Drug Discovery and Development

Molecular Representation Learning

QSAR Modeling

2019.11.15 AI Lab.

장성은

Drug Discovery and Development

- Drug discovery

  the process by which new candidate medications are discovered.

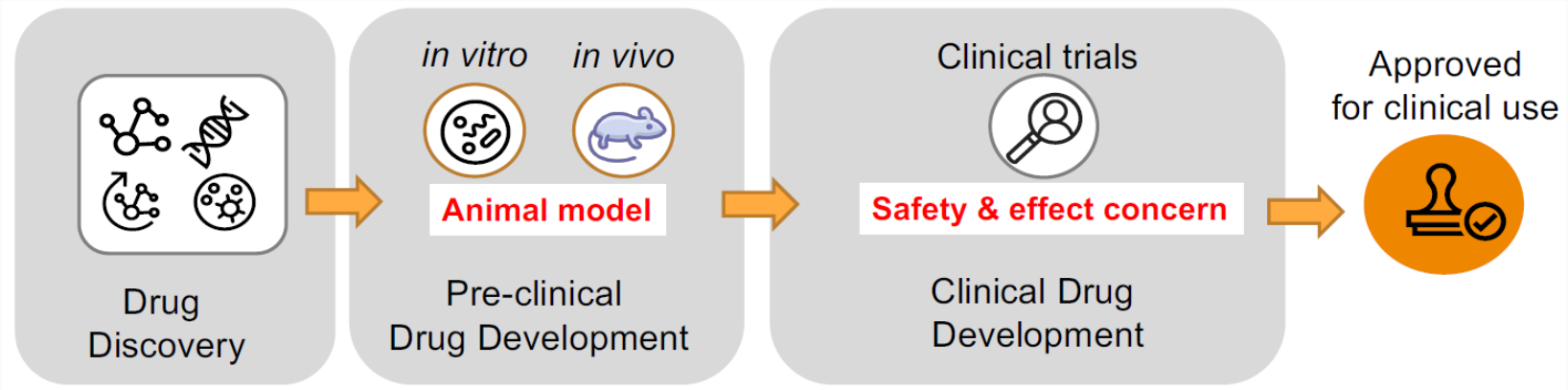
- Drug development

  the process of bringing a new pharmaceutical drug to the market once a lead compound has been identified through the process of drug discovery.

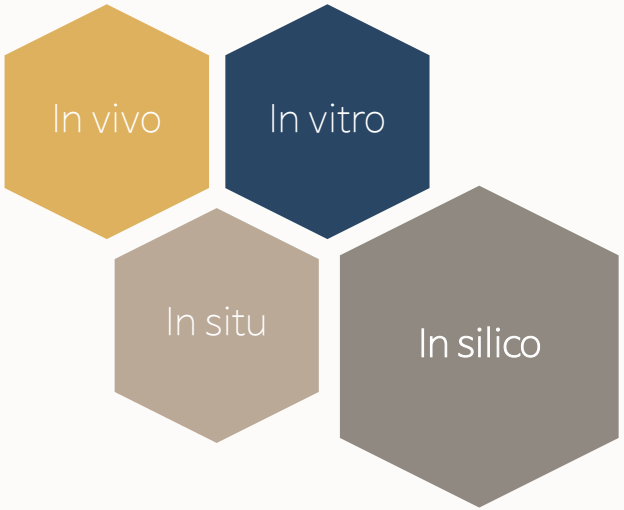
Drug Discovery and Development

Traditional Drug Discovery & Development Process

- Drug discovery and drug development are conducted through various biological and chemical experiments.



Silico modeling

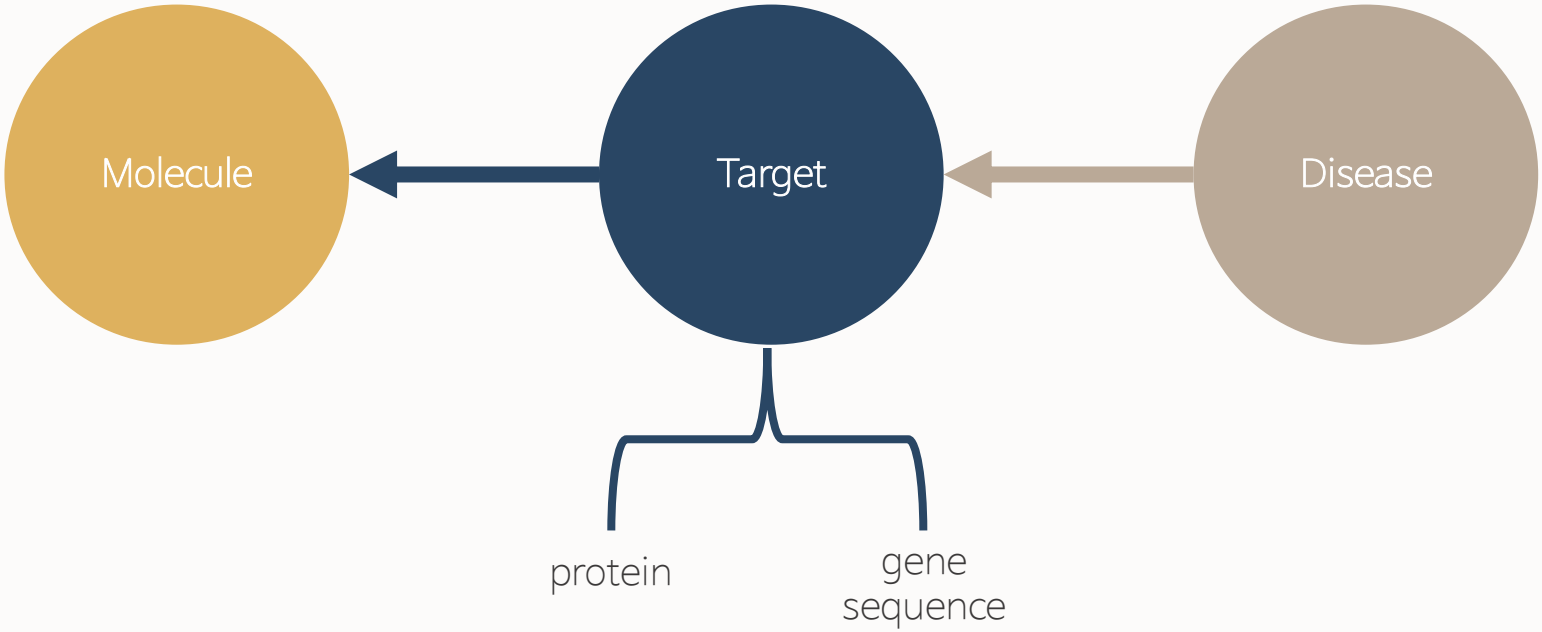


- Perform experiments on computer or via computer simulation
- Speed the rate of discovery
- Reduce the need for expensive lab work and clinical trials.

Source	Representation	Challenges
• Compound databases • Protein databases • Disease Knowledge • Biochemical literature • Clinical trial data	• Feature vectors • Graphs • Sequences • Text	• High dimensional • Small sample • Lack of labels • Complex interaction

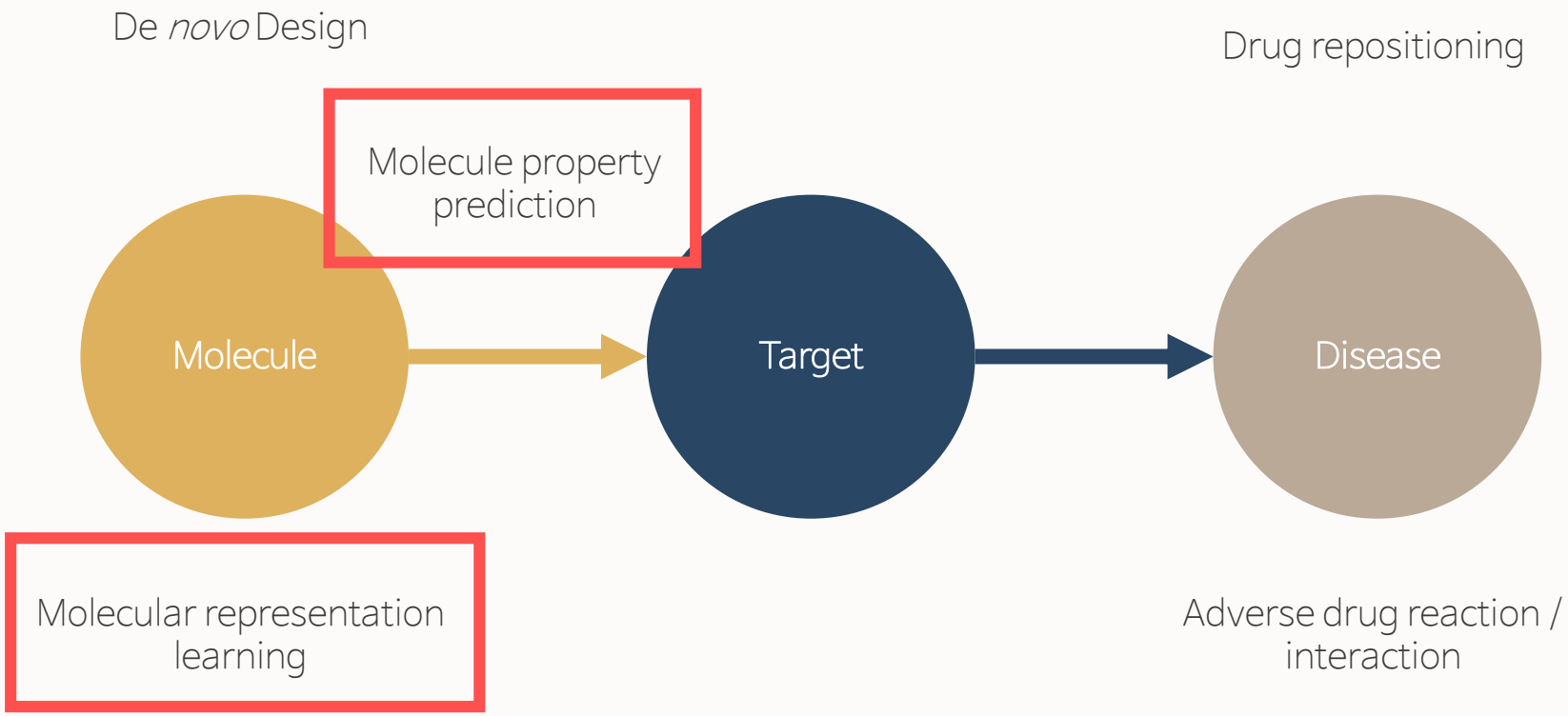
Silico modeling

Entities and Task of Drug Discovery



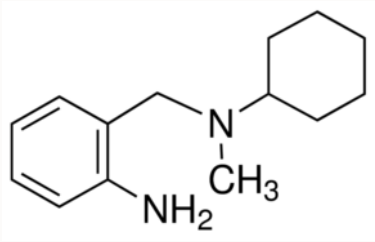
Silico modeling

Entities and Task of Drug Discovery



Molecular Representation Learning

$$f\left(\ \right) = embedding$$

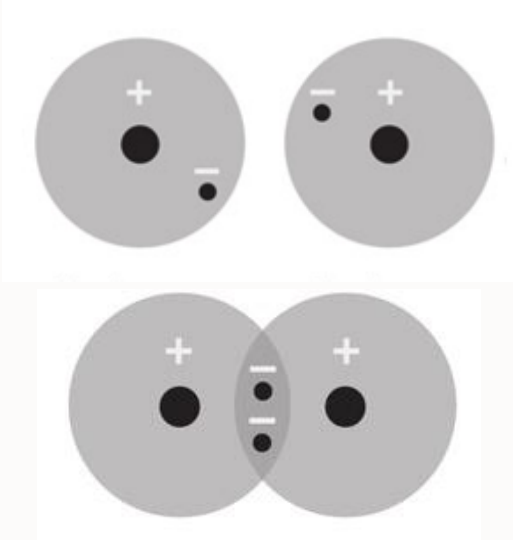


$drug\ molecule$

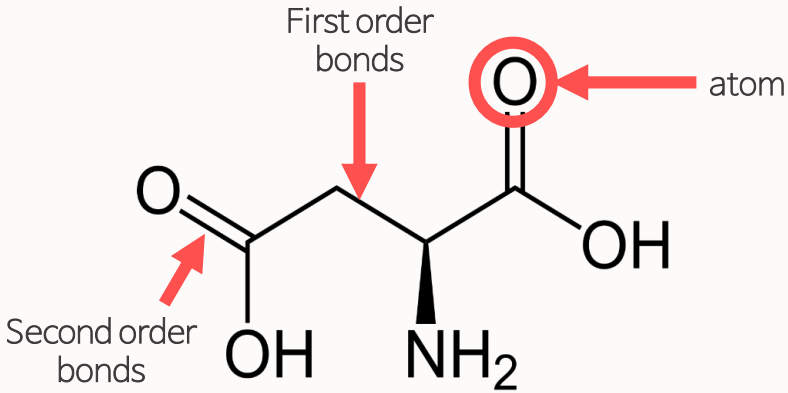
- Fundamental task for in silico modeling
- Map raw drug molecular data to low dimensional embeddings
- Similar molecules are embedded close together

Molecular Representation Learning

Back grounds



Covalent Bond



Molecule

Molecular Representation Learning

Weight Solubility Charge Atom types

Number of rotatable bonds ···

Graph of covalent and aromatic bonds

Atom arrangement in space

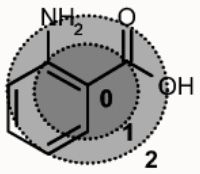
Molecular Representation Learning

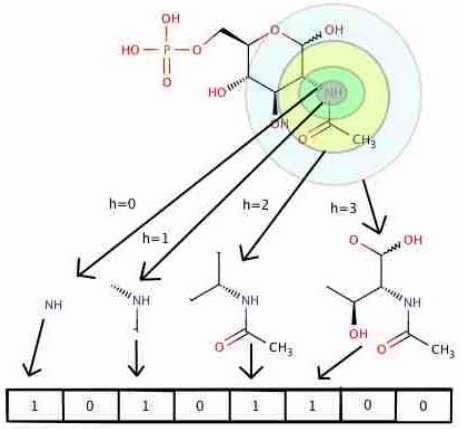
Traditional Approach

- 2D Descriptor
  - Extended circular fingerprints (ECFPx)

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 $\mathbf{f}$





Layer 0
Aromatic carbon (sp^2 orbital)

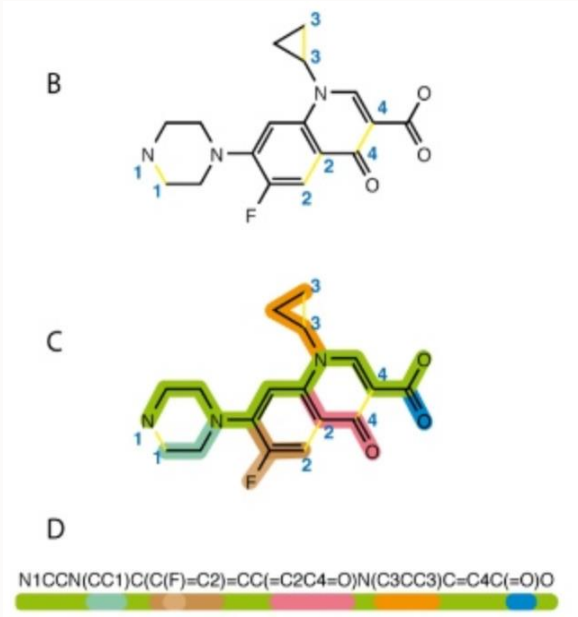
Layer 1
Aromatic carbon (sp^2 orbital)
Aromatic carbon (sp^2 orbital)
Aliphatic carbon (sp^2 orbital)

Layer 2
Aromatic carbon (sp^2 orbital)
Aromatic carbon (sp^2 orbital)
Nitrogen (sp^3 orbital)
Oxygen (sp^2 orbital)
Oxygen (sp^3 orbital)

Molecular Representation Learning

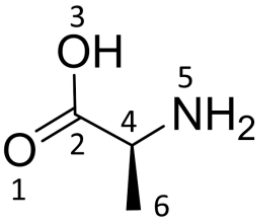
Traditional Approach

- 3D Descriptor



SMILES

Simplified Molecular-input Line-Entry System



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

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

Matrix representation for molecules

Molecular Representation Learning

Mol2Vec

- Overview

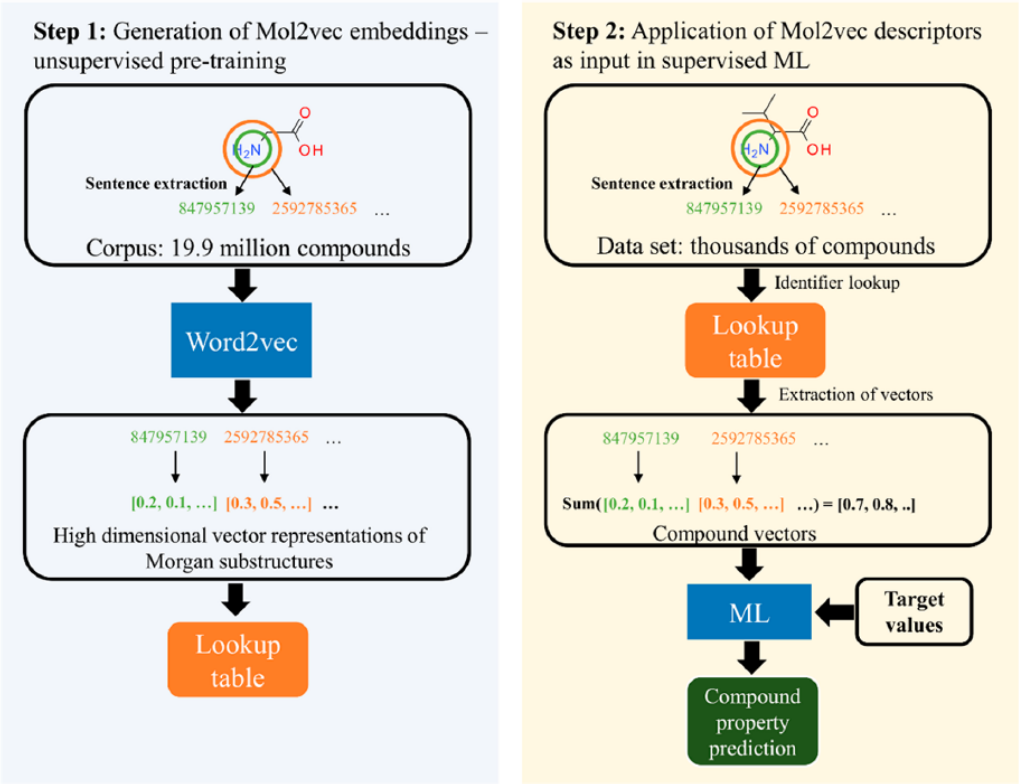


Figure 1. Overview of the generation and usage steps of Mol2vec. Step 1: Mol2vec embeddings (i.e., vector representations of substructures) are generated via an unsupervised pretraining step. Step 2: Application of Mol2vec vectors requires that substructure vectors be retrieved and summed to obtain compound vectors, which can finally be used to train a supervised prediction model.

Molecular Representation Learning

Mol2Vec

- Sentence and words

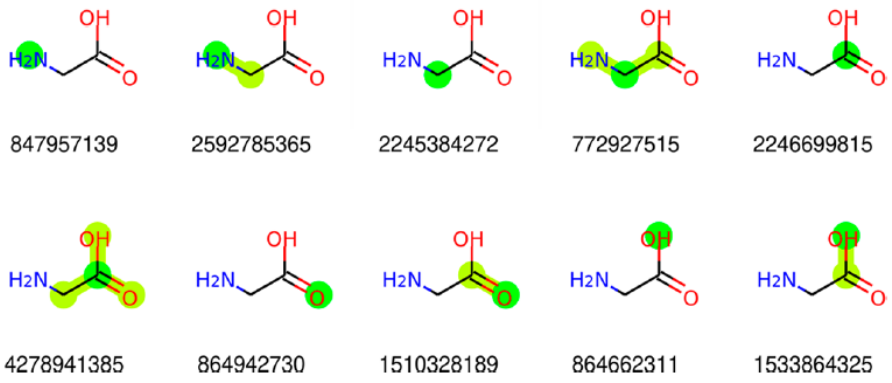


Figure 2. Depiction of identifiers obtained with the Morgan algorithm on the structure of glycine forming a molecular sentence. Identifiers are ordered in the same order as the atoms in the canonical SMILES representation for consistency reasons. If an atom has more than one identifier, the first identifier for that atom is the one for radius 0, followed by radius 1, etc.

Molecular Representation Learning

Mol2Vec

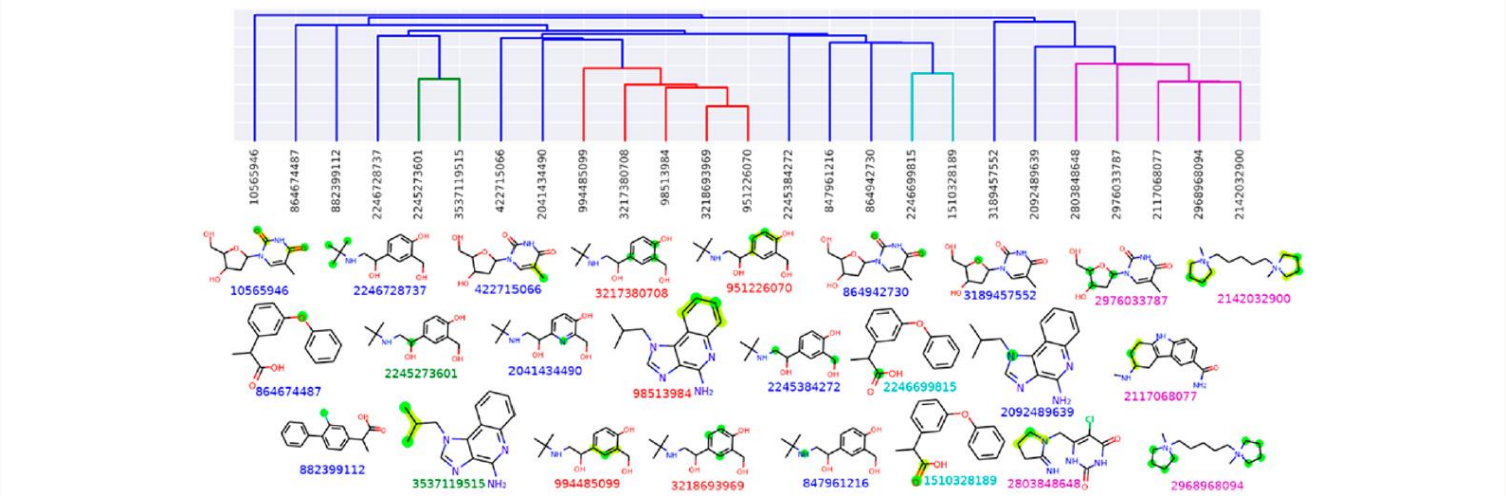


Figure 4. Dendrogram showing relationships among vectors representing the 25 most common substructures in the compound corpus. Substructures are depicted (central atoms in green and surrounding atoms in light green) on a representative compound from a pool of FDA-approved drugs.

Molecular Representation Learning

Mol2Vec

- Experiment

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

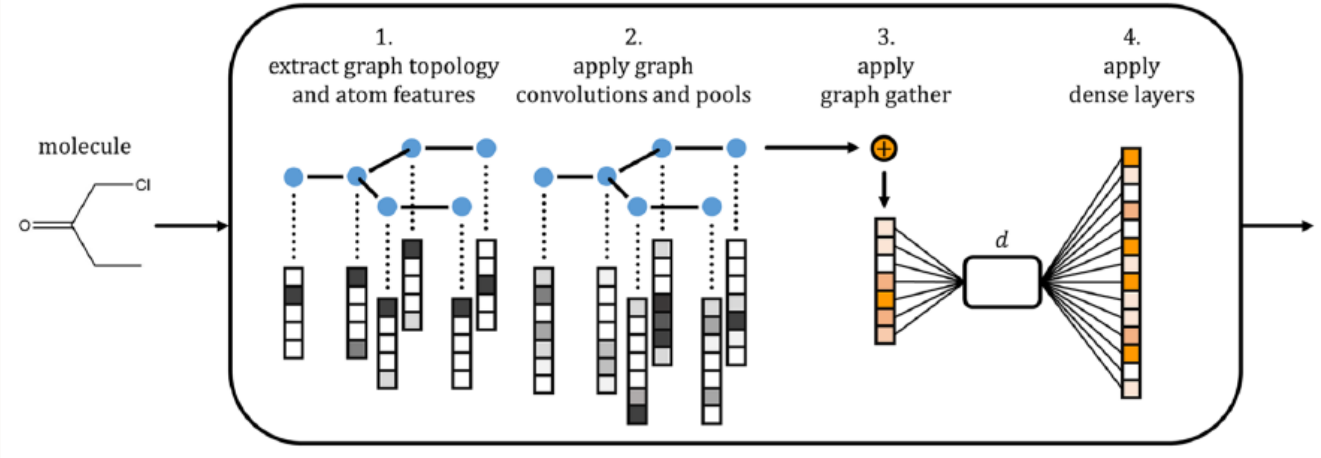
Tox21 dataset

covering 12 targets that were associated with human toxicity and contains a total of 8192 compounds.

Molecular Representation Learning

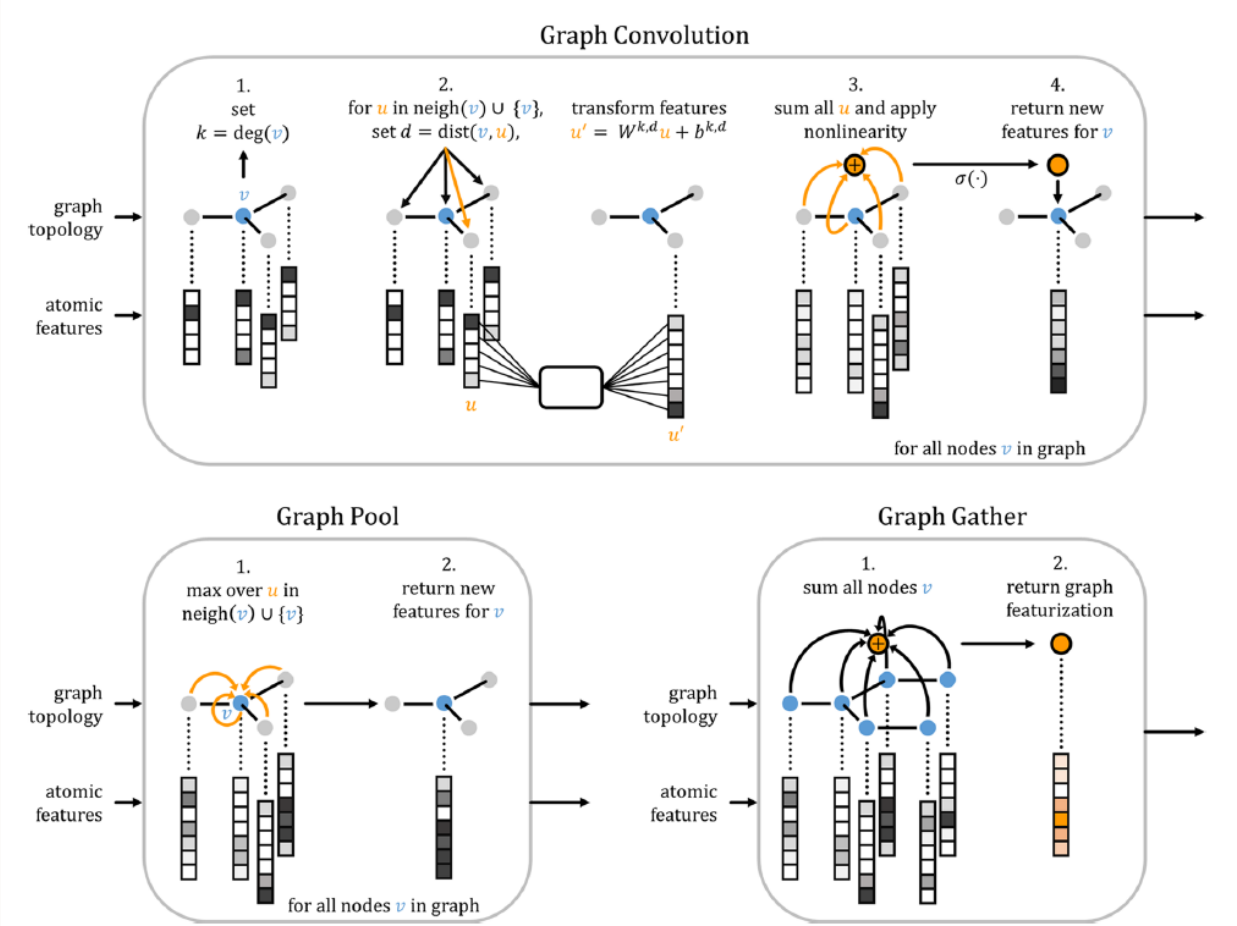
Graph Convolutional Network

- Graph Convolutional Network (GCN)



Molecular Representation Learning

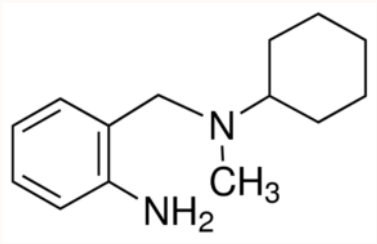
Graph Convolutional Network



QSAR Modeling

Traditional Approach

- QSAR

$$f\left(\ \text{  }\ \right) = molecule\ property$$

$drug\ molecule$

- Quantitative structure – activity relationship models (QSAR models) are regression or classification models that predict molecule property

QSAR Modeling

Deep Learning Approach

- Low Data Drug Discovery with One-shot Learning

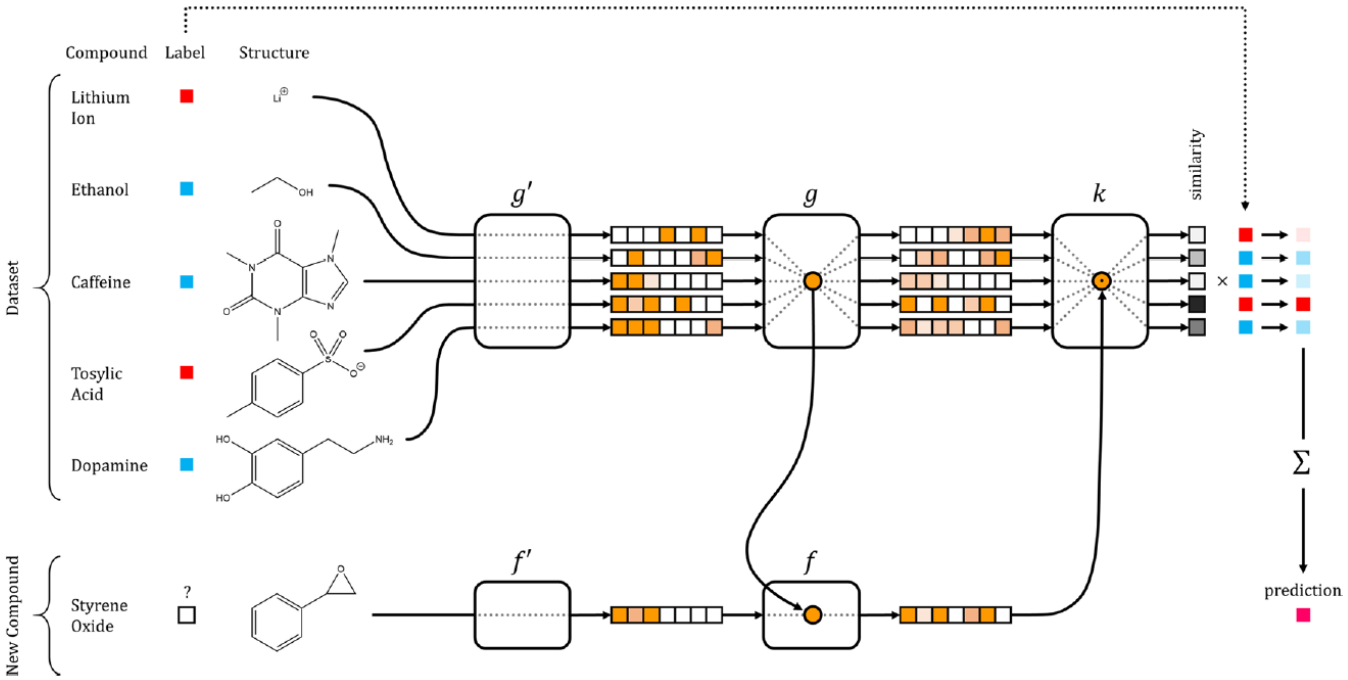


Figure 1. Schematic of Network Architecture for one-shot learning in drug discovery.

QSAR Modeling

Deep Learning Approach

- Low Data Drug Discovery with One-shot Learning

	Initialize	
$\mathbf{r} = g'(S)$	$\delta\mathbf{z} = \mathbf{0}$	$\delta z = 0$
	Repeat L times	
$e = k(f'(x) + \delta z, \mathbf{r})$	$\mathbf{e} = k(\mathbf{r} + \delta\mathbf{z}, g'(S))$	(similarity measures)
$a_j = e_j / \sum_{j=1}^{m} e_{ij}$	$\mathbf{A}_{ij} = \mathbf{e}_{ij} / \sum_{j=1}^{m} \mathbf{e}_{ij}$	(attention mechanism)
$r = a^{\mathrm{T}}\mathbf{r}$	$\mathbf{r} = \mathbf{A}g'(\mathbf{S})$	(expected feature map)
$\delta z = \mathrm{LSTM}([\delta z, r])$	$\delta\mathbf{z} = \mathrm{LSTM}([\delta\mathbf{z}, \mathbf{r}])$	(generate updates)
	Return	
$f(x) = f'(x) + \delta z$	$g(\mathbf{S}) = g'(\mathbf{S}) + \delta\mathbf{z}$	(evolve embeddings)

QSAR Modeling

Deep Learning Approach

- Low Data Drug Discovery with One-shot Learning

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**

[a]Numbers reported are means and standard deviations. Randomness is over the choice of support set; experiment is repeated with 20 support sets. The Appendix contains results for all held-out Tox21 tasks. The result with highest mean in each row is highlighted. The notation 10+/10− indicates supports with 10 positive examples and 10 negative examples.

Reference

- Jaeger, Sabrina, Simone Fulle, and Samo Turk. "Mol2vec: unsupervised machine learning approach with chemical intuition." Journal of chemical information and modeling 58.1 (2018): 27-35.
- Altae-Tran, Han, et al. "Low data drug discovery with one-shot learning." ACS central science 3.4 (2017): 283-293.
- Duvenaud, David K., et al. "Convolutional networks on graphs for learning molecular fingerprints." Advances in neural information processing systems. 2015.
- https://slideplayer.com/slide/5959740/
- http://www.secmem.org/blog/2019/08/17/gnn/
- https://www.slideshare.net/databricks/assessing-drug-safety-using-ai
- https://mc.ai/machine-learning-for-drug-discovery-in-a-nutshell%E2%80%8A-%E2%80%8Apart-ii/

Thank you!