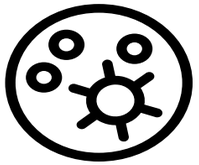
- High-dimensional
- Small sample size
- Lack of labels
- Complex interaction

Agenda

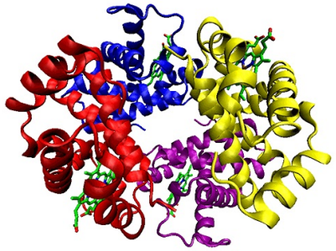
- Motivation
- **Data**
- Tasks
- Future Directions

Entities of Drug Discovery Modeling

Disease



Target



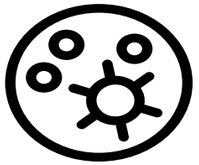
Identify a protein
involved in the disease

Molecule

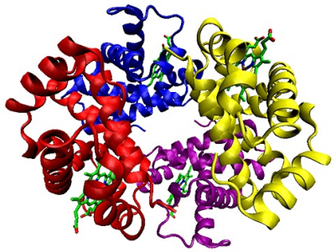
Find a molecule effective
against the target

Entities of Drug Discovery Modeling

Disease



Target

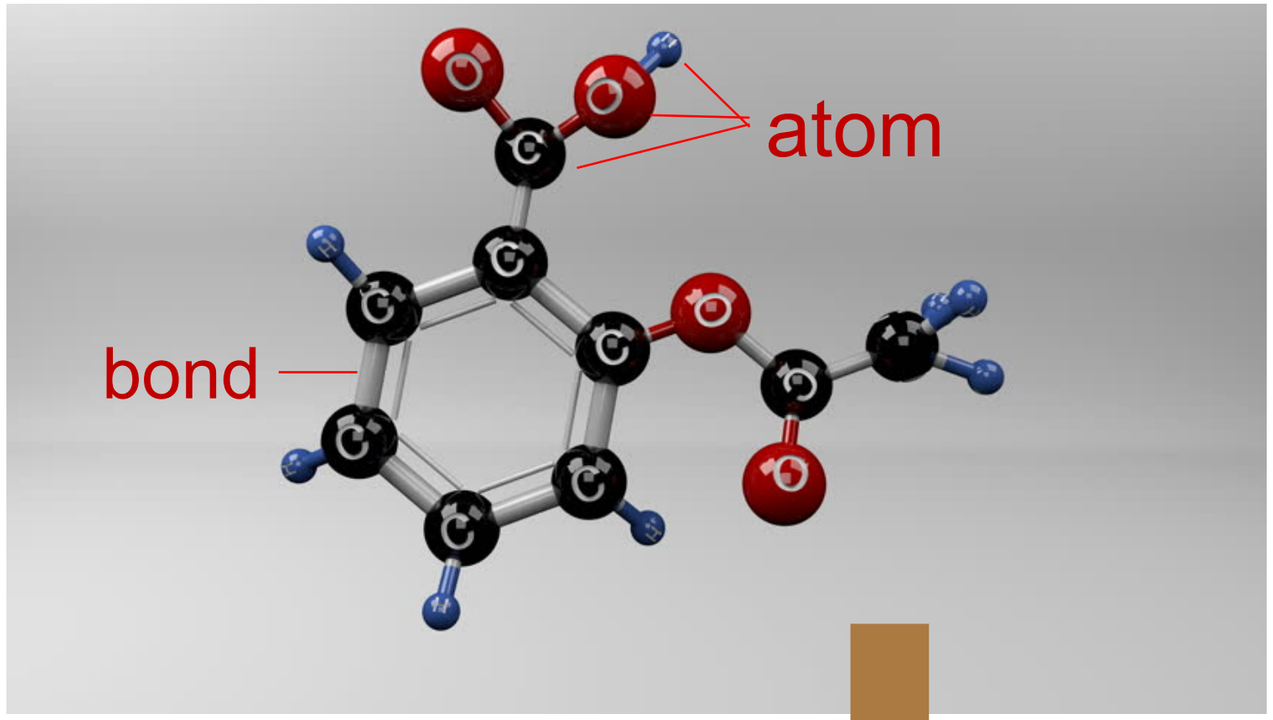


Identify a protein
involved in the disease

Molecule

Find a molecule effective
against the target

Data Encoding: Molecule Compounds (1D)



Aspirin molecule.

Formally, acetylsalicylic acid

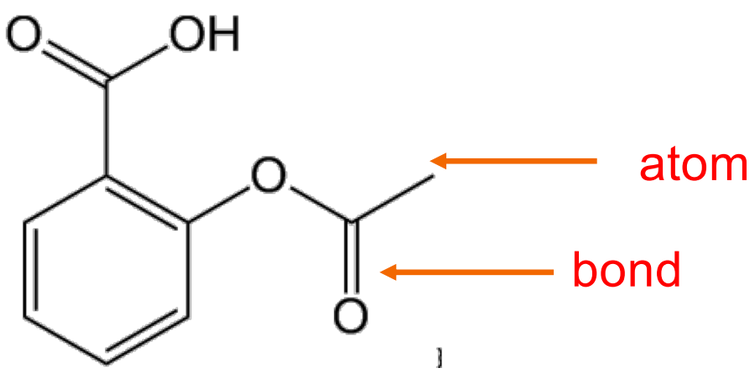
1-D descriptors

Weight, solubility, charge, number of rotatable bonds, atom types, topological polar surface area

Data Encoding: Molecule Compounds (2D)

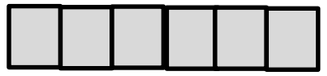
Aspirin molecule.

2D Graph Representation



fingerprinting

Circular fingerprinting

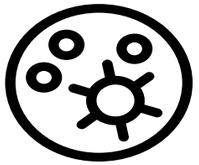


2-D Descriptor

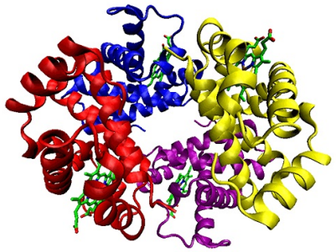
taking into account the graph of covalent and aromatic bonds, but not spatial coordinates.

Entities of Drug Discovery Modeling

Disease



Target



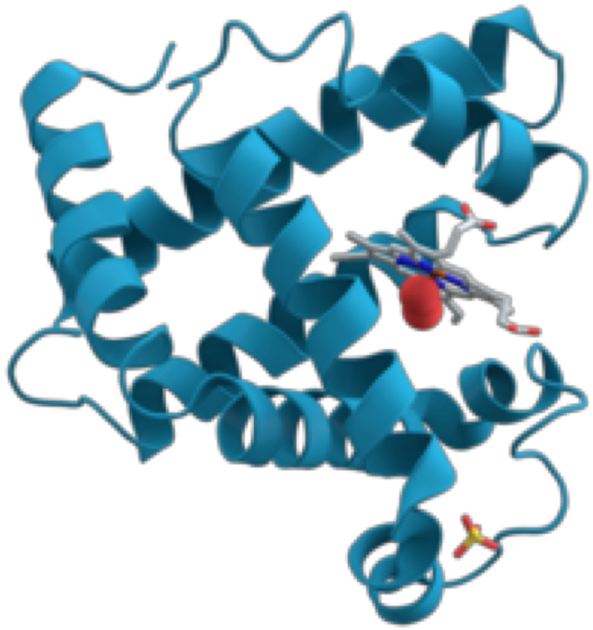
Identify a protein
involved in the disease

Molecule

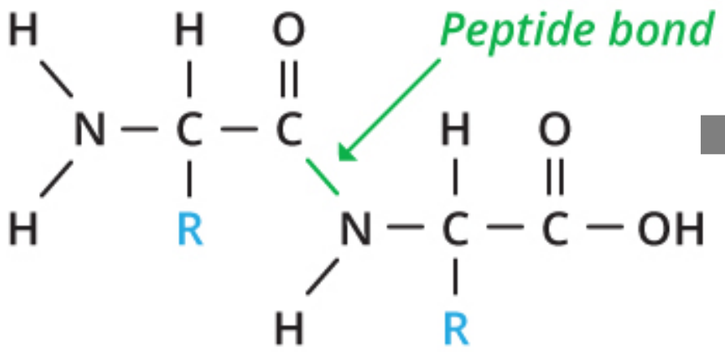
Find a molecule effective
against the target

Protein Targets

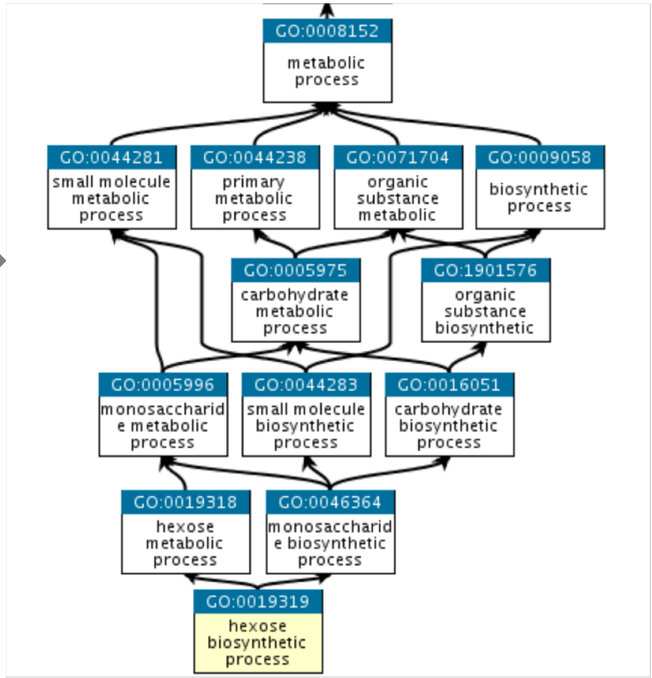
Biology
3-D structure



A Sequence of
23 Amino Acids

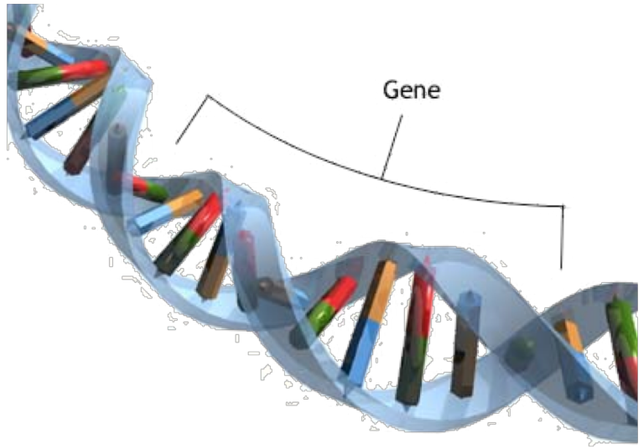


GO term
Protein Function

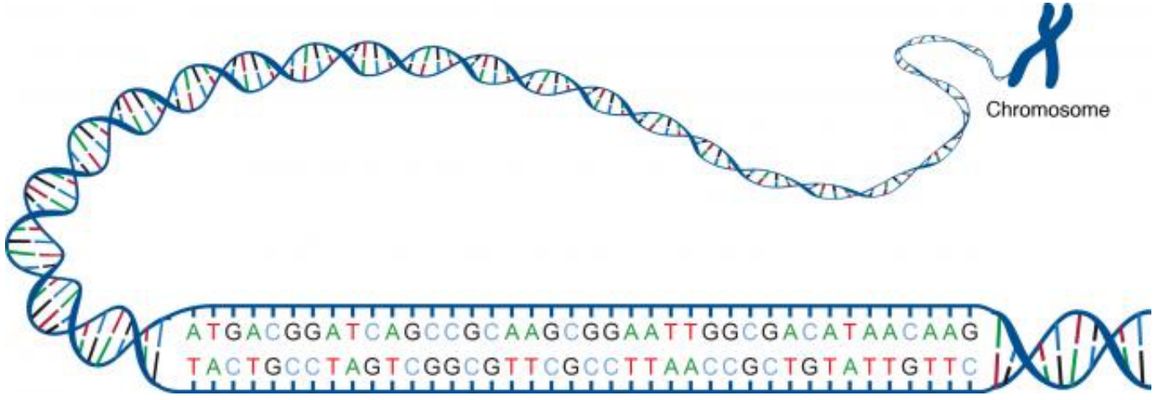


Gene Sequence

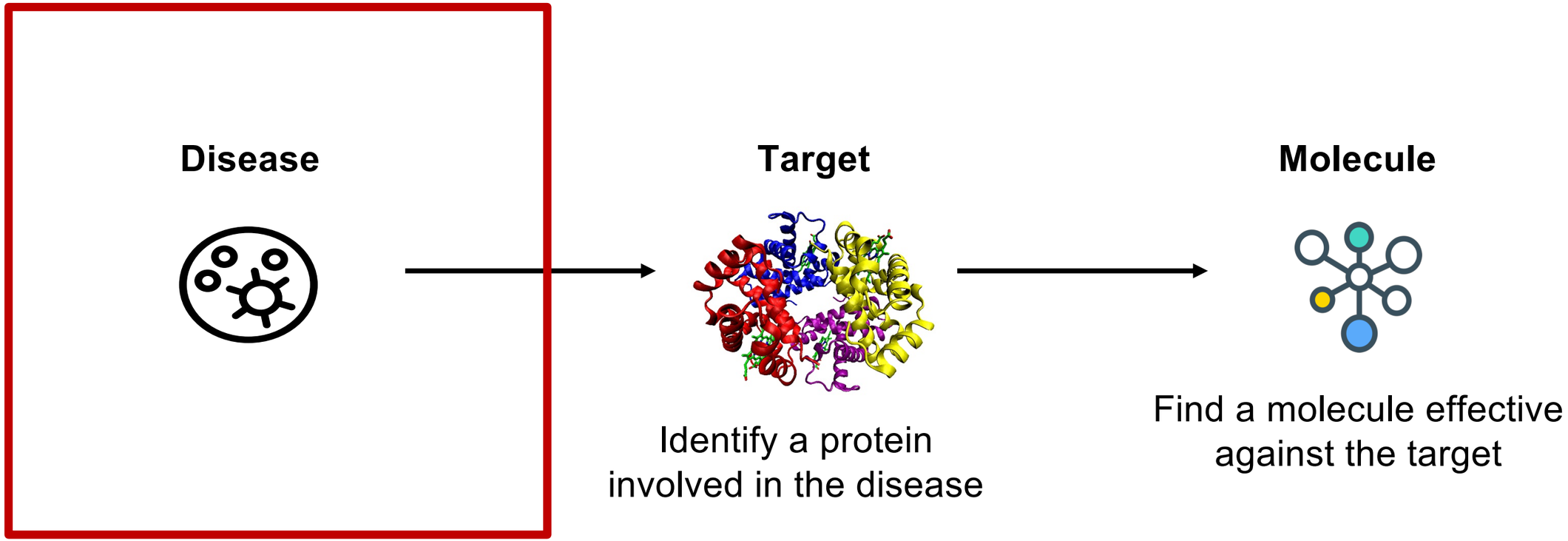
Double Helix DNA structure



A-T, C-G allele representation



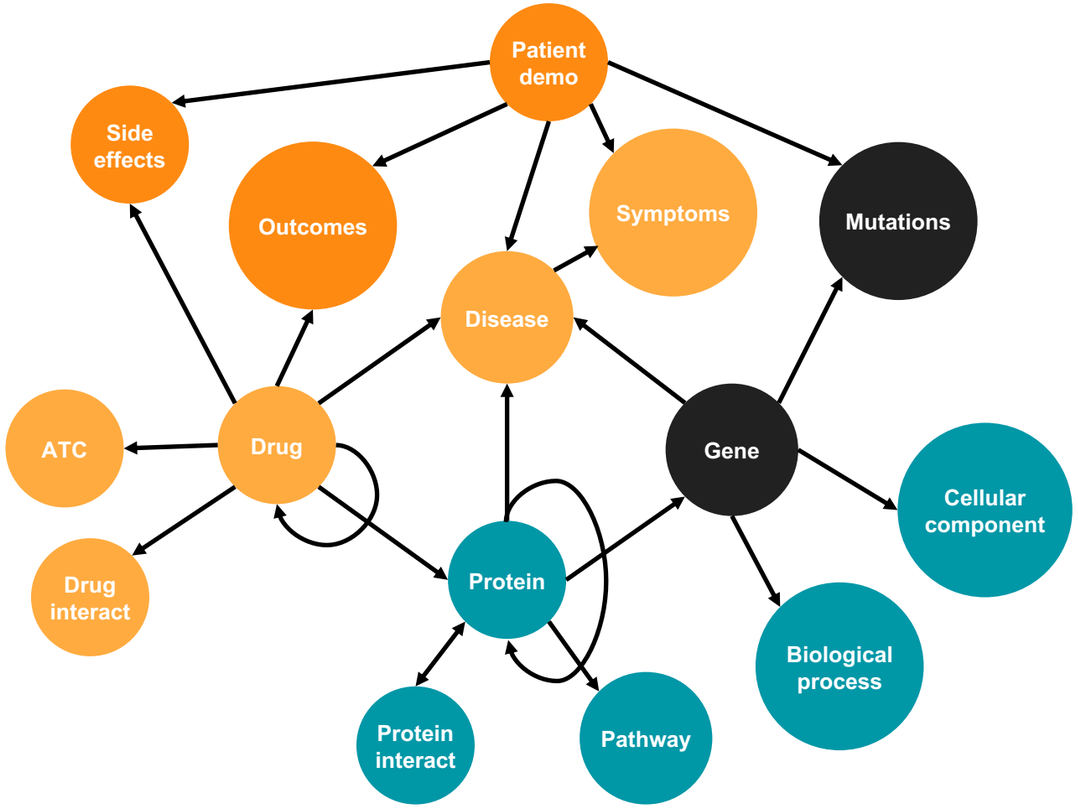
Entities of Drug Discovery Modeling

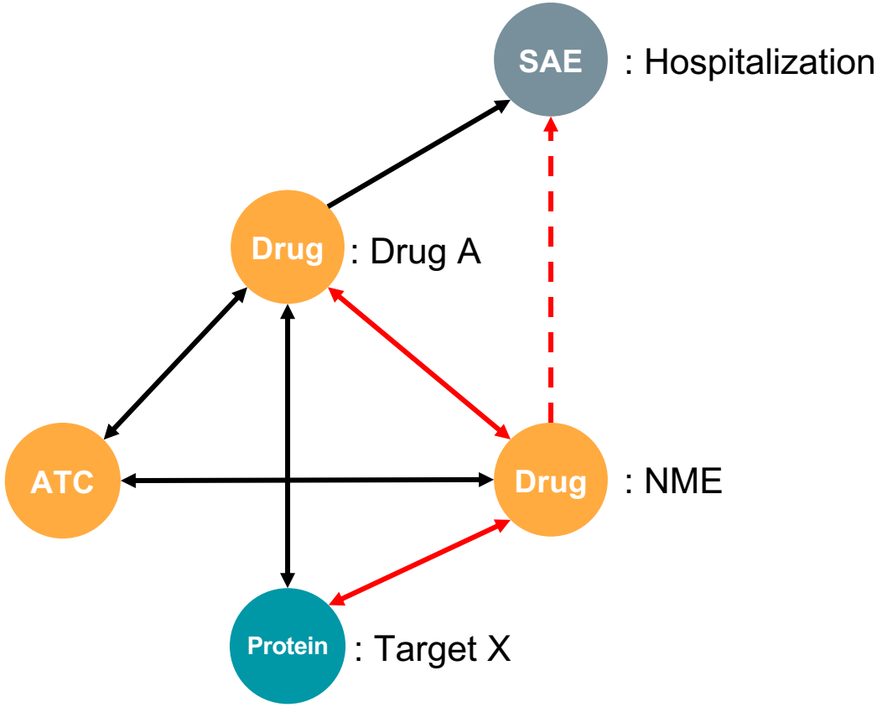


Disease

ICD code	ICD description
Acute rheumatic fever codes	
I00.	Rheumatic fever without mention of heart involvement
I01.0	Acute rheumatic pericarditis
I01.1	Acute rheumatic endocarditis
I01.2	Acute rheumatic myocarditis
I01.8	Other acute rheumatic heart disease
I02.0	Rheumatic chorea with heart involvement
I02.9	Rheumatic chorea without heart involvement

Biomedical Entities in the Knowledge Graph

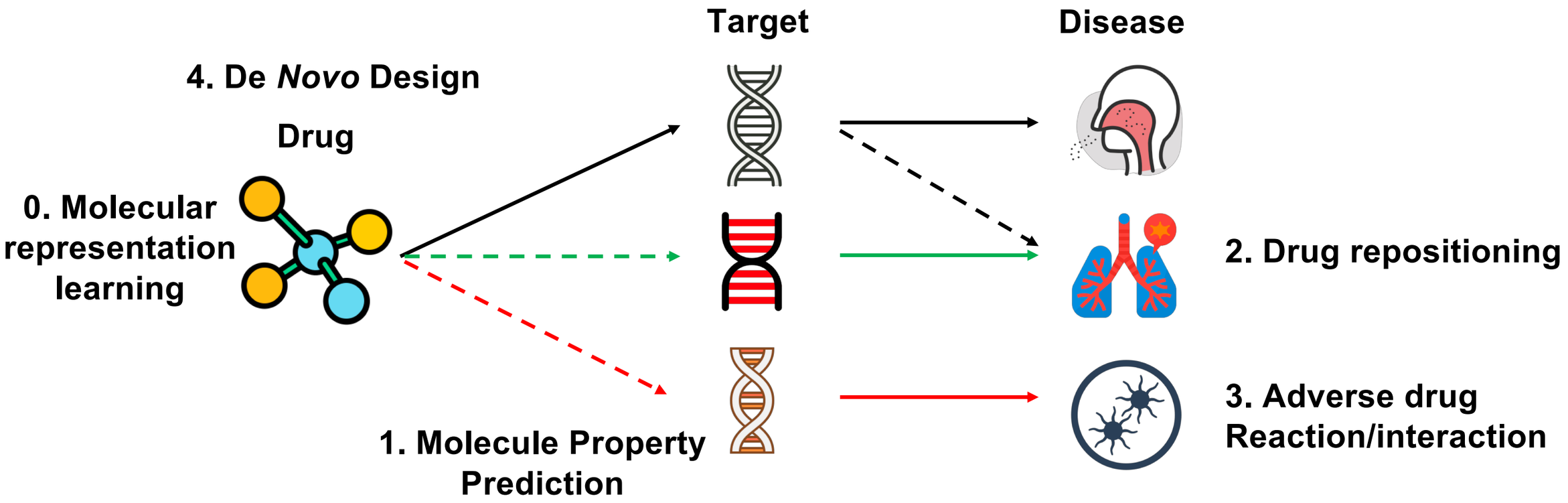




Agenda

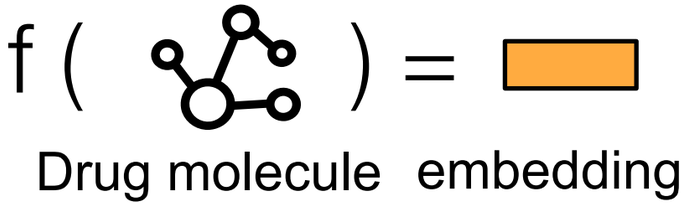
- Motivation
- Data
- **Tasks**
- Future Directions

Modeling Tasks Covered in This Talk

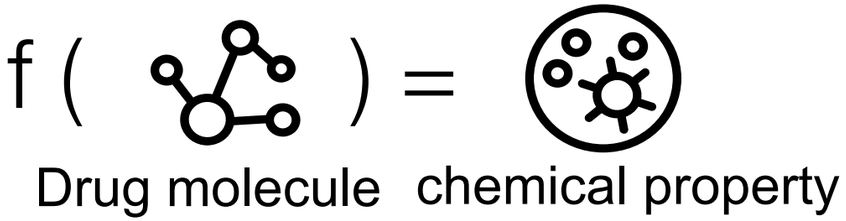


Input and Output of Modeling Tasks

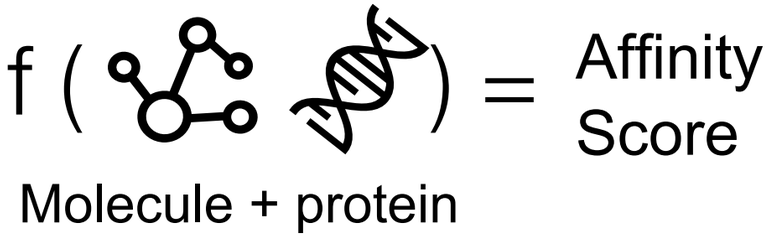
0. Molecular Representation Learning



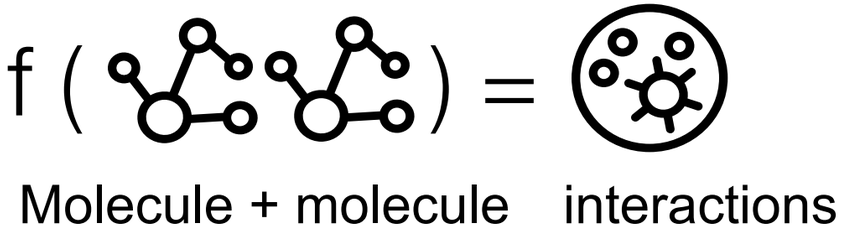
1. Molecule Property Prediction



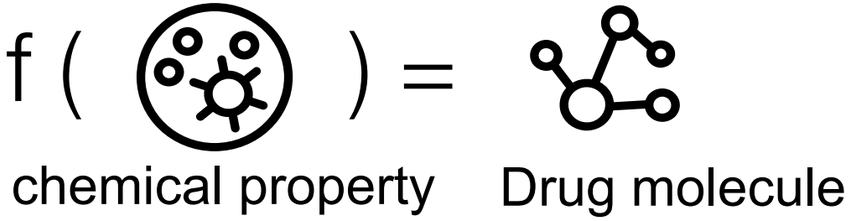
2. Drug repositioning



3. Adverse drug Reaction/interaction

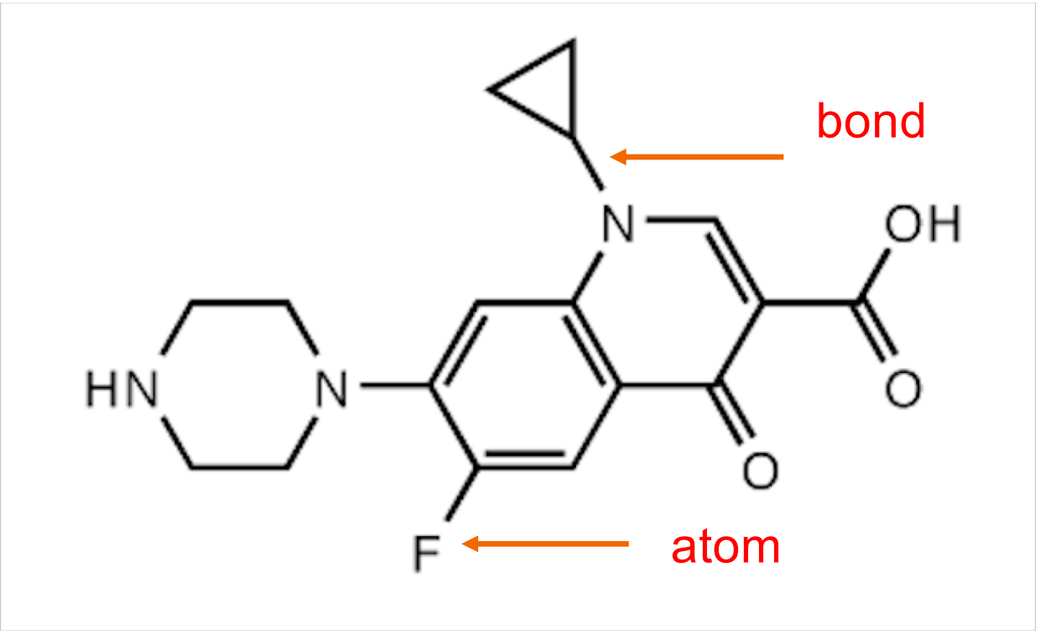


4. De *Novo* Design



0.Molecular Representation Learning

Molecular Graph



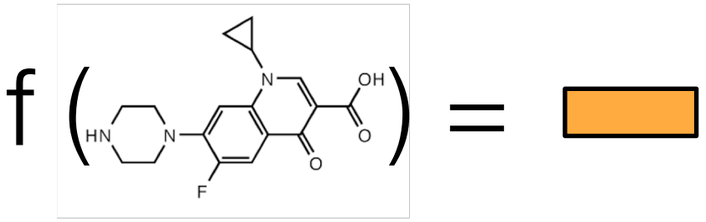
Molecular Graph

A molecule compound is a distinct group of atoms held together by chemical bonds.

Molecules with similar descriptors have similar properties.

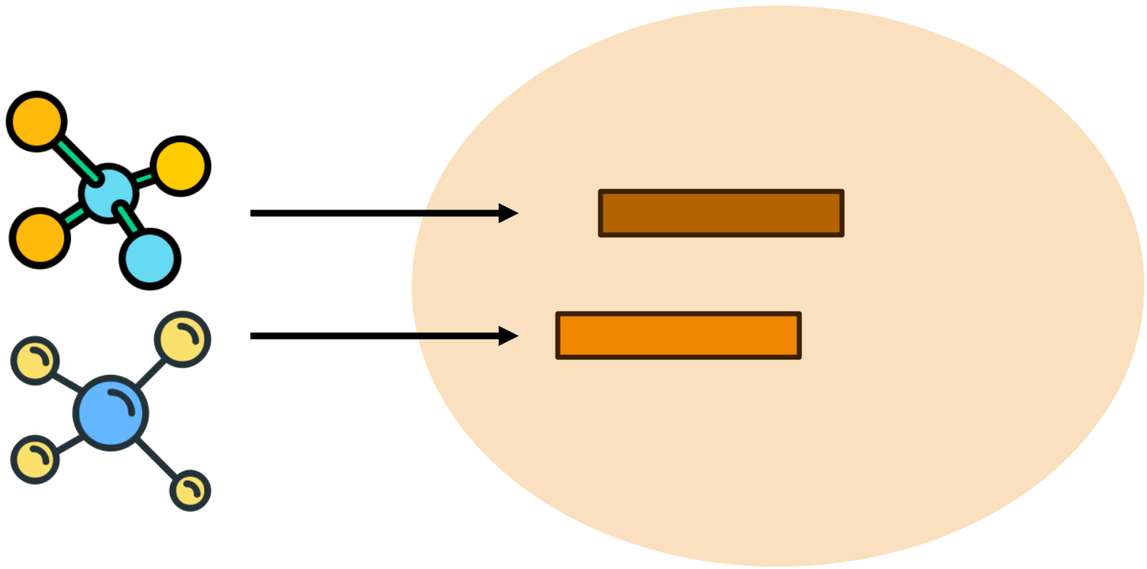
Molecular representation learning is a fundamental task for in silico modeling.

Molecular Graph Representation: Overview



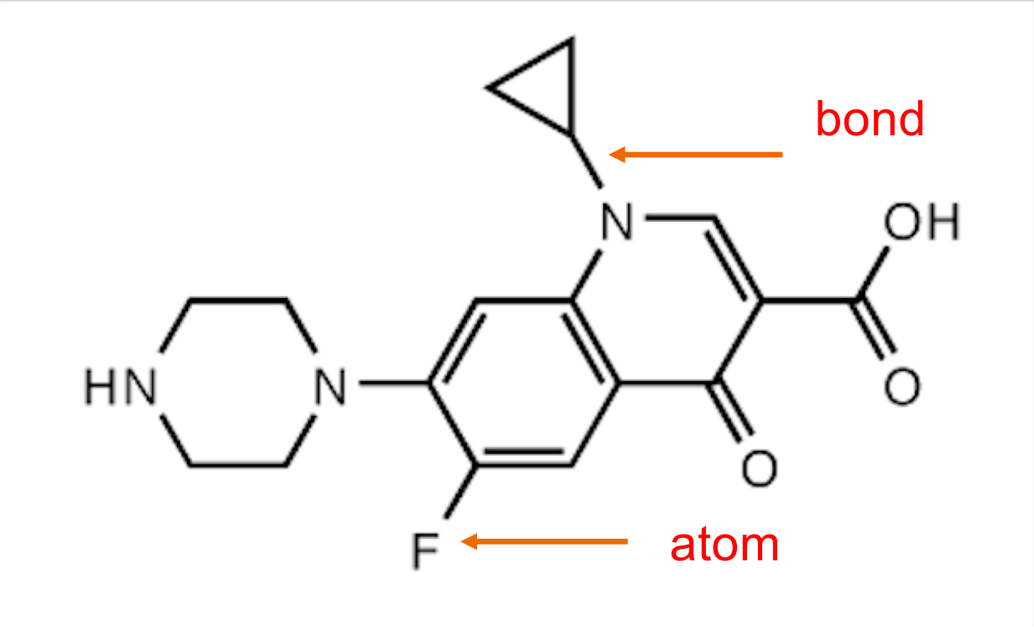
Drug molecular descriptors

Low dimensional embeddings



Intuition: Map raw drug molecular data to low dimensional embeddings such that similar molecules are embedded close together

Traditional Molecular Representation



1. Need to represent a **structure** by a characteristic vector of numbers (descriptors),
2. Should include **property**-relevant aspects
3. Atom arrangement in **space**

Traditional Molecular Representation (1D)

1-D Descriptor

experimental and calculated molecular properties that do not account for a molecule's bond structure: weight, solubility, charge, number of rotatable bonds, atom types, topological polar surface area

1. Need to represent a **structure** by a characteristic vector of numbers (descriptors),
2. Should include **property**-relevant aspects
3. Atom arrangement in **space**

Traditional Molecular Representation (2D)

2-D Descriptor

taking into account the graph of covalent and aromatic bonds, but not spatial coordinates.

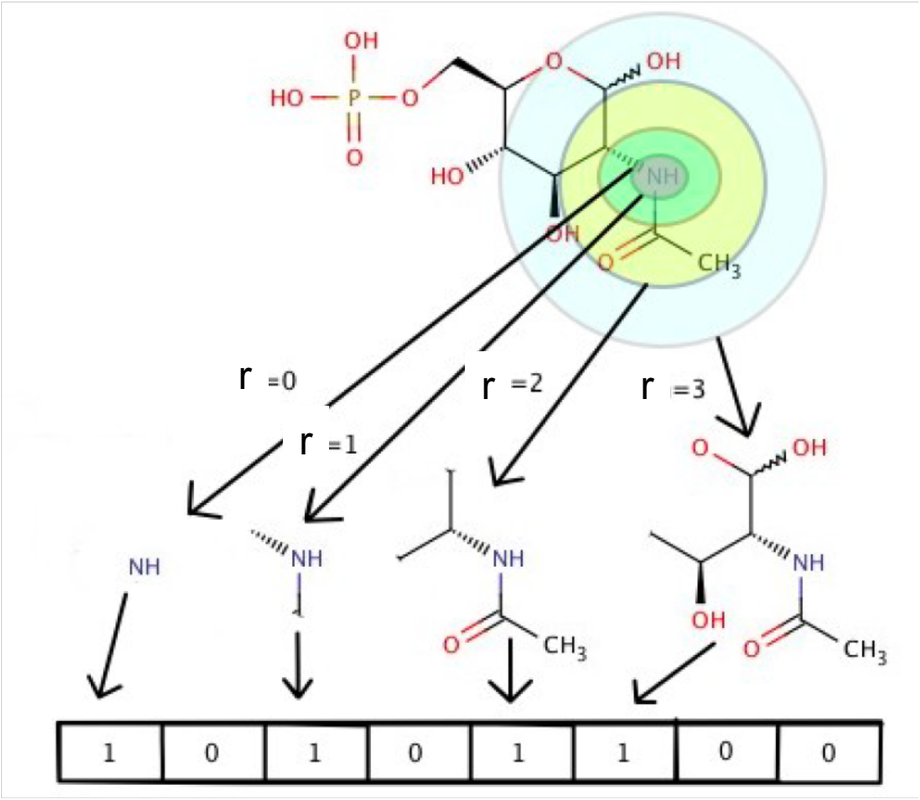
1. Need to represent a **structure** by a characteristic vector of numbers (descriptors),
2. Should include **property**-relevant aspects
3. Atom arrangement in **space**

Example of 2-D Representation (Circular Fingerprints)

Circular fingerprints

Input: molecule, radius R, fingerprint length S
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule **do**
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R **do** ▷ for each layer
  for each atom a in molecule **do**
    $\mathbf{r}_1 \ldots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow [\mathbf{r}_a, \mathbf{r}_1, \ldots, \mathbf{r}_N]$ ▷ concatenate
    $\mathbf{r}_a \leftarrow \text{hash}(\mathbf{v})$ ▷ hash function
    $i \leftarrow \text{mod}(r_a, S)$ ▷ convert to index
    $\mathbf{f}_i \leftarrow 1$ ▷ Write 1 at index
Return: binary vector $\mathbf{f}$

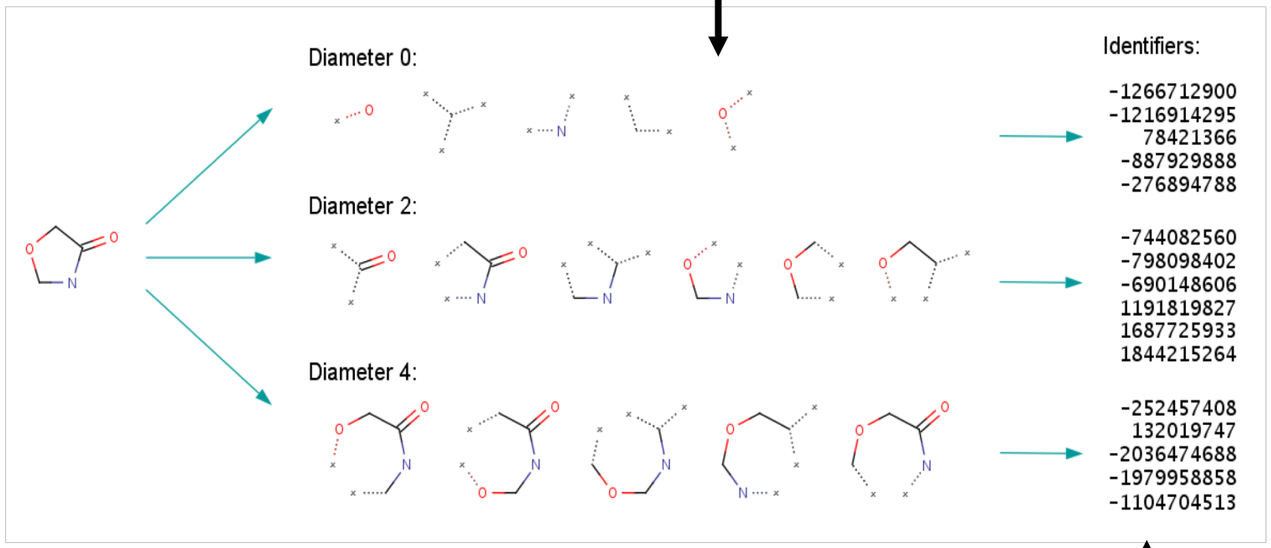
Extended Circular Fingerprint (ECFPx)



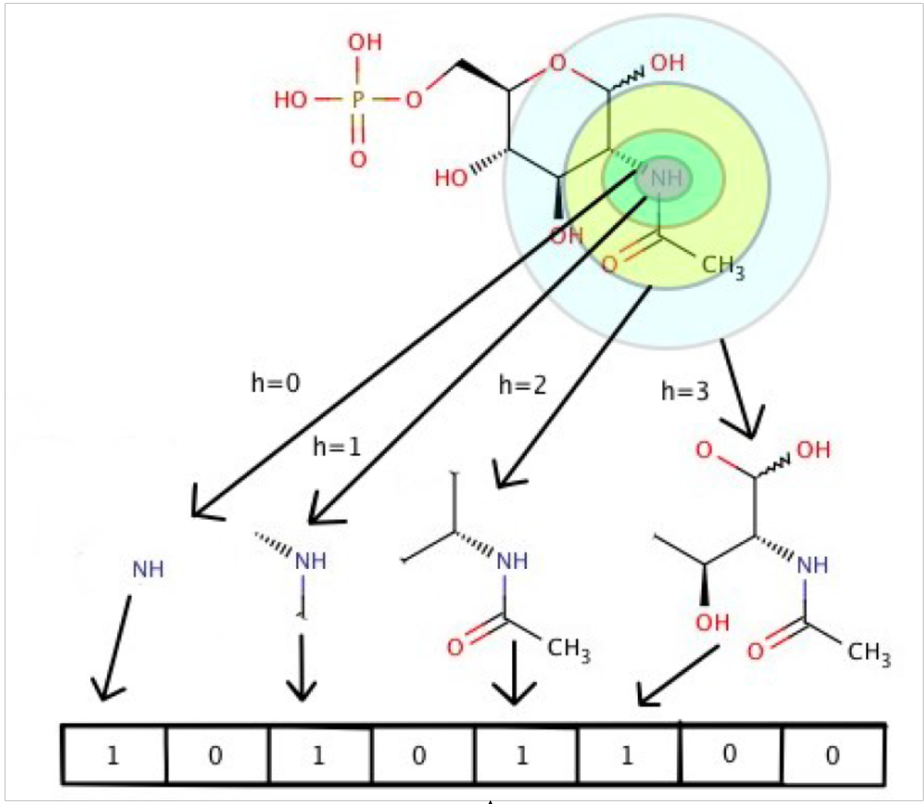
Example of 2-D Representation (Circular Fingerprints)

Extended Circular Fingerprint (ECFPx)

1. Search the partial structures around each atom recurrently



2. Assign an integer identifier to each partial structure



3. Convert to a binary vector using a hash function

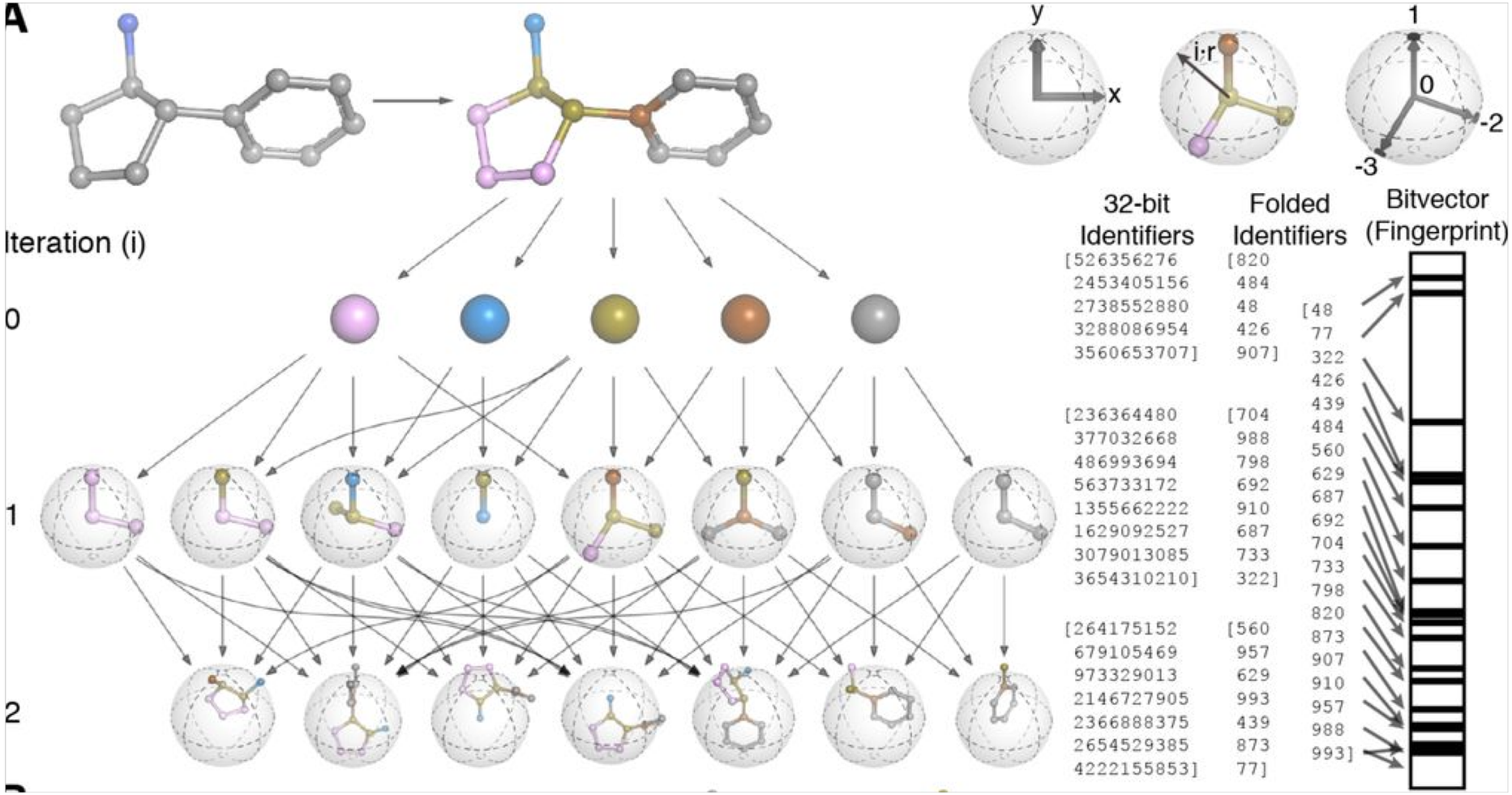
Knowledge Based Molecular Representation Learning (3D)

3-D Descriptor

Further considers spatial coordinates.

1. Need to represent a structure by a characteristic vector of numbers (descriptors), e.g., # N Atoms; # Aromatic Rings.
2. Should include property-relevant aspects, e.g., neighborhood-induced properties, and relative arrangement of atoms.
3. Atom arrangement in space,

Example of 3-D Representation (E3FP)

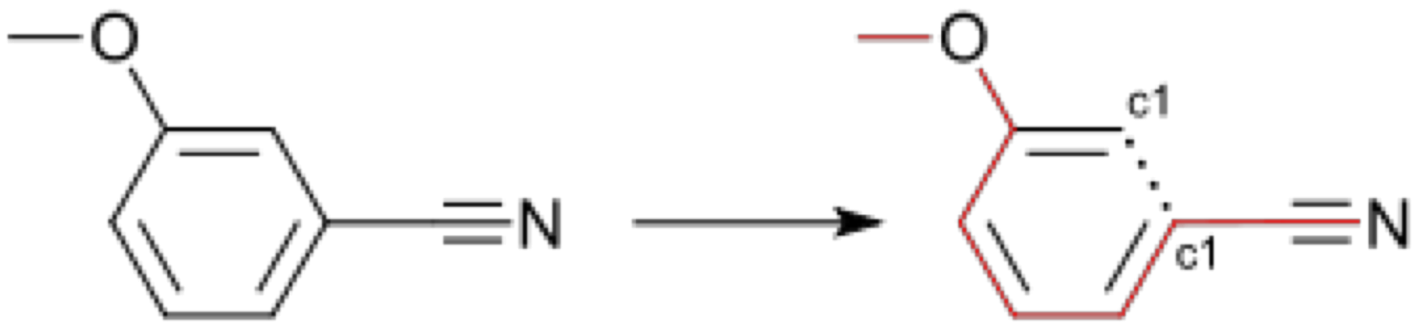


Extended Three-Dimensional Fingerprint (E3FP)

Simplified Molecular-Input Line-Entry System (SMILES)

Construction: Traverse the molecular graph in a depth-first manner following the atom with the smallest label at each branch point.

3-cyanoanisole



() are used to branches

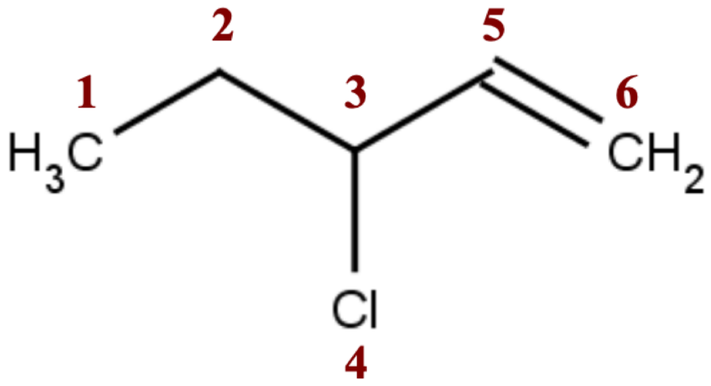
`COc(c1)cccc1C#N`

SMILES

Numbers are used to represent rings

Matrix Representation for Molecules (Bond Adjacency)

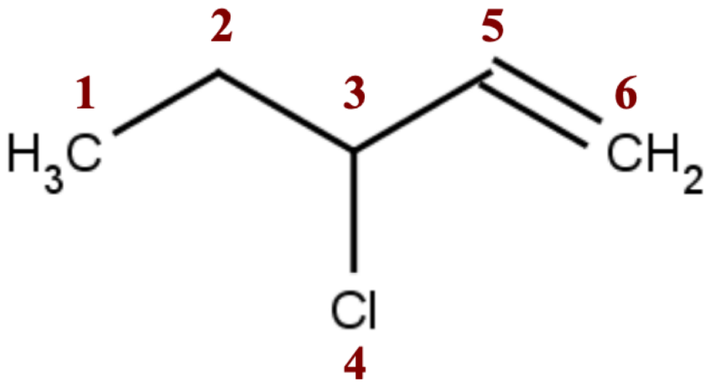
A molecular structure with n atoms may be represented by an n × n matrix (H atoms are often omitted).



	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	0	**2**
6	0	0	0	0	**2**	0

Matrix Representation for Molecules (Topological Distance)

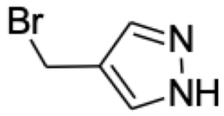
Distance matrix : encodes the distances between atoms.
Topological distance is defined as the number of bonds between atoms on the shortest possible path.



	1	2	3	4	5	6
1	0	1	2	3	3	4
2	1	0	1	2	2	3
3	2	1	0	1	1	2
4	3	2	1	0	2	3
5	3	2	1	2	0	1
6	4	3	2	3	1	0

Molecular Representation Learning using Deep Neural Networks

Graph



SMILES

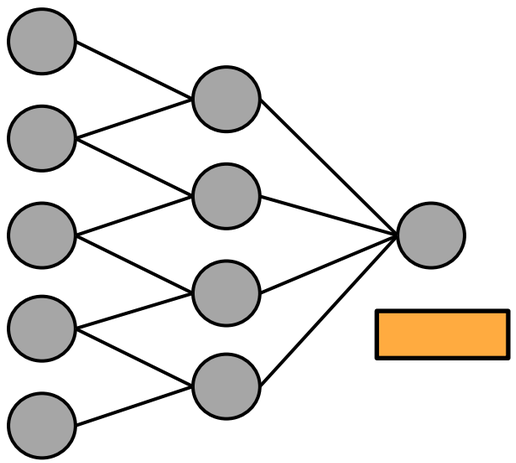
BrCc1c[nH]nc1

One-hot encoding

	Br	C	c	1	c	[	n	H	]	n	c	1
Br	1	0	0	0	0	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0
c	0	0	1	0	1	0	0	0	0	0	1	0
n	0	0	0	0	0	0	1	0	0	1	0	0
1	0	0	0	1	0	0	0	0	0	0	0	1
[	0	0	0	0	0	1	0	0	0	0	0	0
]	0	0	0	0	0	0	0	0	1	0	0	0

vectorize

f

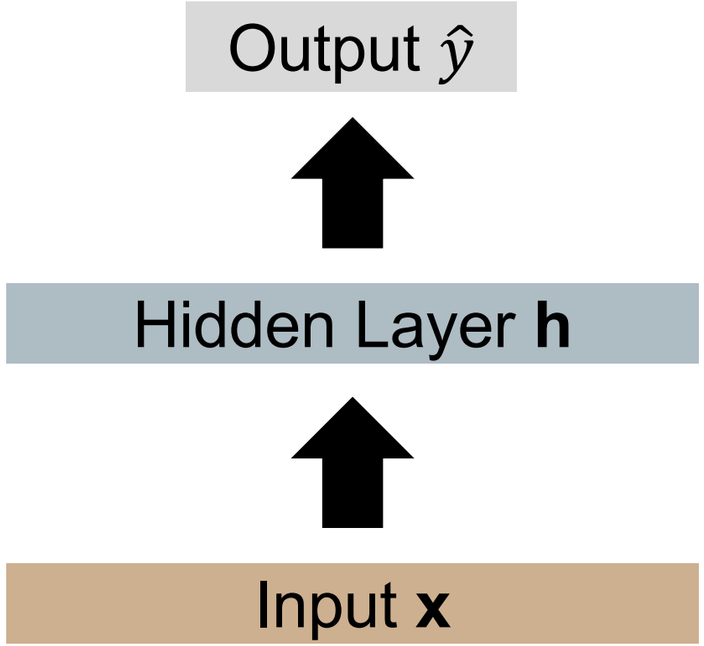


y

molecular property e.g., solubility, toxicity

Data Representation: deep neural networks

- Let's start with a simple Multi-layer Perceptron (MLP)
  - Binary classification: toxicity



$\hat{y} = \boldsymbol{\sigma}(\mathbf{w_o}^{\mathsf{T}}\mathbf{h})$ (generate an outcome 0.0~1.0)

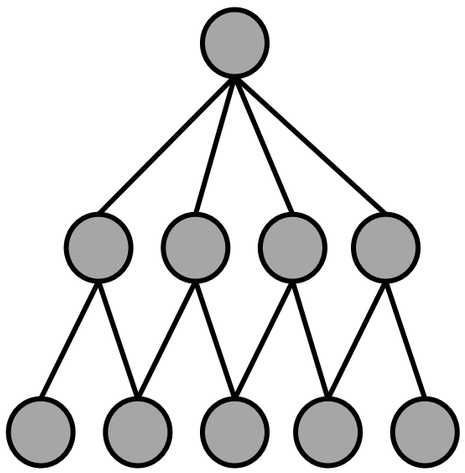
$\mathbf{h} = \boldsymbol{\sigma}(\mathbf{W_h}^{\mathsf{T}}\mathbf{x})$ (transform **x** for an easier prediction)

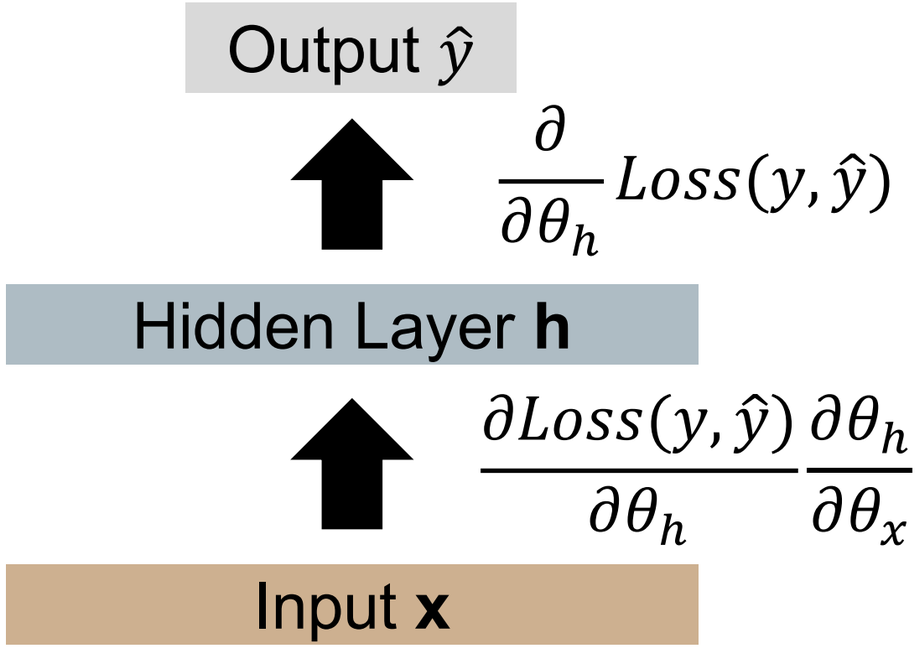
x (a vector representation of the input sample)

Data Representation (deep neural networks)

- Learning the model parameters
  - Backpropagation + Gradient descent

$$Loss(y, \hat{y}) = -(y \log \hat{y} + (1 - y) \log(1 - \hat{y}))$$



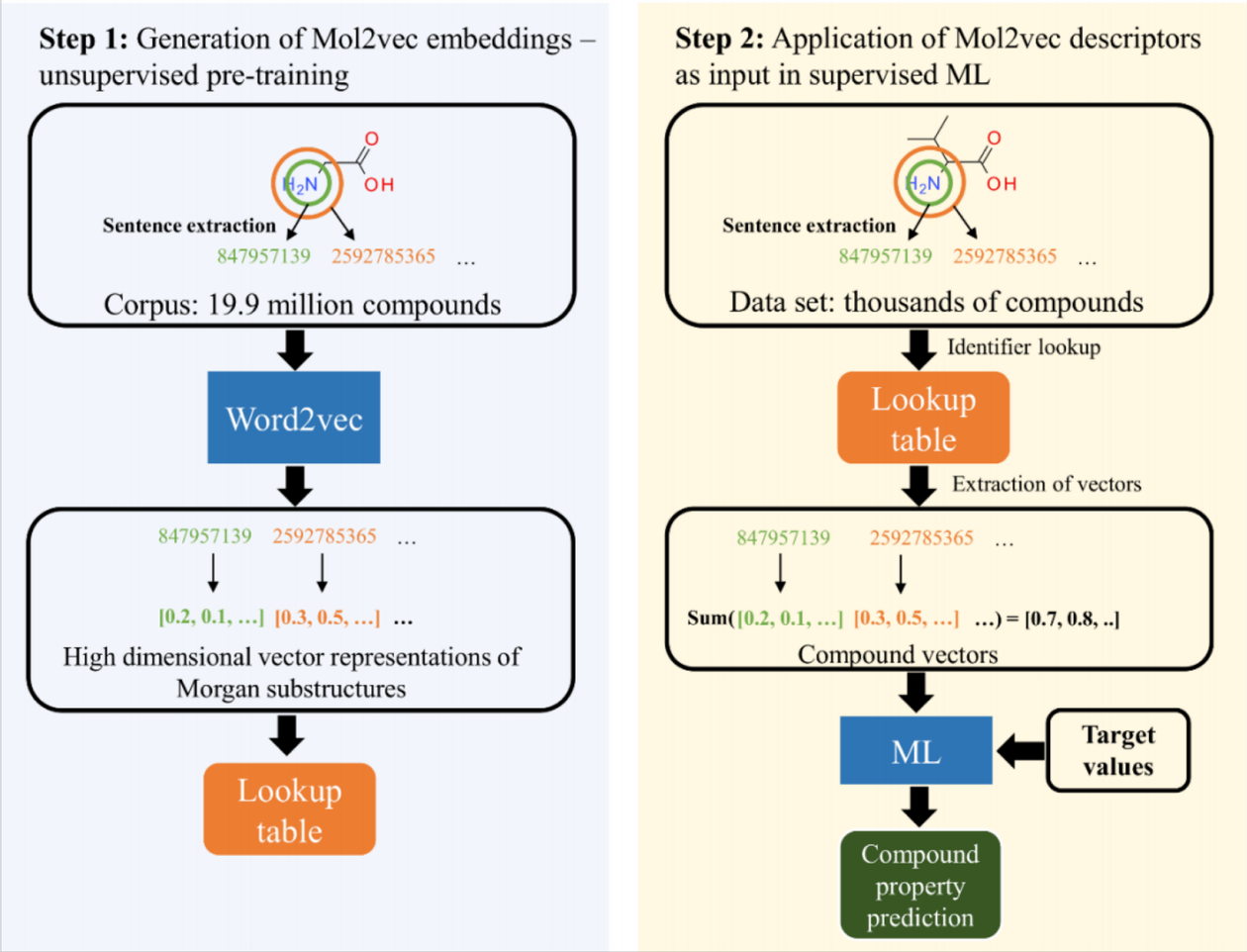


Mol2Vec (JCIM 2018)

learning molecular representation without labeled or with limited number of samples

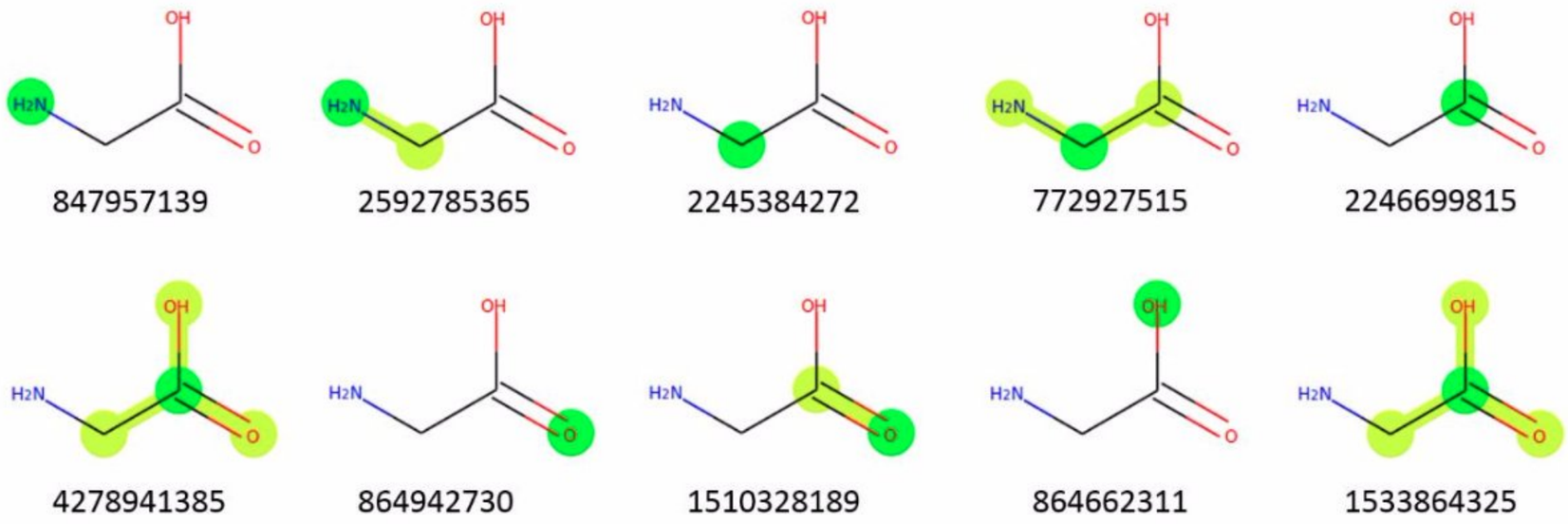
Sentence <-> molecules

words <-> substructure



Jaeger et al., Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. J Chem Inf Model 2018

Mol2Vec (JCIM 2018)



Method: Generate word2vec embedding on ECFP integer identifiers (words). As each identifier corresponds to a substructure, one molecule structure corresponds to a SMILE string (“sentence”)

Jaeger et al., Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. J Chem Inf Model 2018

Mol2Vec (JCIM 2018)

ESOL solubility data set25 : a regression task to predict aqueous solubility of 1144 compounds

Ames mutagenicity data set: a classification task to determine if is mutagenic of 6471 compounds

Tox21: classification task about 12 targets which were associated with human toxicity of 8192 compounds

Mol2Vec (JCIM 2018)

Table 1. Performance of Mol2vec and Other Models on Regression Predictions of the ESOL Data Set

ML features	ML method	R^2_{cv}	MSE	MAE	ref
descriptors	MLR	0.81 ± 0.01	0.82	0.69	28
molecular graph	CNN	0.82	–	–	40
molecular graph	CNN	–	–	0.52 ± 0.07	41
molecular graph	CNN	0.93	0.31 ± 0.03	0.40 ± 0.00	9
molecular graph	RNN	0.92 ± 0.01	0.35	0.43	42
Morgan FPs	GBM	0.66 ± 0.00	1.43 ± 0.00	0.88 ± 0.00	this work
Mol2vec	GBM	0.86 ± 0.00	0.62 ± 0.00	0.60 ± 0.00	this work

Table 2. Performance of Mol2vec and Other Methods on Classification Prediction of the Ames Data Set

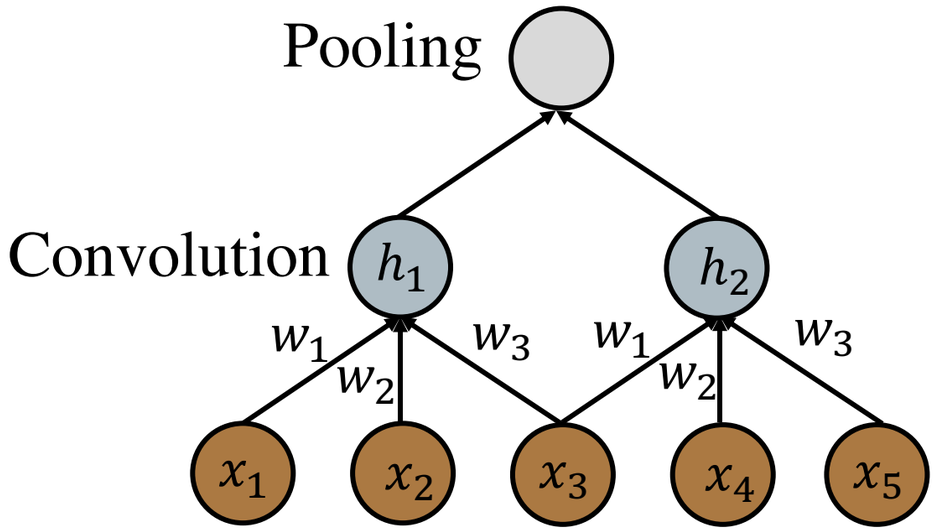
ML features	ML method	AUC	sensitivity	specificity	ref
descriptors	SVM	0.86 ± 0.01	–	–	29
descriptors and Morgan FPs	NBC	0.84 ± 0.01	0.74 ± 0.02	0.81 ± 0.01	43
Morgan FPs	RF	0.88 ± 0.00	0.82 ± 0.00	0.80 ± 0.01	this work
Mol2vec	RF	0.87 ± 0.00	0.80 ± 0.01	0.80 ± 0.01	this work

Table 3. Performance of Mol2vec and Other Methods on Classification Predictions of the Tox21 Data Set

ML features	ML method	AUC	sensitivity	specificity	ref
molecular graph	CNN	0.71 ± 0.13	–	–	9
molecular descriptors and FPs	SVM	0.71 ± 0.13	–	–	5
molecular descriptors and FPs	DNN	0.72 ± 0.13	–	–	5
Morgan FPs	RF	0.83 ± 0.05	0.28 ± 0.14	0.99 ± 0.01	this work
Mol2vec	RF	0.83 ± 0.05	0.20 ± 0.15	1.00 ± 0.01	this work

Jaeger et al., Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. J Chem Inf Model 2018

Data Representation (convolutional neural networks)



- Process data that has a known grid-like structure (e.g., images, waveforms).
- Utilize a specialized linear operation – convolution.
- Advantages: sparse interactions, parameter sharing, and translational invariance.

Neural Fingerprint (NIPS' 15)

Contribution

- provide an end-to-end learning framework:
  - to learn fingerprint with better predictive performance
  - the inputs are graphs with arbitrary size and shape
- Efficient computation
  - Fixed fingerprint must be large to encode all possible substructures
  - Neural fingerprint can be learned to encode relevant features for classification-> reduce the size
- Neural fingerprint is more interpretable-> meaningful

Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints, NIPS' 15

Neural Fingerprint (NIPS' 15)

Circular fingerprints

Input: molecule, radius R, fingerprint length S
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule **do**
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R **do** ▷ for each layer
  for each atom a in molecule **do**
    $\mathbf{r}_1 \ldots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow [\mathbf{r}_a, \mathbf{r}_1, \ldots, \mathbf{r}_N]$ ▷ concatenate
    $\mathbf{r}_a \leftarrow \text{hash}(\mathbf{v})$ ▷ hash function
    $i \leftarrow \text{mod}(r_a, S)$ ▷ convert to index
    $\mathbf{f}_i \leftarrow 1$ ▷ Write 1 at index
Return: binary vector $\mathbf{f}$

Neural graph fingerprints

Input: molecule, radius R, weights $H_1^1 \ldots H_R^5$, output weights $W_1 \ldots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule **do**
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R **do** ▷ for each layer
  for each atom a in molecule **do**
    $\mathbf{r}_1 \ldots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^{N} \mathbf{r}_i$ ▷ sum
    $\mathbf{r}_a \leftarrow \sigma(\mathbf{v} H_L^N)$ ▷ smooth function
    $\mathbf{i} \leftarrow \text{softmax}(\mathbf{r}_a W_L)$ ▷ sparsify
    $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{i}$ ▷ add to fingerprint
Return: real-valued vector $\mathbf{f}$

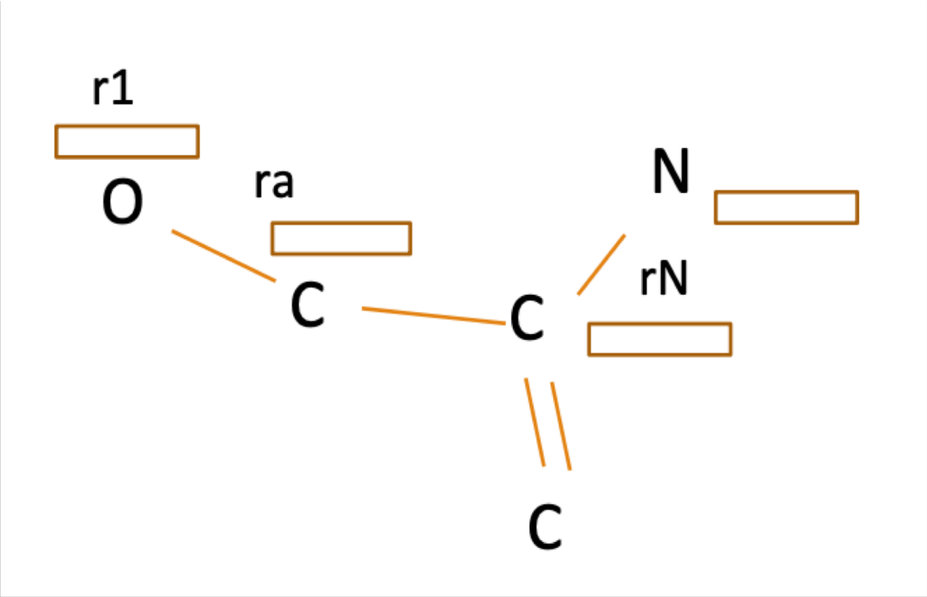
Every non-differentiable operation is replaced with a differentiable analog.

Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints, NIPS' 15

Neural Fingerprint (NIPS' 15)

Neural graph fingerprints

Input: molecule, radius R, weights $H_1^1 \ldots H_R^5$, output weights $W_1 \ldots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule **do**
    $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R **do** ▷ for each layer
    for each atom a in molecule **do**
        $\mathbf{r}_1 \ldots \mathbf{r}_N = \text{neighbors}(a)$
        $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^{N} \mathbf{r}_i$ ▷ sum
        $\mathbf{r}_a \leftarrow \sigma(\mathbf{v} H_L^N)$ ▷ smooth function
        $\mathbf{i} \leftarrow \text{softmax}(\mathbf{r}_a W_L)$ ▷ sparsify
        $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{i}$ ▷ add to fingerprint
Return: real-valued vector $\mathbf{f}$

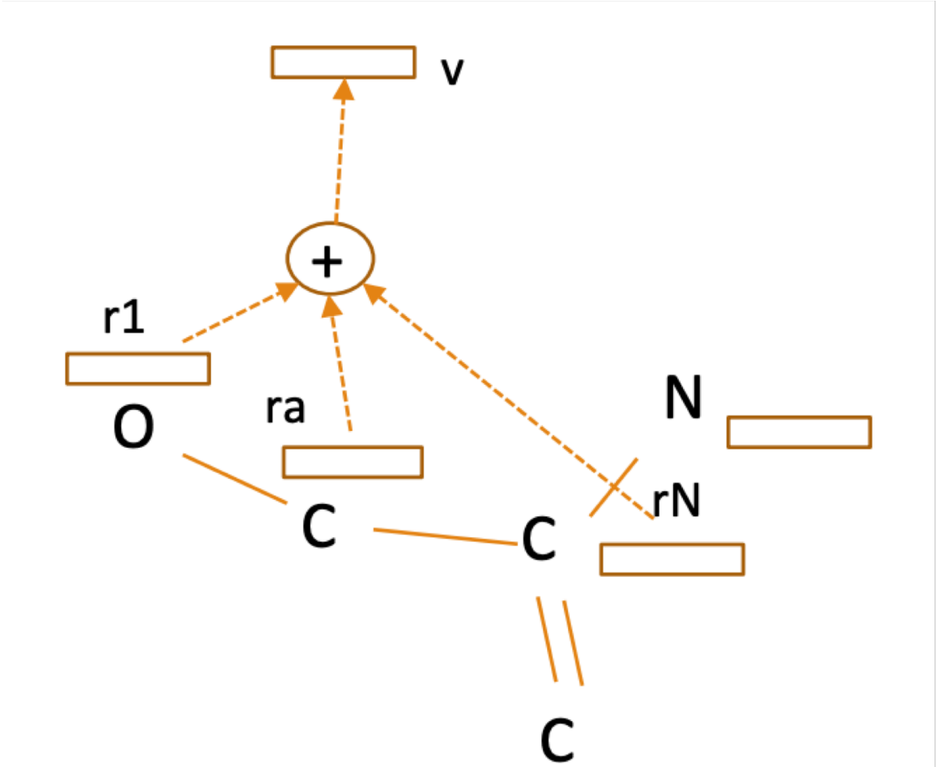


Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints, NIPS' 15

Neural Fingerprint (NIPS' 15)

Neural graph fingerprints

Input: molecule, radius R, weights $H_1^1 \ldots H_R^5$, output weights $W_1 \ldots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule **do**
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R **do** ▷ for each layer
  for each atom a in molecule **do**
    $\mathbf{r}_1 \ldots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^{N} \mathbf{r}_i$ ▷ sum
    $\mathbf{r}_a \leftarrow \sigma(\mathbf{v} H_L^N)$ ▷ smooth function
    $\mathbf{i} \leftarrow \text{softmax}(\mathbf{r}_a W_L)$ ▷ sparsify
    $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{i}$ ▷ add to fingerprint
Return: real-valued vector $\mathbf{f}$

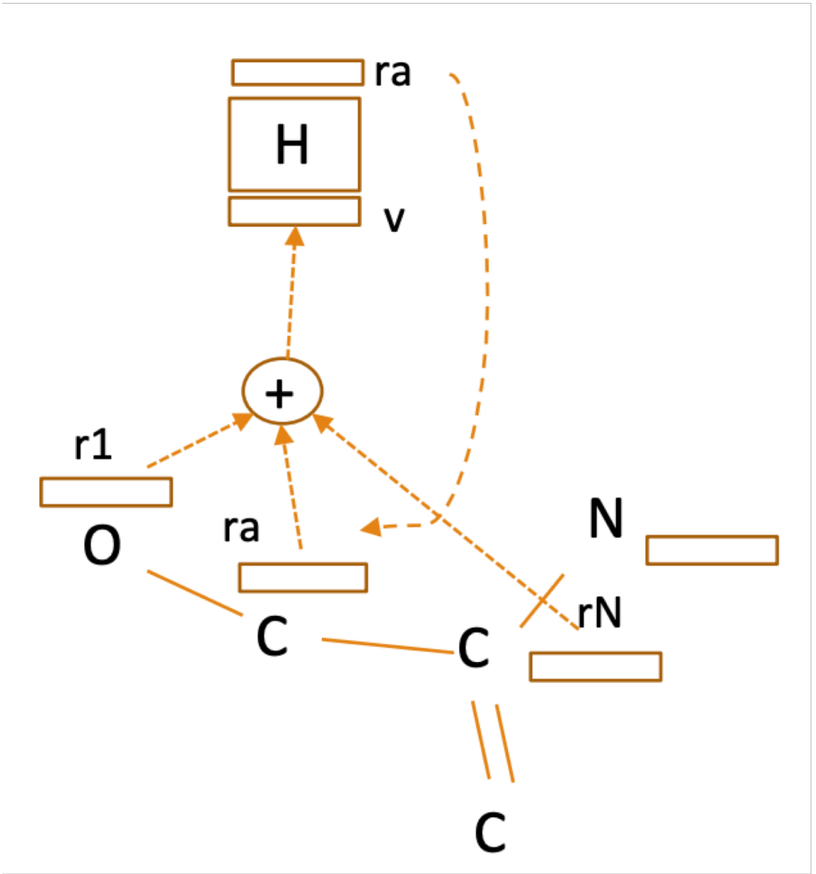


Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints, NIPS' 15

Neural Fingerprint (NIPS' 15)

Neural graph fingerprints

Input: molecule, radius R, weights $H_1^1 \ldots H_R^5$, output weights $W_1 \ldots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule **do**
    $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R **do** ▷ for each layer
    for each atom a in molecule **do**
        $\mathbf{r}_1 \ldots \mathbf{r}_N = \text{neighbors}(a)$
        $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^{N} \mathbf{r}_i$ ▷ sum
        $\mathbf{r}_a \leftarrow \sigma(\mathbf{v} H_L^N)$ ▷ smooth function
        $\mathbf{i} \leftarrow \text{softmax}(\mathbf{r}_a W_L)$ ▷ sparsify
        $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{i}$ ▷ add to fingerprint
Return: real-valued vector $\mathbf{f}$

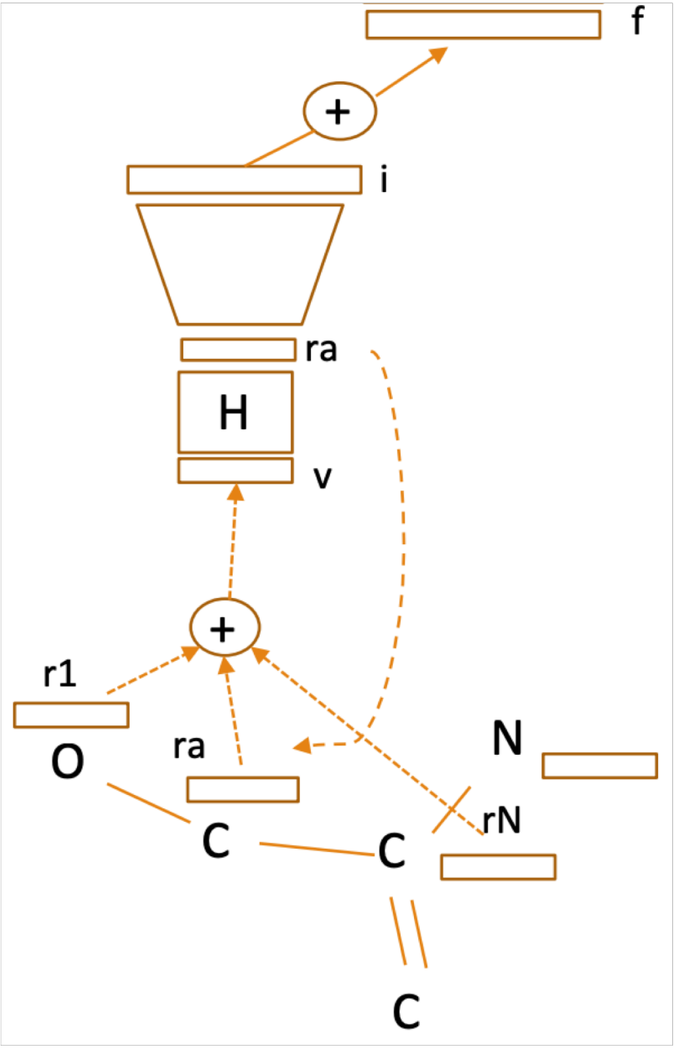


Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints, NIPS' 15

Neural Fingerprint (NIPS' 15)

Neural graph fingerprints

Input: molecule, radius R, weights $H_1^1 \ldots H_R^5$, output weights $W_1 \ldots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule **do**
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R **do** ▷ for each layer
  for each atom a in molecule **do**
    $\mathbf{r}_1 \ldots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^{N} \mathbf{r}_i$ ▷ sum
    $\mathbf{r}_a \leftarrow \sigma(\mathbf{v} H_L^N)$ ▷ smooth function
    $\mathbf{i} \leftarrow \text{softmax}(\mathbf{r}_a W_L)$ ▷ sparsify
    $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{i}$ ▷ add to fingerprint
Return: real-valued vector $\mathbf{f}$



Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints, NIPS' 15

Neural Fingerprint (NIPS' 15)

Neural graph fingerprints

Input: molecule, radius R, weights $H_1^1 \ldots H_R^5$, output weights $W_1 \ldots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule **do**
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R **do** ▷ for each layer
  for each atom a in molecule **do**
    $\mathbf{r}_1 \ldots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^{N} \mathbf{r}_i$ ▷ sum
    $\mathbf{r}_a \leftarrow \sigma(\mathbf{v} H_L^N)$ ▷ smooth function
    $\mathbf{i} \leftarrow \text{softmax}(\mathbf{r}_a W_L)$ ▷ sparsify
    $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{i}$ ▷ add to fingerprint
Return: real-valued vector $\mathbf{f}$

This process is repeated many times to extract substructures with different levels

Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints, NIPS' 15

Neural Fingerprint (NIPS' 15)

Experiment: predict solubility, drug efficacy, and organic photovoltaic efficiency

Dataset Units	Solubility log Mol/L	Drug efficacy EC_{50} in nM	Photovoltaic efficiency percent
Predict mean	4.29 ± 0.40	1.47 ± 0.07	6.40 ± 0.09
Circular FPs + linear layer	1.84 ± 0.08	$\mathbf{1.13 \pm 0.03}$	2.62 ± 0.07
Circular FPs + neural net	1.40 ± 0.15	1.24 ± 0.03	2.04 ± 0.07
Neural FPs + linear layer	0.74 ± 0.09	$\mathbf{1.16 \pm 0.03}$	2.71 ± 0.13
Neural FPs + neural net	$\mathbf{0.53 \pm 0.07}$	$\mathbf{1.17 \pm 0.03}$	$\mathbf{1.44 \pm 0.11}$

Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints, NIPS' 15

Summary of Molecular Representation

Traditional Molecular Representation

- 1D: properties
- 2D: Circular fingerprints – e.g. ECFP

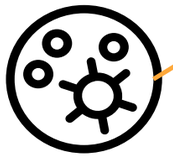
Neural network based molecular representation

- Mol2Vec
- Neural Fingerprint

Quantitative structure-activity relationship (QSAR) modeling

QSAR: Quantitative structure–activity relationship

Molecule Property Prediction

f () = 

Drug molecule

Molecule property

<u>Intrinsic Properties</u>

Molar volume, molecular weight, connectivity indices

<u>Chemical Properties</u>

pKa, Log P, Solubility, Stability

<u>Biological Properties</u>

Activity, Toxicity, Pharmacokinetics

Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships.

Ma, Junshui, Robert P. Sheridan, Andy Liaw, George E. Dahl, and Vladimir Svetnik. 2015. *Journal of Chemical Information and Modeling* 55 (2): 263–74

2012

MERCK
Be well

Merck Molecular Activity Challenge

Help develop safe and effective medicines by predicting molecular activity.

$40,000 · 236 teams · 7 years ago

Overview Data Discussion Leaderboard Rules

Overview

Description

Prizes

Evaluation

Visualization-Prospect

Submission-Instructions

Winners

Help enable the development of safe, effective medicines.

When developing new medicines it is important to identify molecules that are highly active toward their intended targets but not toward other targets that might cause side effects. The objective of this competition is to identify the best statistical techniques for predicting biological activities of different molecules, both on- and off-target, given numerical descriptors generated from their chemical structures

The challenge is based on 15 molecular activity data sets, each for a biologically relevant target. Each row corresponds to a molecule and contains descriptors derived from that molecule's chemical structure.

In addition to the prediction competition, Merck is also hosting a visualization challenge with a $2,000 prize for the most insightful and elegant graphical representations of the data.

Prizes total **$40,000**.

Ma et al. 2015. “Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships.” *Journal of Chemical Information and Modeling* 55 (2): 263–74.

Deep learning for Quantitative Structure-Activity Relationships (QSAR)

f(X = molecular descriptors (features))= y = activity

Ma et al. 2015. “Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships.” *Journal of Chemical Information and Modeling* 55 (2): 263–74.

Datasets and tasks

15 tasks/datasets (Kaggle competition)

+ **15** additional datasets

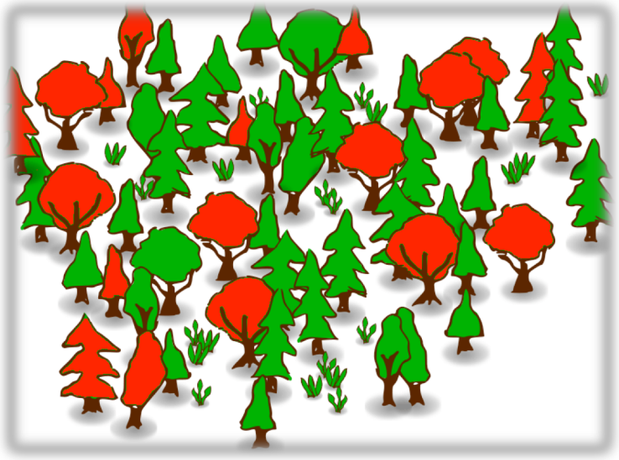
Largest dataset has **318,795** molecules

and **12,508** descriptors/features

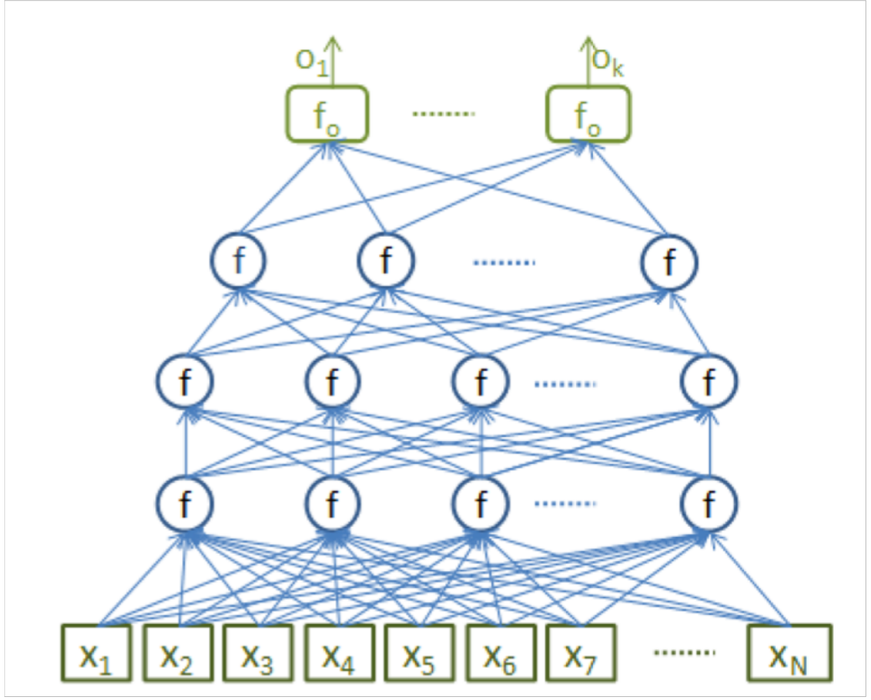
data set	type	description	number of molecules	number of unique AP, descriptors
		Kaggle Data Sets		
3A4	ADME	CYP P450 3A4 inhibition −log(IC50) M	50000	9491
CB1	target	binding to cannabinoid receptor 1 −log(IC50) M	11640	5877
DPP4	target	inhibition of dipeptidyl peptidase 4 −log(IC50) M	8327	5203
HIVINT	target	inhibition of HIV integrase in a cell based assay −log(IC50) M	2421	4306
HIVPROT	target	inhibition of HIV protease −log(IC50) M	4311	6274
LOGD	ADME	logD measured by HPLC method	50000	8921
METAB	ADME	percent remaining after 30 min microsomal incubation	2092	4595
		kinin1 (substance P) receptor binding −log(IC50) M	13482	5803
		1 receptor −log(K_i) M	7135	4730
		2 receptor −log(K_i) M	14875	5790
		oprotein log(BA/AB)	8603	5135
		tein binding log(bound/unbound)	11622	5470
		ity) at 2 mg/kg	7821	5698
		4 inhibitions log(IC50 without NADPH/IC50 with	5559	5945
		hibition −log(IC50) M	6924	5552
		Additional Data Sets		
2C8	ADME	CYP P450 2C8 inhibition −log(IC50) M	29958	8217
		50) M	189670	11730
		50) M	50000	9729
		−log(IC50) M	2763	5242
		50) M	17469	6200
			50000	8959
		(clearance) μL/min·mg	23292	6782
		50) M	12843	6596
		M	9536	6136
FASSIF	ADME	solubility in simulated gut conditions log(solubility) mol/L	89531	9541
HERG	ADME	inhibition of hERG channel −log(IC50) M	50000	9388
HERG (full data set)	ADME	inhibition of hERG ion channel −log(IC50) M	318795	12508
NAV	ADME	inhibition of Nav1.5 ion channel −log(IC50) M	50000	8302
PAPP	ADME	apparent passive permeability in PK1 cells log(permeability) cm/s	30938	7713
PXR	ADME	induction of 3A4 by pregnane X receptor; percentage relative to rifampicin	50000	9282

Methods

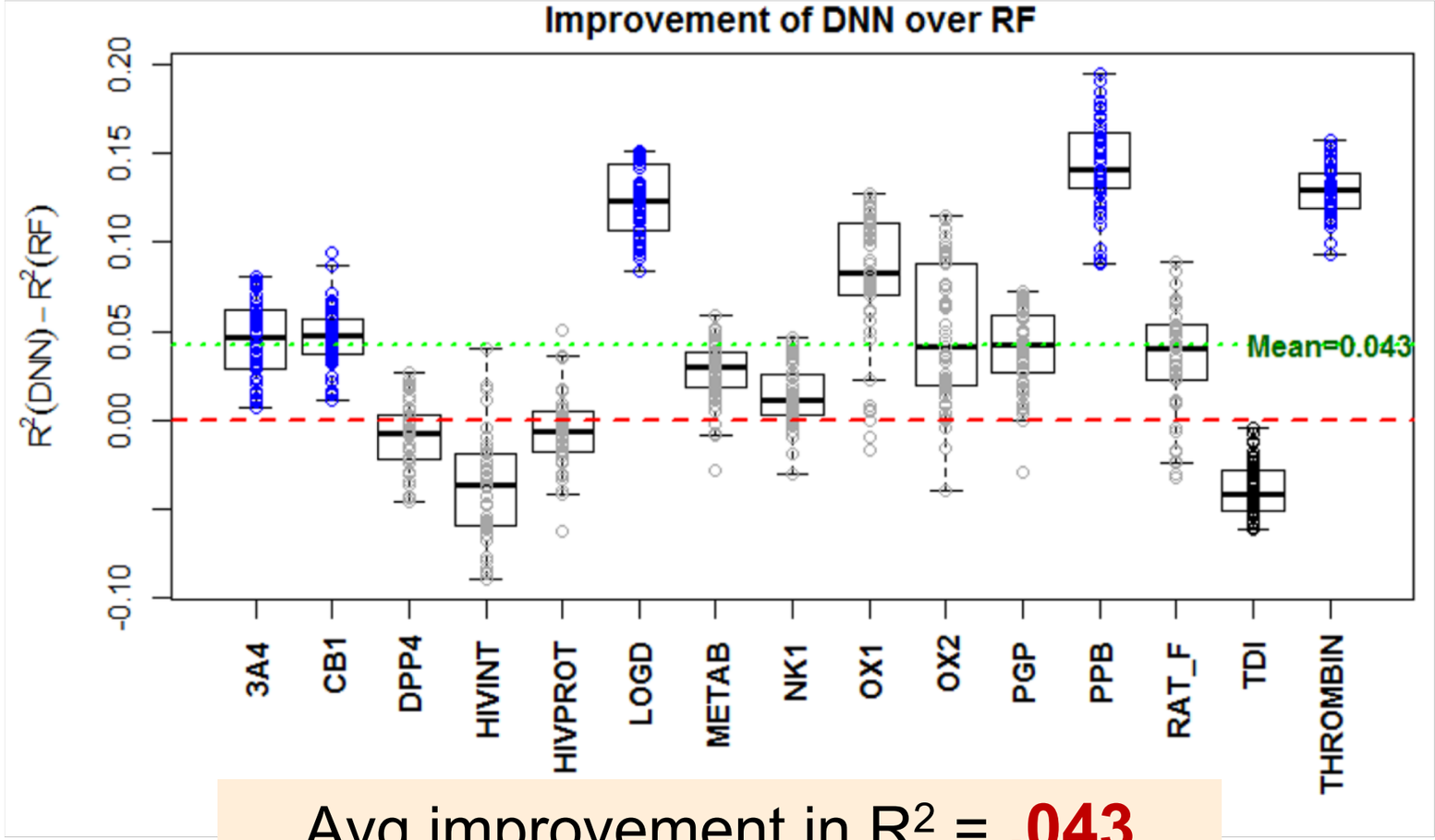
- Previous state of the art
  - Random forest (RF)



- Fully connected neural networks with 1 or 2 hidden layers



Results on Kaggle competition data



Avg improvement in R^2 = **.043**

RF = .423

DNN = .496

Results on additional data

Avg improvement **13.9%**
RF = .361
DNN = .411

Table 3. Comparing RF with DNN Trained Using Recommended Parameter Settings on 15 Additional Datasets

data set	random forest (R^2)	individual DNN (R^2)
2C8	0.158	**0.255**
2C9BIG	0.279	**0.363**
2D6	0.130	**0.195**
A-II	0.805	**0.812**
BACE	0.629	**0.644**
CAV	0.399	**0.463**
CLINT	0.393	**0.554**
ERK2	**0.257**	0.198
FACTORXIA	0.241	**0.244**
FASSIF	**0.294**	0.271
HERG	0.305	**0.352**
HERGfull	0.294	**0.367**
NAV	0.277	**0.347**
PAPP	0.621	**0.678**
PXR	0.333	**0.416**
mean	0.361	***0.411***

Automatic Generation of Complementary Descriptors with Molecular Graph Networks

J. Chem. Inf. Model. **2005,** *45,* 1159−1168 **1159**

Automatic Generation of Complementary Descriptors with Molecular Graph Networks

Christian Merkwirth*,†,‡ and Thomas Lengauer*,†

Computational Biology and Applied Algorithmics Group, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, and Department for Information Technology, Faculty of Physics, Astronomy, and Applied Computer Science, Jagiellonian University, Reymonta 4, 30-059 Kraków, Poland

Received December 20, 2004

Motivation: How to automatically generate predictive chemical descriptors

- ***Chemical descriptors*** can quantify properties or characteristics of molecules, but expensive **feature engineering** is required

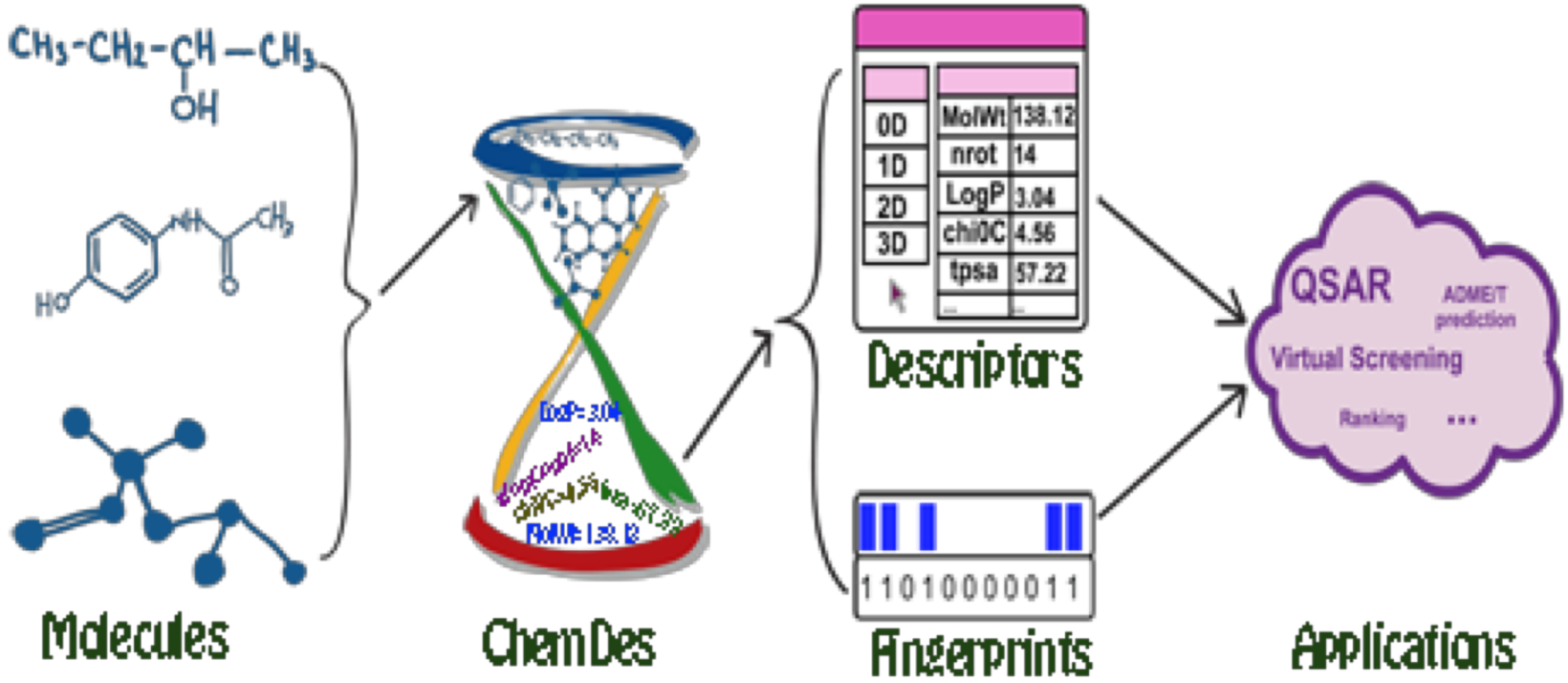


Figure source: http://www.scbdd.com/chemdes/

Molecular graph network (MGN)

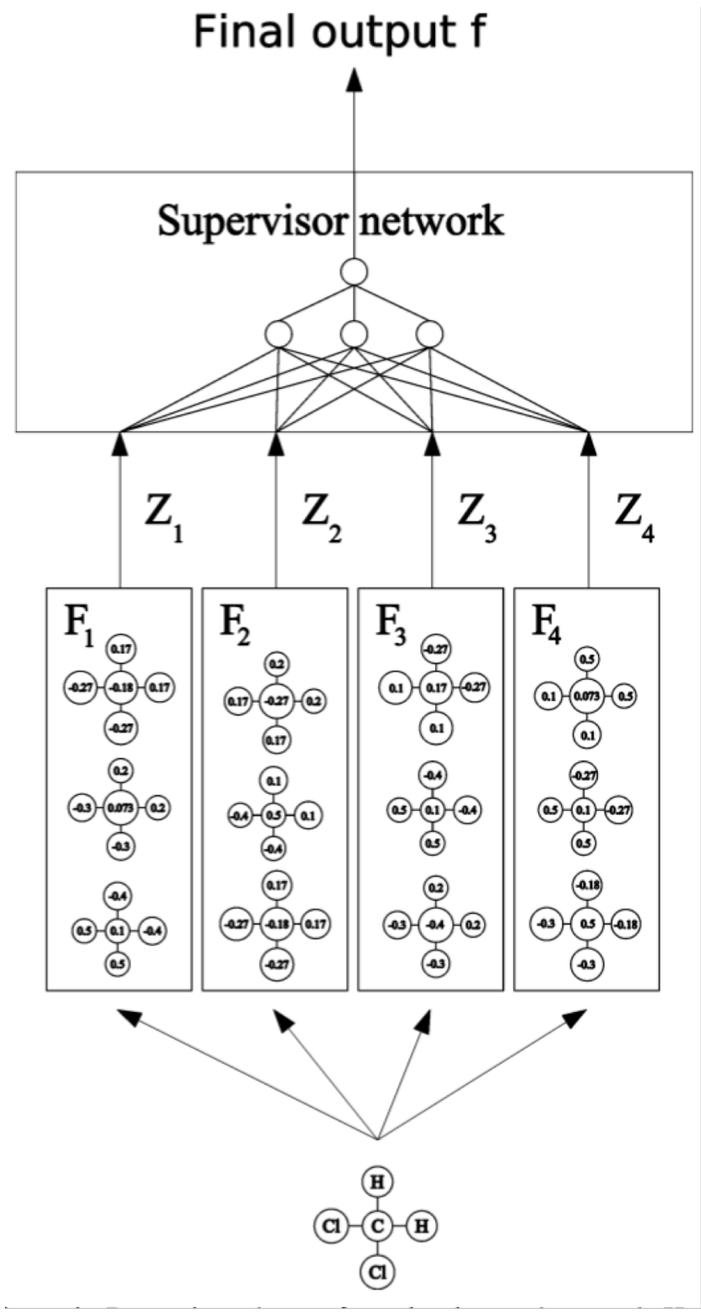
- Molecule structure –
  atoms as nodes, bonds as edges

- Feature net – Graph neural network (1D node embedding)

$$x_i^{t+1} = \sum_{\text{atom } j \text{ adjacent to } i} A^t_{e_i, B_{ij}} y_j^t + c^t_{e_i}$$

$$y_i^{t+1} = \sigma(x_i^{t+1})$$

- Supervisor network – Fully connected feed forward neural net



Result

- Data: 42 000 compounds from the Developmental Therapeutics Program AIDS antiviral screen data set
  - 41,179 compounds: confirmed inactive (CI)
  - 1080 compounds: confirmed moderately active (CM)
  - 423 compounds: confirmed active (CA)
- Confusion matrix

	predicted class		
actual class	CI	CM	CA
CI	0.835	0.126	0.038
CM	0.408	0.380	0.212
CA	0.124	0.187	0.690

Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction

Connor W. Coley, Regina Barzilay, William H. Green, Tommi S. Jaakkola, and Klavs F. Jensen

JOURNAL OF CHEMICAL INFORMATION AND MODELING

July 2017

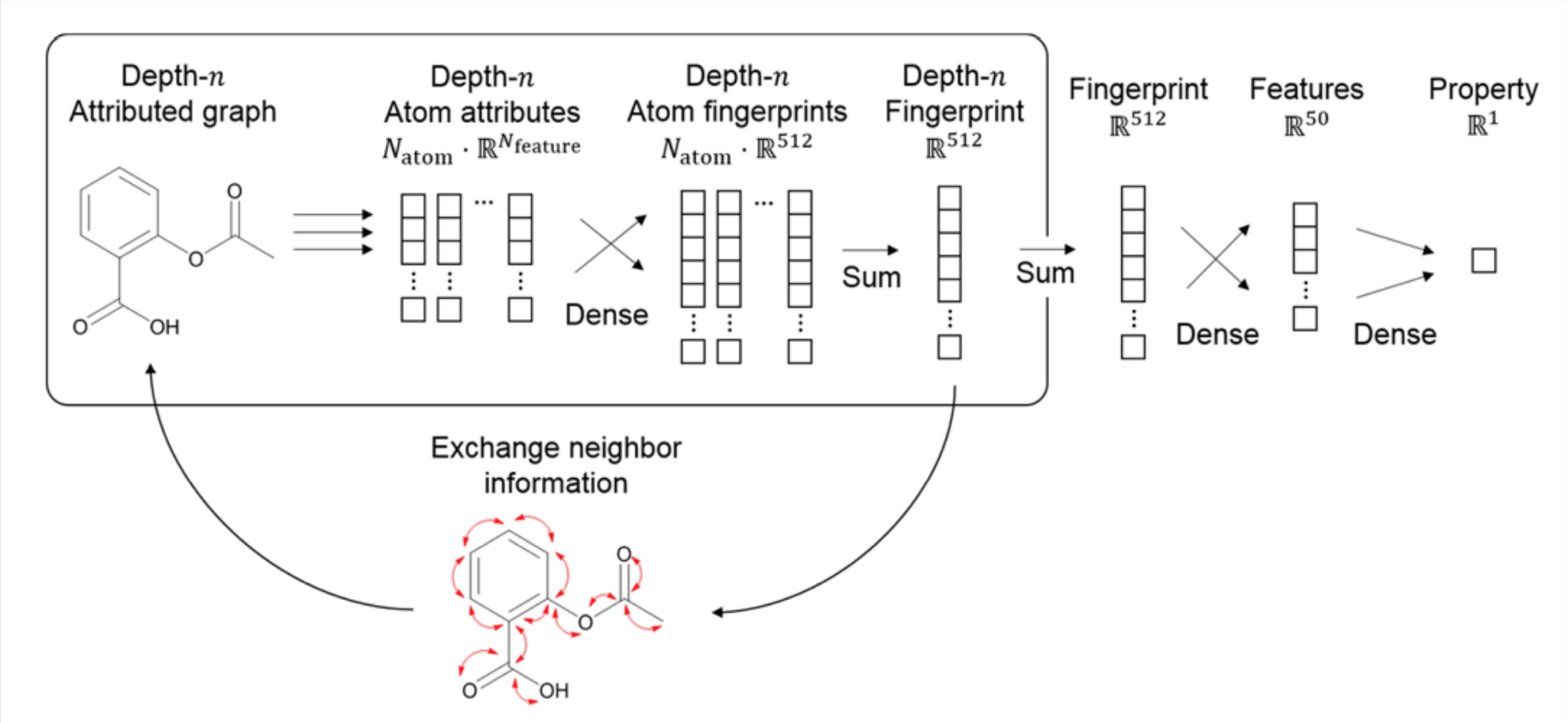
Convolutional Embedding of Molecular Graphs

Task: Given the molecular graph, predict Octanol solubility, Aqueous solubility, Melting point.

Intuition: “Predictive models can assist in lead optimization and in determining whether drug candidates should proceed to later development stages.” To replace experimental High-Throughput Screening (HTS) to virtual HTS.

Coley et al. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction

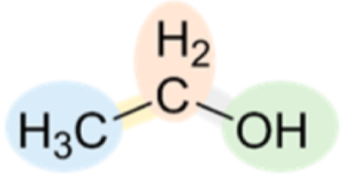
Convolutional Embedding of Molecular Graphs



Molecular Graphs as attribute graphs

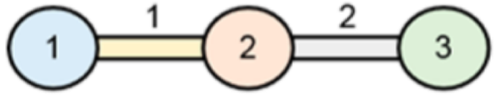
Undirected graphs with features on nodes and edges

Structure



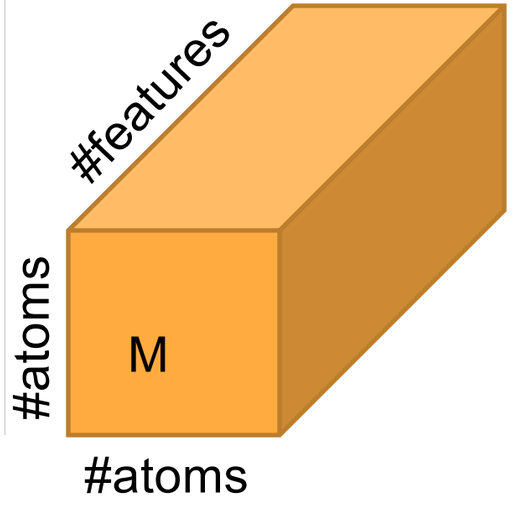
	Z	*# neighbors*	*# hydrogens*	*Formal charge*
Node 1	6	1	3	0
Node 2	6	2	2	0
Node 3	8	1	1	0

Attributed Graph



	Order	*Aromatic*	*Conjugated*	*In ring*	*Connects*
Edge 1	Single	No	No	No	(1, 2)
Edge 2	Single	No	No	No	(2, 3)

$$M_{ethanol} = \begin{bmatrix} 6,1,3,0,0,0,0,0,0 & 6,2,2,0,1,0,0,0,1 & 0,0,0,0,0,0,0,0,0 \\ 6,1,3,0,1,0,0,0,1 & 6,2,2,0,0,0,0,0,0 & 8,1,1,0,1,0,0,0,1 \\ 0,0,0,0,0,0,0,0,0 & 6,2,2,0,1,0,0,0,1 & 8,1,1,0,0,0,0,0,0 \end{bmatrix}$$



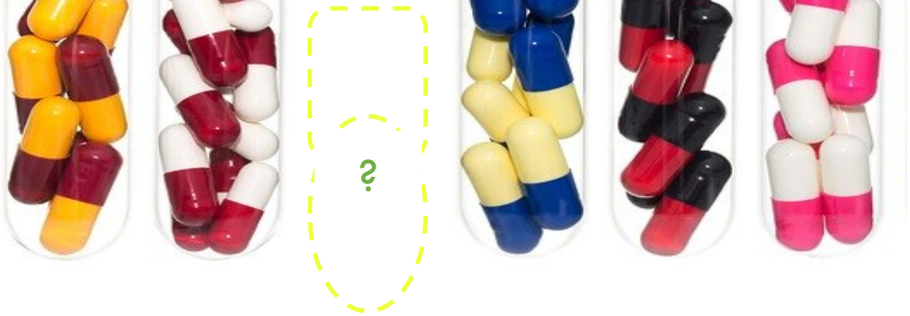
Convolutional Embedding of Molecular Graphs

Table 3. 5-Fold CV Performance on the Abraham Octanol Solubility Dataset, Averaged over Three Runs[a]

Model	Required data	Number of samples	MSE	MAE	SD
Best SVM baseline	–	245[c]	0.467 ± 0.019	0.520 ± 0.008	0.680 ± 0.013
GSE[34]	Melting point	223			0.71
Abraham and Acree, no m.p.[33]	Four empirical descriptors	282[b]			0.63
Abraham and Acree, m.p.[33]	Four empirical descriptors and melting point	282[b]			0.47
CNN-Ab-oct-representative[d]	–	245[c]	0.413 ± 0.018	0.496 ± 0.014	0.641 ± 0.011
CNN-Ab-oct-representative	–	245[c]	0.338 ± 0.005	0.455 ± 0.007	0.581 ± 0.005
CNN-Ab-oct-consensus	–	245[c]	0.328 ± 0.022	0.455 ± 0.015	0.573 ± 0.019

Table 4. 5-Fold CV Performance on the Delaney Aqueous Solubility Dataset, Averaged over Three Runs[a]

Model	Number of samples	MSE	MAE	SD
Best SVM baseline	1116[b]	1.255 ± 0.011	0.821 ± 0.006	1.117 ± 0.004
Lusci et al.[18]	1144	0.34	0.43	
Duvenaud et al.[19]	1144		0.52	
CNN-De-aq-representative[c]	1116[b]	0.334 ± 0.011	0.424 ± 0.005	0.577 ± 0.010
CNN-De-aq-representative	1116[b]	0.312 ± 0.003	0.401 ± 0.002	0.559 ± 0.003
CNN-De-aq-consensus	1116[b]	0.314 ± 0.008	0.403 ± 0.005	0.560 ± 0.007



Polyadic Regression and its Application to Chemogenomics

Ioakeim Perros, Fei Wang, Ping Zhang, Peter Walker, Richard Vuduc, Jyotishman Pathak, and Jimeng Sun

In Proceedings of the 2017 SIAM International Conference on Data Mining (SDM 2017)

SunLab Enlighten the World

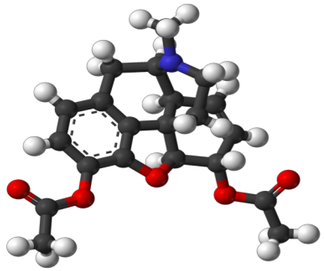
Georgia Tech

Weill Cornell Medicine

IBM Research

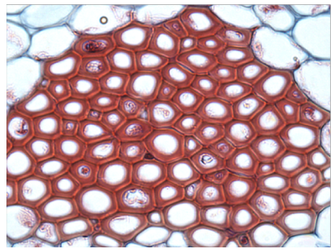
NAVAL MEDICAL RESEARCH CENTER U.S. NAVY

Chemogenomics methodology:
Drug-induced, cell-specific gene expression analysis



Drug i_1

Gene i_2



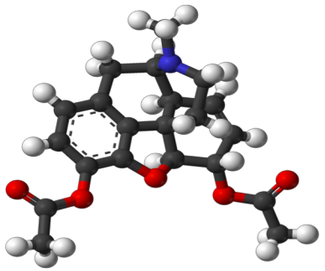
Tissue i_3

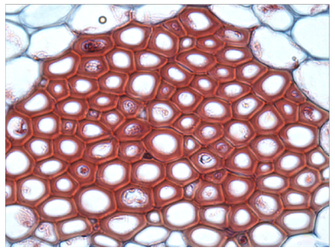
Lab measurement y_{i_1,i_2,i_3} indicates effectiveness of drug i_1 towards treating tissue i_3, w.r.t. gene i_2.

Important for drug repositioning and revealing drug mechanisms

Challenges

1. Not all drugs are measured across all tissues
   - Expensive or impossible to measure

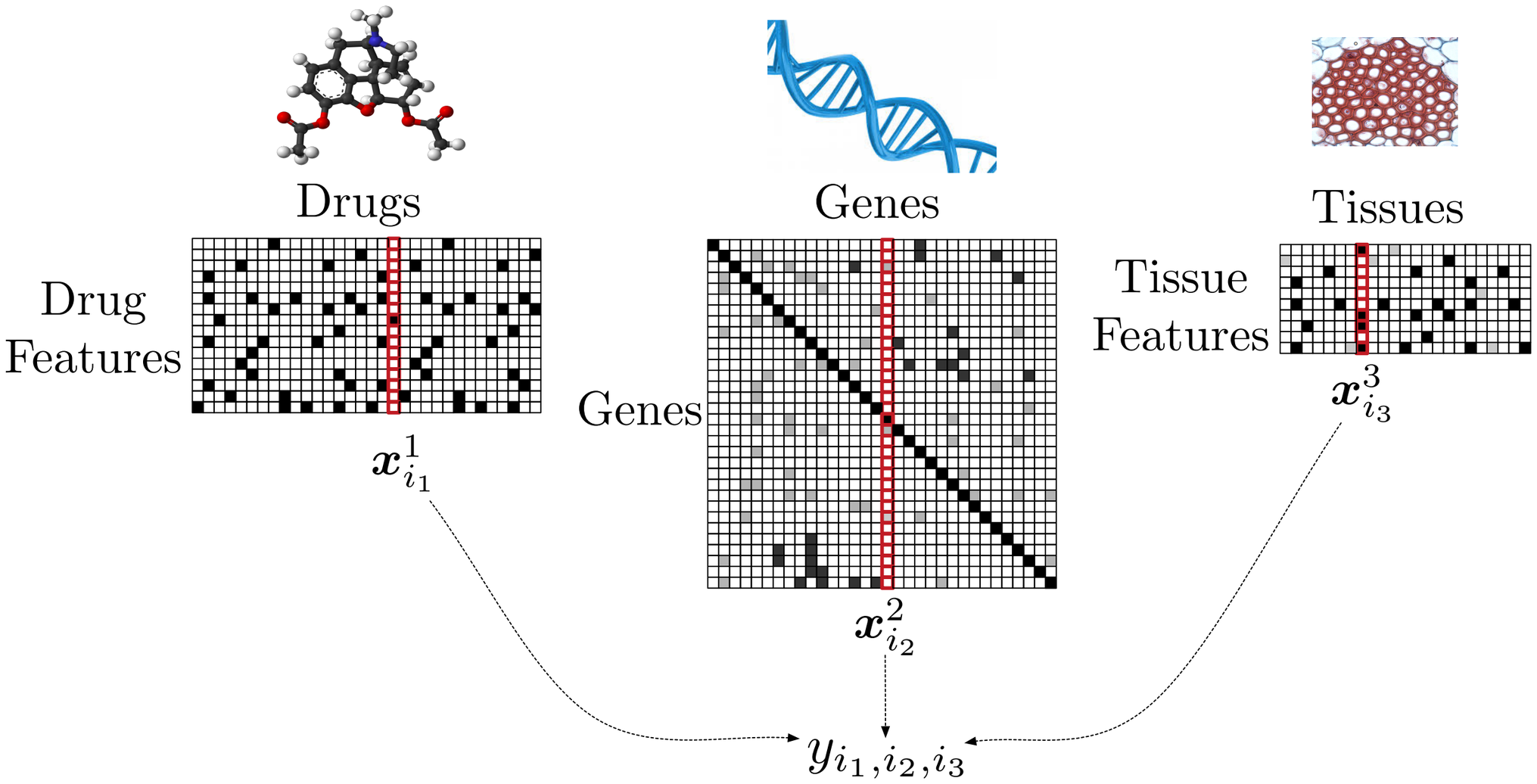




2. Exploit external knowledge and estimate expression for **new drugs,** for which we have **no measurements**
   - Focus on a small subset of targeted lab trials and cut down the costs

Problem: Polyadic Prediction

Predicted measurements are associated with ordered tuple of objects

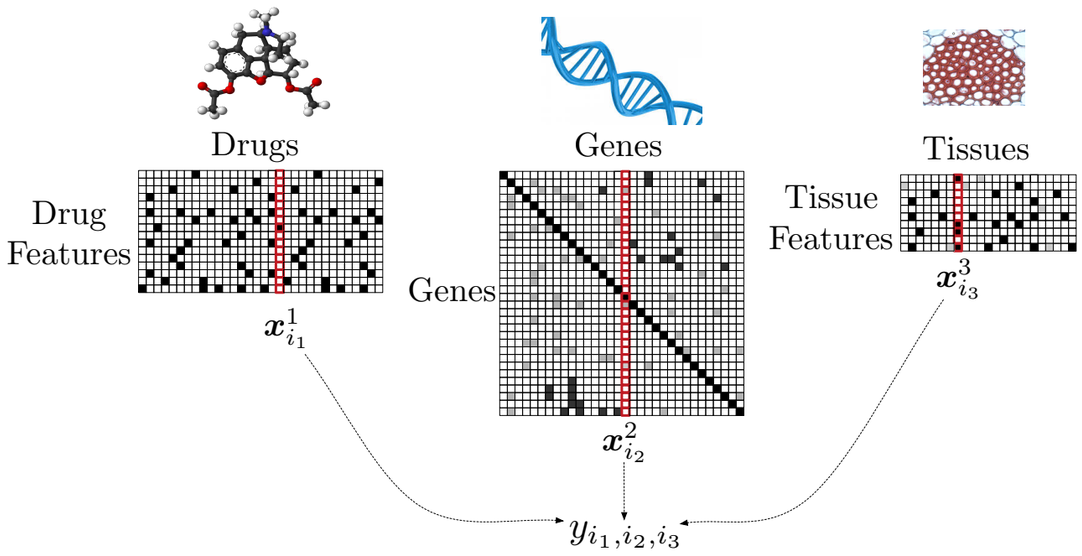


$\mathcal{S} \times_1 \boldsymbol{x}^1_{i_1} \times_2 \boldsymbol{x}^2_{i_2} \cdots \times_K \boldsymbol{x}^K_{i_K}$: vector-tensor analogue of vector-matrix multiplication

Polyadic Regression core model

$$
\begin{aligned}
& f(\boldsymbol{x}^1_{i_1}, \boldsymbol{x}^2_{i_2}, \cdots, \boldsymbol{x}^K_{i_K}) \\
&= b + \underbrace{\sum_{k=1}^{K} (\boldsymbol{w}^k)^\top \boldsymbol{x}^k_{i_k}}_{\text{linear terms}} + \underbrace{\sum_{uv} (\boldsymbol{x}^u_{i_u})^\top \mathbf{S}^{uv} \boldsymbol{x}^v_{i_v}}_{\text{dyadic interactions}} \\
&\quad + \underbrace{\sum_{uvr} \mathcal{S}^{uvr} \times_1 \boldsymbol{x}^u_{i_u} \times_2 \boldsymbol{x}^v_{i_v} \times_3 \boldsymbol{x}^r_{i_r}}_{\text{triadic interactions}} + \cdots \\
&\quad + \underbrace{\mathcal{S} \times_1 \boldsymbol{x}^1_{i_1} \times_2 \boldsymbol{x}^2_{i_2} \cdots \times_K \boldsymbol{x}^K_{i_K}}_{\text{general polyadic interactions}}
\end{aligned}
$$

- Explores all inter-aspect interactions
- High model complexity



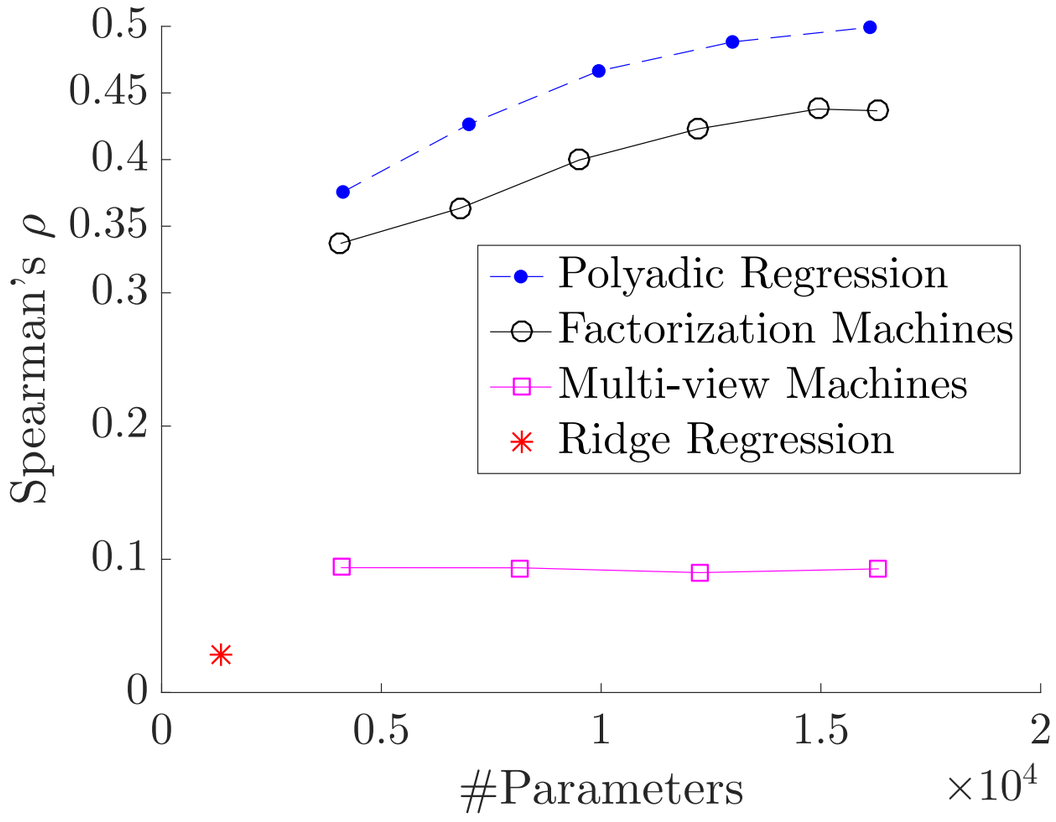
Real data description

- LINCS L1000 *publicly-available* drug-gene-tissue data: ~1000 genes, known to be maximally predictive
- 10 tissues with most expression profiles
- 850 genes for which we have similarity information (GOSemSim package in R)
- Drug features from PubChem: chemical structure of each drug

	Drugs	Genes	Tissues	Density	Values
Missing value	81	850	10	100%	688,500
New drug	500	850	10	44%	1,870,850

New drug experiment: constrain the train, validation and test sets to have no common drugs

Task 1: Estimating missing measurements



- Linear terms: low predictive value
- MVMs: joint factorization and regularization (even if linear terms are irrelevant)
- FMs: competitive, but do not include 3-order interactions
- Polyadic Regression: **highest accuracy**

Task 2: Predicting measurements for new drugs

Method	Spearman's ρ	#Parameters
Polyadic Regression	0.23025 ± 0.0063886	4471
Factorization Machines	0.1252 ± 0.0083942	4417
Multi-view Machines	0.0669 ± 0.017242	4425
Ridge Regression	0.0061	1473

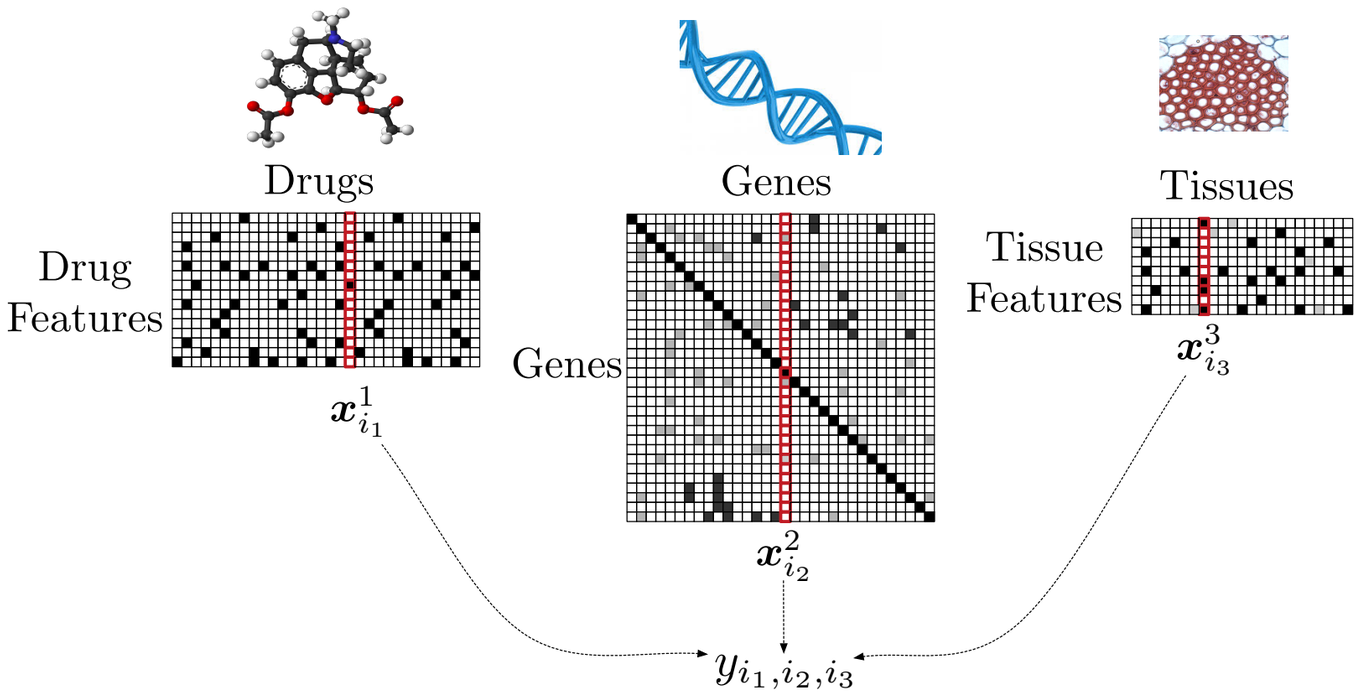
- **Polyadic Regression** achieves **0.1 increase in correlation** between the predicted and the true vector of measurements
- **Robustness** for different initialization of parameters

Summary: Polyadic Regression

Drug-perturbed, cell-specific gene expression

Predicting interactions among multiple data aspects

SIAM Data Mining (SDM) 2017



1) Estimate effectiveness of drug on tissues w.r.t. available genes

2) Predict for drugs unseen during training

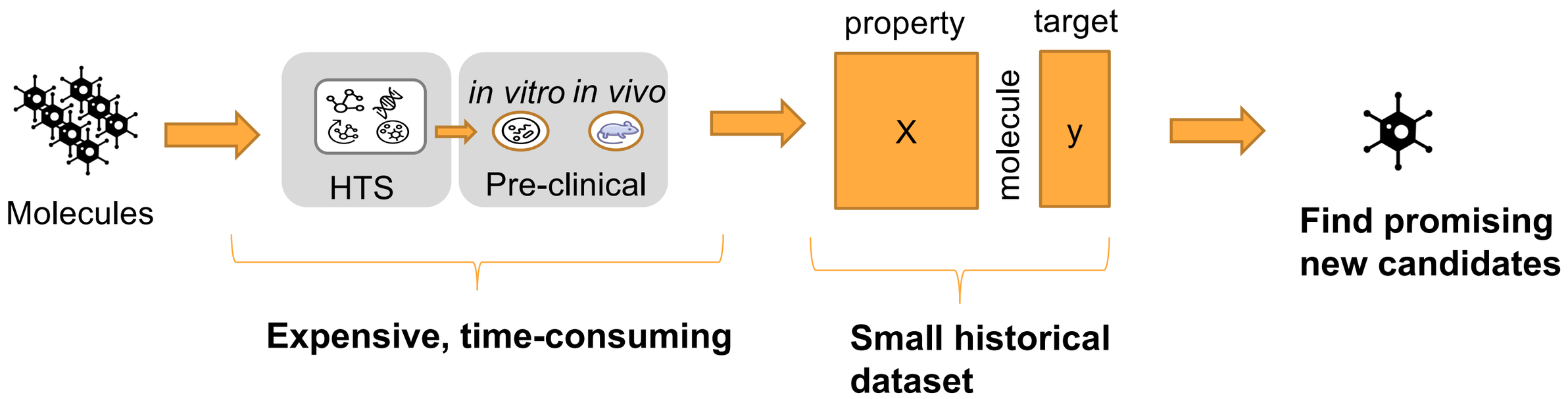
Perros et al 2017. "Polyadic Regression and Its Application to Chemogenomics." SDM'17

Low Data Drug Discovery with One-Shot Learning

Altae-Tran, Han, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande.

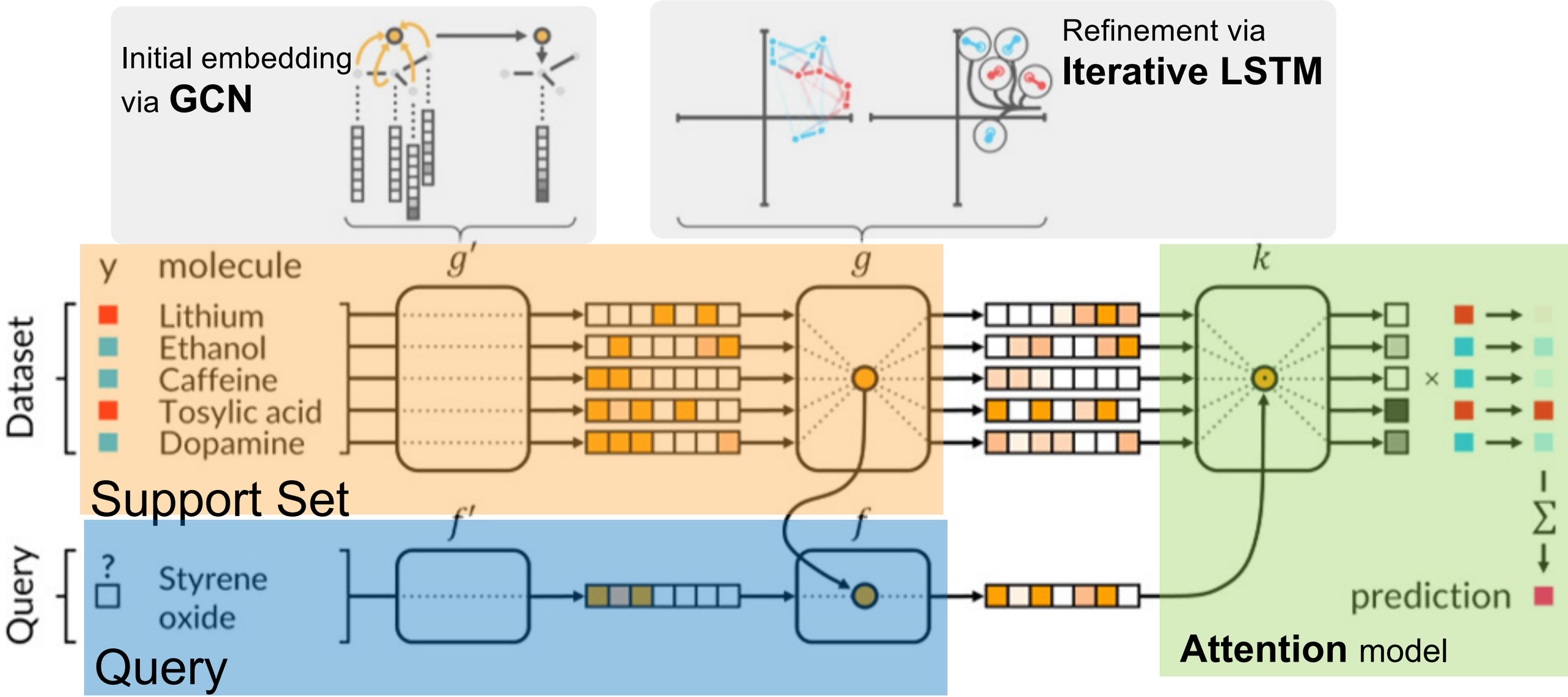
2017. *ACS Central Science* 3 (4): 283–93.

Motivation of one-shot learning for compound activity prediction



- How to find promising new drug candidates?
- How to find the candidates that are similar to the small number of active molecules?

One shot learning for compound activity prediction



Great prediction results on two datasets with very limited training data

Table 1. ROC-AUC Scores of Models on Median Held-out Task for Each Model on Tox21[a]

Tox21	RF (100 trees)	Graph Conv	Siamese	AttnLSTM	IterRefLSTM
10+/10−	0.586 ± 0.056	0.648 ± 0.029	0.820 ± 0.003	0.801 ± 0.001	**0.823 ± 0.002**
5+/10−	0.573 ± 0.060	0.637 ± 0.061	0.823 ± 0.004	0.753 ± 0.173	**0.830 ± 0.001**
1+/10−	0.551 ± 0.067	0.541 ± 0.093	**0.726 ± 0.173**	0.549 ± 0.088	0.724 ± 0.008
1+/5−	0.559 ± 0.063	0.595 ± 0.086	0.687 ± 0.210	0.593 ± 0.153	**0.795 ± 0.005**
1+/1−	0.535 ± 0.056	0.589 ± 0.068	0.657 ± 0.222	0.507 ± 0.079	**0.827 ± 0.001**

Table 2. ROC-AUC Scores of Models on Median Held-out Task for Each Model on SIDER[a]

SIDER	RF (100 trees)	Graph Conv	Siamese	AttnLSTM	IterRefLSTM
10+/10−	0.535 ± 0.036	0.483 ± 0.026	**0.687 ± 0.089**	0.553 ± 0.058	0.669 ± 0.007
5+/10−	0.533 ± 0.030	0.473 ± 0.029	0.648 ± 0.070	0.534 ± 0.053	**0.704 ± 0.002**
1+/10−	0.540 ± 0.034	0.447 ± 0.016	0.544 ± 0.056	0.506 ± 0.016	**0.556 ± 0.011**
1+/5−	0.529 ± 0.028	0.457 ± 0.029	0.530 ± 0.050	0.505 ± 0.022	**0.644 ± 0.012**
1+/1−	0.506 ± 0.039	0.468 ± 0.045	0.510 ± 0.016	0.501 ± 0.022	**0.697 ± 0.002**

10 active molecules, 10 inactive molecules

SIAMESE network, AttentionLSTM, and IterRefLSTM perform great

But inconsistent/poor performance on some dataset

Table 3. ROC-AUC Scores of Models on Median Held-out Task for Each Model on MUV[a]

MUV	RF (100 trees)	Graph Conv	Siamese	AttnLSTM	IterRefLSTM
10+/10−	**0.754 ± 0.064**	0.568 ± 0.085	0.601 ± 0.041	0.504 ± 0.058	0.499 ± 0.053
5+/10−	**0.730 ± 0.063**	0.565 ± 0.068	0.655 ± 0.166	0.507 ± 0.052	0.663 ± 0.019
1+/10−	0.556 ± 0.084	0.569 ± 0.061	**0.602 ± 0.118**	0.504 ± 0.044	0.569 ± 0.012
1+/5−	0.598 ± 0.067	0.554 ± 0.089	0.514 ± 0.053	0.515 ± 0.021	**0.632 ± 0.011**
1+/1−	**0.559 ± 0.095**	0.552 ± 0.084	0.500 ± 0.0001	0.500 ± 0.027	0.479 ± 0.037

- MUV dataset select structurally distinct positive examples.
- Poor performance on models leveraging structural similarity (SIAMESE, AttnLSTM, IterRefLSTM)

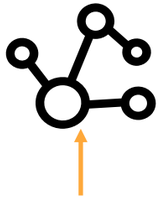
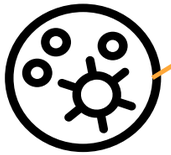
And generalization across datasets is poor

Table 4. ROC-AUC Scores of Models Trained on Tox21 on Median SIDER Task for Each Model on SIDER[a]

SIDER from Tox21	Siamese	AttnLSTM	IterRefLSTM
10+/10−	0.511 ± 0.031	0.509 ± 0.014	0.509 ± 0.012

- Even on the datasets which can be trained toward accurate models on themselves, those models do NOT generalize across datasets

Summary: QSAR: Quantitative structure–activity relationship

f () = 

Drug molecule

Molecule property

<u>Intrinsic Properties</u>
Molar volume, molecular weight, connectivity indices

<u>Chemical Properties</u>
pKa, Log P, Solubility, Stability

<u>Biological Properties</u>
Activity, Toxicity, Pharmacokinetics

- Deep neural networks
- Graph neural networks
- one-shot learning

Agenda

- Motivation
- Data
- **Tasks**
- Future Directions