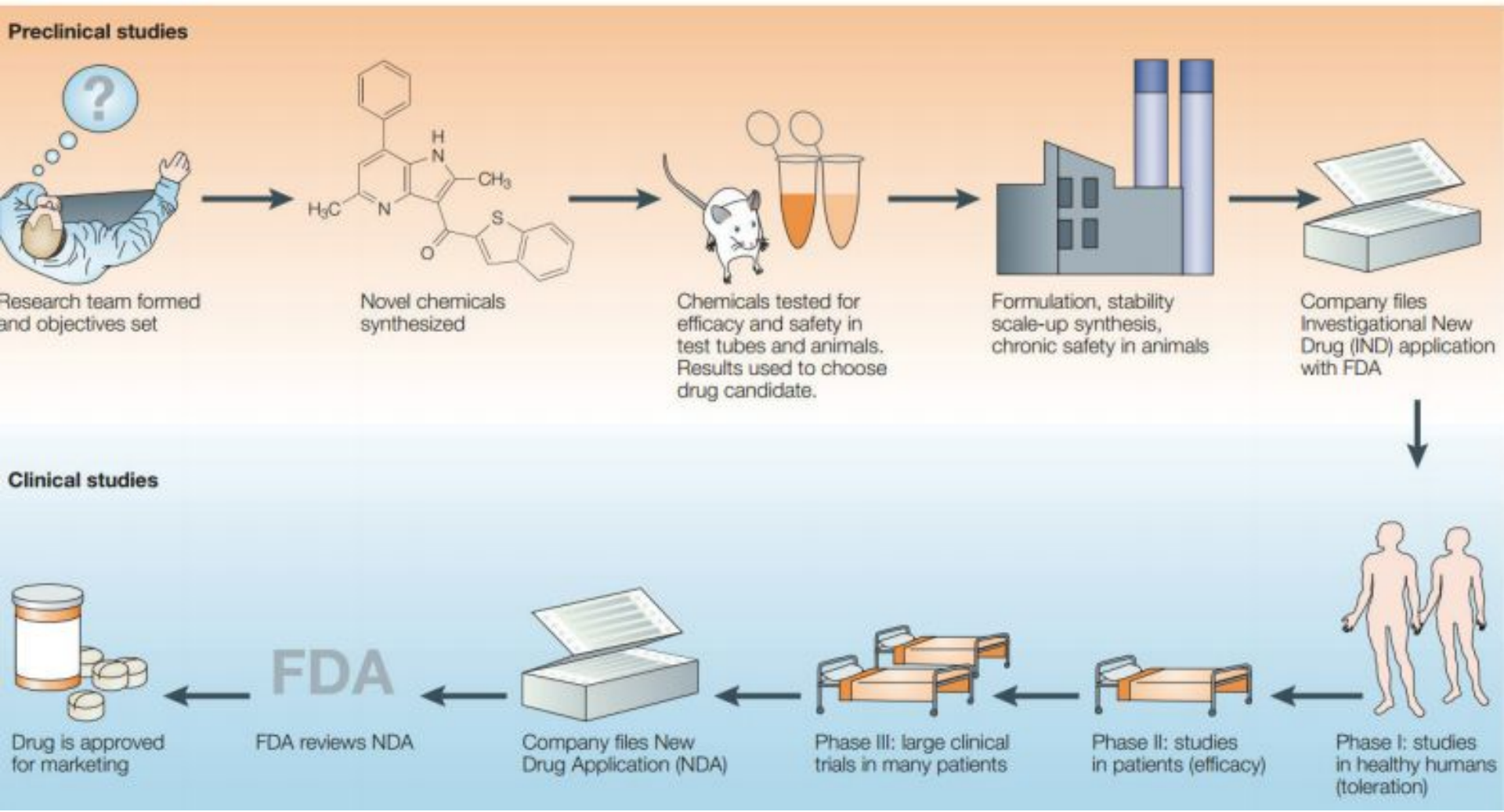
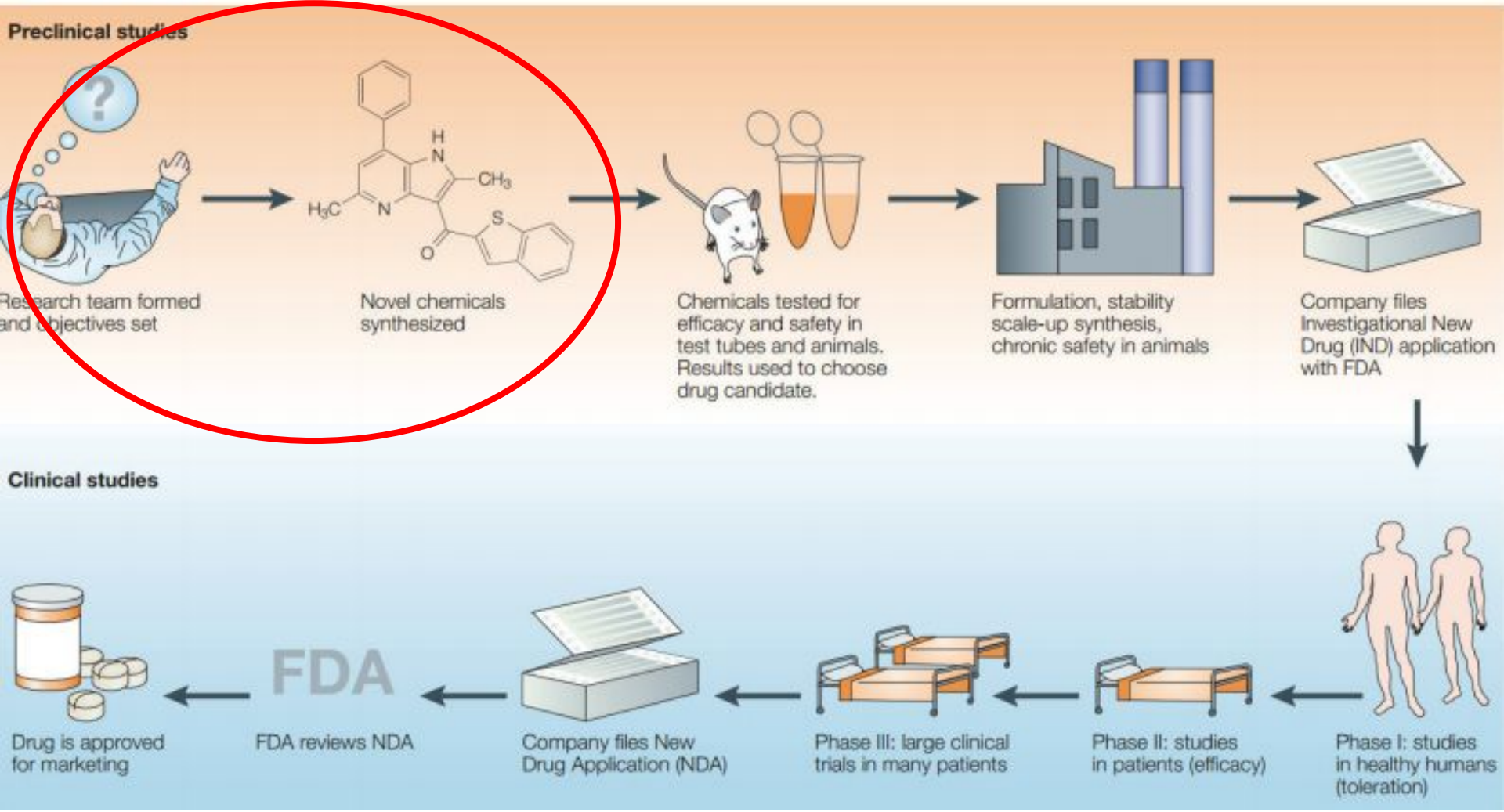
Drug Discovery - **Properties prediction**

Fanyun, 2019.04.02

What and Why?

- Costly
  - Costs about $2.6bn to develop a new drug on average
  - It requires 12-15 years of R&D from start to market
- Challenging - High failure rate
  - 97% of drug programmes fail
  - Only less that 40% of known diseases are currently treatable


Preclinical studies
Research team formed and objectives set
Novel chemicals synthesized
Chemicals tested for efficacy and safety in test tubes and animals. Results used to choose drug candidate.
Formulation, stability scale-up synthesis, chronic safety in animals
Company files Investigational New Drug (IND) application with FDA
Clinical studies
Phase I: studies in healthy humans (toleration)
Phase II: studies in patients (efficacy)
Phase III: large clinical trials in many patients
Company files New Drug Application (NDA)
FDA
FDA reviews NDA
Drug is approved for marketing

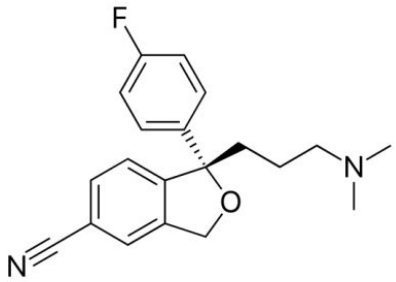

Preclinical studies
Research team formed and objectives set
Novel chemicals synthesized
Chemicals tested for efficacy and safety in test tubes and animals. Results used to choose drug candidate.
Formulation, stability scale-up synthesis, chronic safety in animals
Company files Investigational New Drug (IND) application with FDA
Clinical studies
Phase I: studies in healthy humans (toleration)
Phase II: studies in patients (efficacy)
Phase III: large clinical trials in many patients
Company files New Drug Application (NDA)
FDA
FDA reviews NDA
Drug is approved for marketing

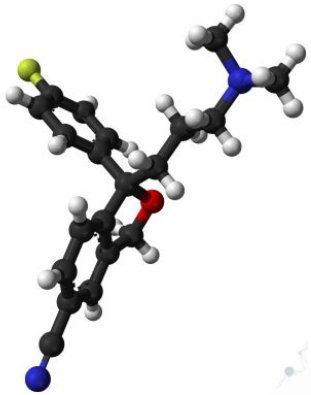
106 STARTUPS TRANSFORMING HEALTHCARE WITH AI

ML tasks for Drug Design

- Property prediction
  - Search space too large → Improve decision-making
- Generative chemistry
  - Given a set of goals, generate one or more new molecule that best optimize those goals
- Side-Effect predictions

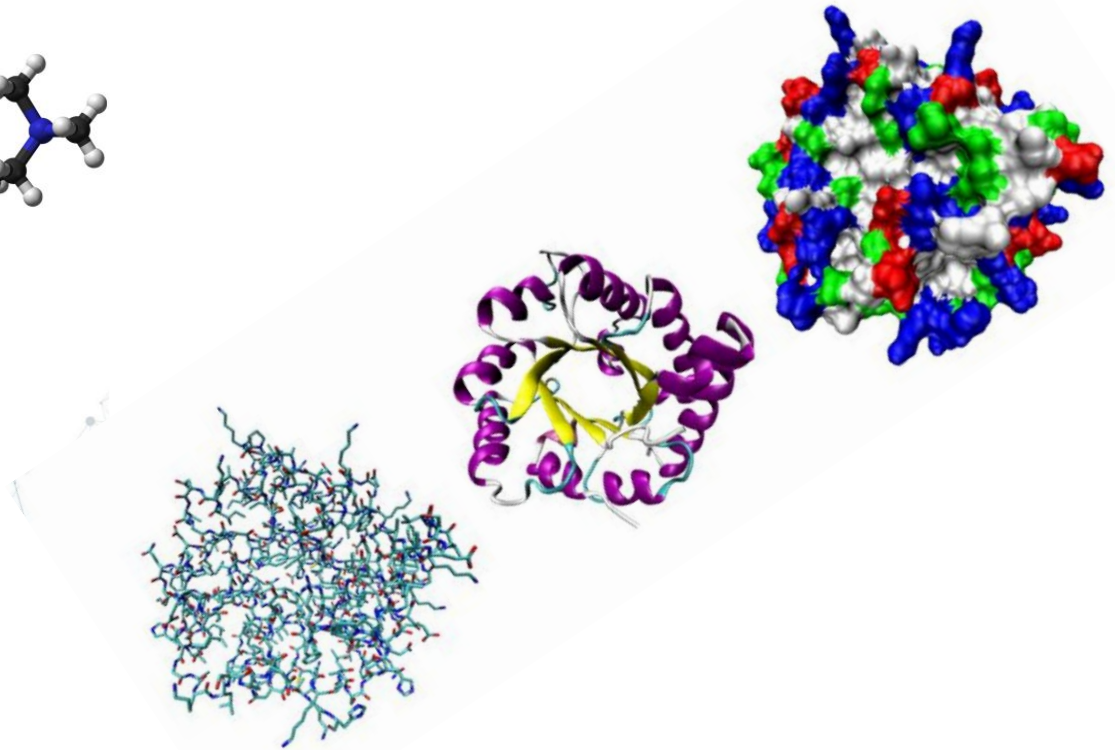
Drugs are molecules





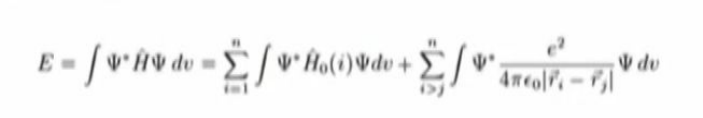
https://en.wikipedia.org/wiki/Escitalopram

Proteins are biomolecules



Density Functional Theory (DFT)

- Numerical method for approximating many properties of a molecule
- 1998 nobel prize in Chemistry, 2 of the top 10 most cited paper on google scholar(> 70000 citations
- Good balance of speed and accuracy (compared to other methods



$$E = \int \Psi^* \hat{H} \Psi \, dv = \sum_{i=1}^{n} \int \Psi^* \hat{H}_0(i) \Psi dv + \sum_{i>j}^{n} \int \Psi^* \frac{e^2}{4\pi\epsilon_0 |\vec{r}_i - \vec{r}_j|} \Psi \, dv$$

- **Still too slow for large searches** (~ 1 hour for molecules with ~ 20 atoms)

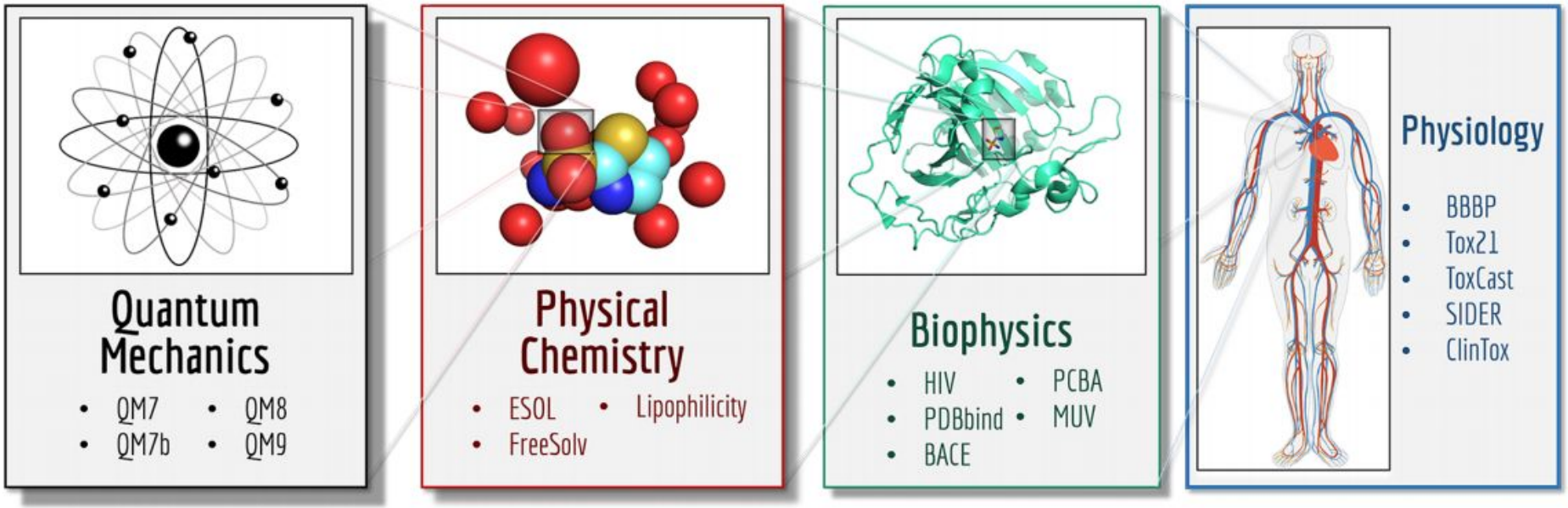


Figure 2: Tasks in different datasets focus on different levels of properties of molecules.

Milestones

- Extended-Connectivity Fingerprints (ECFP) - 2010
- Convolutional Networks on Graphs for Learning Molecular Fingerprints (Neural graph fingerprints) - NIPS 2015
- Neural Message Passing for Quantum Chemistry - ICML 2017
- MoleculeNet: A Benchmark for Molecular Machine Learning - Chemical Science Journal 2018
- **How Powerful are Graph Neural Networks?** - ICLR 2019

Milestones

- Extended-Connectivity Fingerprints (ECFP) - 2010
- Convolutional Networks on Graphs for Learning Molecular Fingerprints (Neural graph fingerprints) - NIPS 2015
- Neural Message Passing for Quantum Chemistry - ICML 2017
- MoleculeNet: A Benchmark for Molecular Machine Learning - Chemical Science Journal 2018
- **How Powerful are Graph Neural Networks?** - ICLR 2019

Extended-Connectivity Fingerprints (ECFP) - 2010

Convolutional Networks on Graphs for Learning Molecular Fingerprints (Neural graph fingerprints) - NIPS 2015

Algorithm 1 Circular fingerprints

Input: molecule, radius R, fingerprint length S
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R ▷ for each layer
  for each atom a in molecule
    $\mathbf{r}_1 \dots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow [\mathbf{r}_a, \mathbf{r}_1, \dots, \mathbf{r}_N]$ ▷ 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 $\mathbf{f}$

Algorithm 2 Neural graph fingerprints

Input: molecule, radius R, hidden weights $H_1^1 \dots H_R^5$, output weights $W_1 \dots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R ▷ for each layer
  for each atom a in molecule
    $\mathbf{r}_1 \dots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^{N} \mathbf{r}_i$ ▷ 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 $\mathbf{f}$

Milestones

- Extended-Connectivity Fingerprints (ECFP) - 2010
- Convolutional Networks on Graphs for Learning Molecular Fingerprints (Neural graph fingerprints) - NIPS 2015
- Neural Message Passing for Quantum Chemistry - ICML 2017
- MoleculeNet: A Benchmark for Molecular Machine Learning - Chemical Science Journal 2018
- **How Powerful are Graph Neural Networks?** - ICLR 2019

MoleculeNet: A Benchmark for Molecular Machine Learning

Aims to be ImageNet of molecular ML

Table 1: Dataset Details: number of compounds and tasks, recommended splits and metrics

Category	Dataset	Data Type	# Tasks	Task Type	# Compounds	Rec - Split	Rec - Metric
Quantum Mechanics	QM7	SMILES, 3D coordinates	1	Regression	7160	Stratified	MAE
	QM7b	3D coordinates	14	Regression	7210	Random	MAE
	QM8	SMILES, 3D coordinates	12	Regression	21786	Random	MAE
	QM9	SMILES, 3D coordinates	12	Regression	133885	Random	MAE
Physical Chemistry	ESOL	SMILES	1	Regression	1128	Random	RMSE
	FreeSolv	SMILES	1	Regression	642	Random	RMSE
	Lipophilicity	SMILES	1	Regression	4200	Random	RMSE
Biophysics	PCBA	SMILES	128	Classification	437929	Random	PRC-AUC
	MUV	SMILES	17	Classification	93087	Random	PRC-AUC
	HIV	SMILES	1	Classification	41127	Scaffold	ROC-AUC
	PDBbind	SMILES, 3D coordinates	1	Regression	11908	Time	RMSE
	BACE	SMILES	1	Classification	1513	Scaffold	ROC-AUC
Physiology	BBBP	SMILES	1	Classification	2039	Scaffold	ROC-AUC
	Tox21	SMILES	12	Classification	7831	Random	ROC-AUC
	ToxCast	SMILES	617	Classification	8575	Random	ROC-AUC
	SIDER	SMILES	27	Classification	1427	Random	ROC-AUC
	ClinTox	SMILES	2	Classification	1478	Random	ROC-AUC

Feature Engineering

- ECFP
- Coulomb Matrix
- Grid Featurizer
- Symmetry Function
- Graph Convolutions
- Weave

Models

- Conventional Models
  - Logistic Regression
  - SVM
  - Kernel Ridge Regression
  - Random Forests
  - Gradient Boosting
- Graph-based Models
  - Graph Convolution Models
  - Weave models
  - Directed Acyclic Graph models
  - Deep Tensor Neural Networks
  - ANI-1
  - Message Passing Neural Networks

Table 3: Summary of performances(test subset): conventional methods versus graph-based methods. Graph-based models outperform conventional methods on 11/17 datasets.

Category	Dataset	Metric	Best performances - conventional methods	Best performances - graph-based methods
Quantum Mechanics	QM7	MAE	KRR(CM): 10.22	**ANI-1: 2.86**
	QM7b	MAE	KRR(CM): 1.05	**DTNN: 1.77***
	QM8	MAE	Multitask: 0.0150	**MPNN: 0.0143**
	QM9	MAE	Multitask(CM): 4.35	**DTNN: 2.35**
Physical Chemistry	ESOL	RMSE	XGBoost: 0.99	**MPNN: 0.58**
	FreeSolv	RMSE	XGBoost: 1.74	**MPNN: 1.15**
	Lipophilicity	RMSE	XGBoost: 0.799	**GC: 0.655**
Biophysics	PCBA	AUC-PRC	Logreg: 0.129	**GC: 0.136**
	MUV	AUC-PRC	**Multitask: 0.184**	Weave: 0.109
	HIV	AUC-ROC	**KernelSVM: 0.792**	GC: 0.763
	BACE	AUC-ROC	**RF: 0.867**	Weave: 0.806
	PDBbind(full)	RMSE	**RF(grid): 1.25**	GC: 1.44
Physiology	BBBP	AUC-ROC	**KernelSVM: 0.729**	GC: 0.690
	Tox21	AUC-ROC	KernelSVM: 0.822	**GC: 0.829**
	ToxCast	AUC-ROC	Multitask: 0.702	**Weave: 0.742**
	SIDER	AUC-ROC	**RF: 0.684**	GC: 0.638
	ClinTox	AUC-ROC	Bypass: 0.827	**Weave: 0.832**

* As discussed in section 4.4, DTNN outperforms KRR(CM) on 14/16 tasks in QM7b while the mean-MAE is skewed due to different magnitudes of labels.

Quick Review - Neural Message Passing for Quantum Chemistry

Milestones

- Extended-Connectivity Fingerprints (ECFP) - 2010
- Convolutional Networks on Graphs for Learning Molecular Fingerprints (Neural graph fingerprints) - NIPS 2015
- Neural Message Passing for Quantum Chemistry - ICML 2017
- MoleculeNet: A Benchmark for Molecular Machine Learning - Chemical Science Journal 2018
- **How Powerful are Graph Neural Networks?** - ICLR 2019

How Powerful are Graph Neural Networks? (ICLR 2019)

*Keyulu Xu**, *Weihua Hu**, *Jure Leskovec*, *Stefanie Jegelka*

- Provide a theoretical framework for analyzing the expressive power of GNNs
  - GNNs are at most as powerful as the WL test in distinguishing graph structures
- Develop a simple architecture - **Graph Isomorphism Network (GIN)** and show that its discriminative/representational power is equal to WL test

Theoretical framework

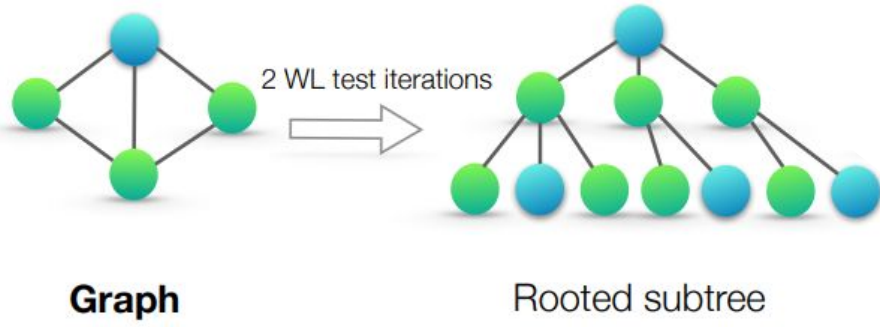
- Assign each feature vector a unique label {a, b, c}
- Feature vectors of a set of neighbors form a ***multiset***
- ***GNN*** →
  - ***AGGREGATE neighbor features***
  - ***COMBINE with self feature***
  - ***READOUT (permutation invariant)***

WL test

For all nodes v_i

- Obtain multiset of nodes and their neighbor nodes' features
  - **ex: {b},{g, g, g} (1-st iteration)**
- Update node feature with *a*n **injective** hash function
  - **ex: v_i = hash({b},{g, g, g}) (1-st iteration)**

Repeat k-steps or until convergence



GNNs are at most as powerful as the WL test in distinguishing graph structures

WL-TEST injective → always re-labels different multisets of neighboring nodes into different new labels !

Proof:

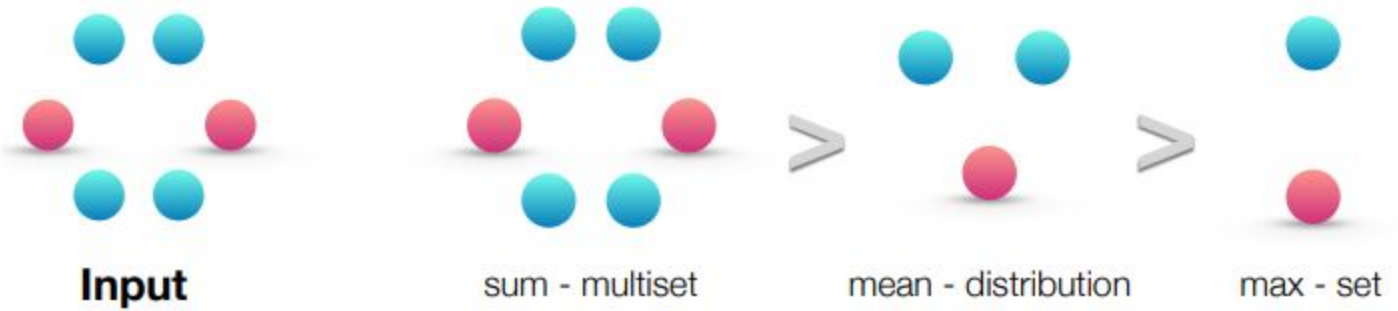
1. Suppose a GNN can decide G_1 and G_2 are non-isomorphic but not WL-test
2. If $\boldsymbol{WL(G_1) = WL(G_2)}$ → it follows that the multiset of nodes are the same → GNN must have same output on both graphs
3. Contradiction!

→ a GNN is as powerful as WL-test when AGGREGATE,COMBINE,READOUT are **injective**

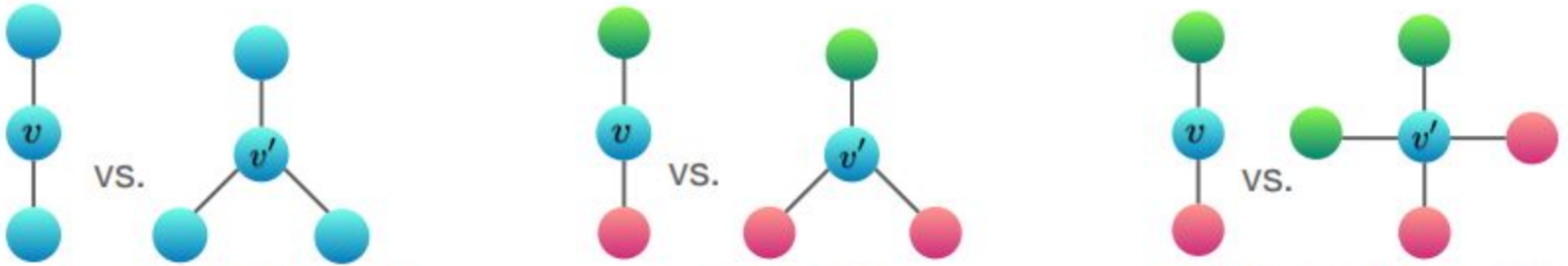
How to model injective multiset functions?

- AGGREGATOR + COMBINE: ?
- READOUT: ?

Ranking AGGREGATOR's expressive power



- Mean learns distributions
- Max learns sets with distinct elements



(a) Mean and Max both fail (b) Max fails (c) Mean and Max both fail

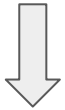
Figure 3: **Examples of graph structures that mean and max aggregators fail to distinguish.**

- Mean learns distributions
- Max learns sets with distinct elements

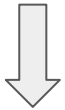
How to model injective multiset functions?

- AGGREGATOR + COMBINE: ?
  - AGGREGATOR: function on **multiset**
  - AGGREGATOR+COMBINE: function on **(element, multiset)** pair
- READOUT: sum

Lemma 5. *Assume $\mathcal{X}$ is countable. There exists a function $f : \mathcal{X} \to \mathbb{R}^n$ so that $h(X) = \sum_{x \in X} f(x)$ is unique for each multiset $X \subset \mathcal{X}$ of bounded size. Moreover, any multiset function g can be decomposed as $g(X) = \phi\left(\sum_{x \in X} f(x)\right)$ for some function ϕ.*



Corollary 6. *Assume $\mathcal{X}$ is countable. There exists a function $f : \mathcal{X} \to \mathbb{R}^n$ so that for infinitely many choices of ϵ, including all irrational numbers, $h(c, X) = (1 + \epsilon) \cdot f(c) + \sum_{x \in X} f(x)$ is unique for each pair (c, X), where $c \in \mathcal{X}$ and $X \subset \mathcal{X}$ is a multiset of bounded size. Moreover, any function g over such pairs can be decomposed as $g(c, X) = \varphi\left((1 + \epsilon) \cdot f(c) + \sum_{x \in X} f(x)\right)$ for some function φ.*



$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1 + \epsilon^{(k)}\right) \cdot h_v^{(k-1)} + \sum\nolimits_{u \in \mathcal{N}(v)} h_u^{(k-1)}\right).$$

Lemma 5.

Lemma 5. *Assume $\mathcal{X}$ is countable. There exists a function $f : \mathcal{X} \to \mathbb{R}^n$ so that $h(X) = \sum_{x \in X} f(x)$ is unique for each multiset $X \subset \mathcal{X}$ of bounded size.*

Proof:

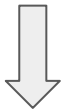
Bounded size → exist a number **N** bigger than the cardinality of all input multiset

Z: mapping of element to natural number [0, N-1]

$$f(x) = N^{-Z(x)} \quad \Longrightarrow \quad h(X) = a_1 * N^1 + a_2 * N^2 + \ldots + a_n * N^n$$

Corollary 6.

Corollary 6. *Assume $\mathcal{X}$ is countable. There exists a function $f : \mathcal{X} \to \mathbb{R}^n$ so that for infinitely many choices of ϵ, including all irrational numbers, $h(c, X) = (1+\epsilon) \cdot f(c) + \sum_{x \in X} f(x)$ is unique for each pair (c, X), where $c \in \mathcal{X}$ and $X \subset \mathcal{X}$ is a multiset of bounded size. Moreover, any function g over such pairs can be decomposed as $g(c, X) = \varphi\left((1+\epsilon) \cdot f(c) + \sum_{x \in X} f(x)\right)$ for some function φ.*



$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right) \cdot h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)}\right).$$

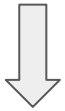
Graph Isomorphism Network **(GIN)**

- AGGREGATOR + COMBINE $\longrightarrow$ $h_v^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right)\cdot h_v^{(k-1)} + \sum_{u\in\mathcal{N}(v)} h_u^{(k-1)}\right).$
- READOUT: **sum**
  - Use **skip connection** to incorporate information from different depth/iterations

$$h_G = \mathrm{CONCAT}\Big(\mathrm{READOUT}\Big(\Big\{h_v^{(k)} | v \in G\Big\}\Big) \mid k = 0, 1, \ldots, K\Big).$$

1-layer MLP is not enough

Corollary 6. *Assume $\mathcal{X}$ is countable. There exists a function $f : \mathcal{X} \to \mathbb{R}^n$ so that for infinitely many choices of ϵ, including all irrational numbers, $h(c, X) = (1+\epsilon) \cdot f(c) + \sum_{x \in X} f(x)$ is unique for each pair (c, X), where $c \in \mathcal{X}$ and $X \subset \mathcal{X}$ is a multiset of bounded size. Moreover, any function g over such pairs can be decomposed as $g(c, X) = \varphi\left((1+\epsilon) \cdot f(c) + \sum_{x \in X} f(x)\right)$ for some function φ.*



$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right) \cdot h_v^{(k-1)} + \sum\nolimits_{u \in \mathcal{N}(v)} h_u^{(k-1)}\right).$$

1-layer MLP is not enough

Lemma 7. *There exist finite multisets* $X_1 \neq X_2$ *so that for any linear mapping* W, $\sum_{x \in X_1} \mathrm{ReLU}(Wx) = \sum_{x \in X_2} \mathrm{ReLU}(Wx)$.

- For example, 1-layer MLP may not be able to distinguish {1,1,1,1,1} and {2,3} due to the linearity

Graph Isomorphism Network **(GIN)**

- AGGREGATOR: **sum**
- COMBINE: $\longrightarrow$>1 layer MLPs $h_v^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right)\cdot h_v^{(k-1)} + \sum_{u\in\mathcal{N}(v)} h_u^{(k-1)}\right).$
- READOUT: **sum**
  - Use **skip connection** to incorporate information from different depth/iterations

$$h_G = \mathrm{CONCAT}\Big(\mathrm{READOUT}\Big(\Big\{h_v^{(k)} | v \in G\Big\}\Big) \mid k = 0, 1, \ldots, K\Big).$$

Experiments

	Datasets	IMDB-B	IMDB-M	RDT-B	RDT-M5K	COLLAB	MUTAG	PROTEINS	PTC	NCI1
Datasets	# graphs	1000	1500	2000	5000	5000	188	1113	344	4110
	# classes	2	3	2	5	3	2	2	2	2
	Avg # nodes	19.8	13.0	429.6	508.5	74.5	17.9	39.1	25.5	29.8
Baselines	WL subtree	73.8 ± 3.9	50.9 ± 3.8	81.0 ± 3.1	52.5 ± 2.1	78.9 ± 1.9	90.4 ± 5.7	75.0 ± 3.1	59.9 ± 4.3	$\mathbf{86.0 \pm 1.8}$ *
	DCNN	49.1	33.5	–	–	52.1	67.0	61.3	56.6	62.6
	PATCHYSAN	71.0 ± 2.2	45.2 ± 2.8	86.3 ± 1.6	49.1 ± 0.7	72.6 ± 2.2	$\mathbf{92.6 \pm 4.2}$ *	75.9 ± 2.8	60.0 ± 4.8	78.6 ± 1.9
	DGCNN	70.0	47.8	–	–	73.7	85.8	75.5	58.6	74.4
	AWL	74.5 ± 5.9	51.5 ± 3.6	87.9 ± 2.5	54.7 ± 2.9	73.9 ± 1.9	87.9 ± 9.8	–	–	–
GNN variants	SUM–MLP (**GIN-0**)	$\mathbf{75.1 \pm 5.1}$	$\mathbf{52.3 \pm 2.8}$	$\mathbf{92.4 \pm 2.5}$	$\mathbf{57.5 \pm 1.5}$	$\mathbf{80.2 \pm 1.9}$	$\mathbf{89.4 \pm 5.6}$	$\mathbf{76.2 \pm 2.8}$	$\mathbf{64.6 \pm 7.0}$	$\mathbf{82.7 \pm 1.7}$
	SUM–MLP (**GIN-ϵ**)	$\mathbf{74.3 \pm 5.1}$	$\mathbf{52.1 \pm 3.6}$	$\mathbf{92.2 \pm 2.3}$	$\mathbf{57.0 \pm 1.7}$	$\mathbf{80.1 \pm 1.9}$	$\mathbf{89.0 \pm 6.0}$	$\mathbf{75.9 \pm 3.8}$	63.7 ± 8.2	$\mathbf{82.7 \pm 1.6}$
	SUM–1-LAYER	74.1 ± 5.0	$\mathbf{52.2 \pm 2.4}$	90.0 ± 2.7	55.1 ± 1.6	$\mathbf{80.6 \pm 1.9}$	$\mathbf{90.0 \pm 8.8}$	$\mathbf{76.2 \pm 2.6}$	63.1 ± 5.7	82.0 ± 1.5
	MEAN–MLP	73.7 ± 3.7	$\mathbf{52.3 \pm 3.1}$	50.0 ± 0.0	20.0 ± 0.0	79.2 ± 2.3	83.5 ± 6.3	75.5 ± 3.4	$\mathbf{66.6 \pm 6.9}$	80.9 ± 1.8
	MEAN–1-LAYER (GCN)	74.0 ± 3.4	51.9 ± 3.8	50.0 ± 0.0	20.0 ± 0.0	79.0 ± 1.8	85.6 ± 5.8	76.0 ± 3.2	64.2 ± 4.3	80.2 ± 2.0
	MAX–MLP	73.2 ± 5.8	51.1 ± 3.6	–	–	–	84.0 ± 6.1	76.0 ± 3.2	64.6 ± 10.2	77.8 ± 1.3
	MAX–1-LAYER (GraphSAGE)	72.3 ± 5.3	50.9 ± 2.2	–	–	–	85.1 ± 7.6	75.9 ± 3.2	63.9 ± 7.7	77.7 ± 1.5

Other settings

- Large Scale **Multitask** Learning
  - Addition of more tasks and data helps with generalisation of models
  - Shared latent representation may be informative and addition tasks act as regularization
  - However, for some tasks single-task model performs better as some tasks require customized feature learning layers
  - Reference: Massively Multitask Networks for Drug Discovery
- **One-Shot** Learning
  - Reference: Low Data Drug Discovery with One-shot Learning

All References

- Overview slides by BenevolenAI. http://www.ymer.org/papers/files/2017-London-ML-Meetup.pdf
- Deep Reinforcement Learning for Pre-Clinical Drug Development. http://compgen.illinois.edu/files/2017/10/Russell-CompGen-Lightning-talk-sept-2017.pdf
- ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity? https://hal.archives-ouvertes.fr/hal-01577696v1/document
- Opportunities and obstacles for deep learning in biology and medicine https://greenelab.github.io/deep-review/manuscript.pdf
- Machine learning for therapeutic research: http://www.cazencott.info/dotclear/public/lectures/2017-10-12-azencott.pdf
- Hongming Chen et al. The rise of deep learning in drug discovery. https://reader.elsevier.com/reader/sd/pii/S1359644617303598?token=334CE8BFDB1C5306DEB9F15AB5DA0CE980F5B55B439EE9DB549FB5C01C0AF9AF19275A174A5EC19088A007EE31ED5756
- AI for Drug Discovery Landscape Overview 2017. http://analytics.dkv.global/data/pdf/AIDrugDiscoveryLandscapeOverview2017.pdf
- Machine learning in chemoinformatics and drug discovery. https://reader.elsevier.com/reader/sd/pii/S1359644617304695?token=6BBC6CAE1D25534FB13054CEE5D8015A48AAC9B52E05677E23165F5606B10E7E5B593BD89726E01E7FD9115658E27C93
- Neural Message Passing for Quantum Chemistry (presentation) https://vimeo.com/238221016

All References (continued)

- Deep Graph Kernels. https://users.soe.ucsc.edu/~vishy/pubs/YanVis15.pdf
- Discriminative Embeddings of Latent Variable Models for Structured Data. https://arxiv.org/abs/1603.05629
- Neural Message Passing for Quantum Chemistry https://arxiv.org/abs/1704.01212
- Using Deep Reinforcement Learning to Generate Rationales for Molecules
- Hierarchical modeling of molecular energies using a deep neural network https://arxiv.org/pdf/1710.00017.pdf
- SchNet: A continuous-filter convolutional neural network for modeling quantum interactions https://arxiv.org/abs/1706.08566
- COVARIANT COMPOSITIONAL NETWORKS FOR LEARNING GRAPHS https://arxiv.org/pdf/1801.02144.pdf

Q&A