Workshop: Graph Neural Network Modeling in Drug Discovery Using PyTorch

Toronto Machine Learning Summit

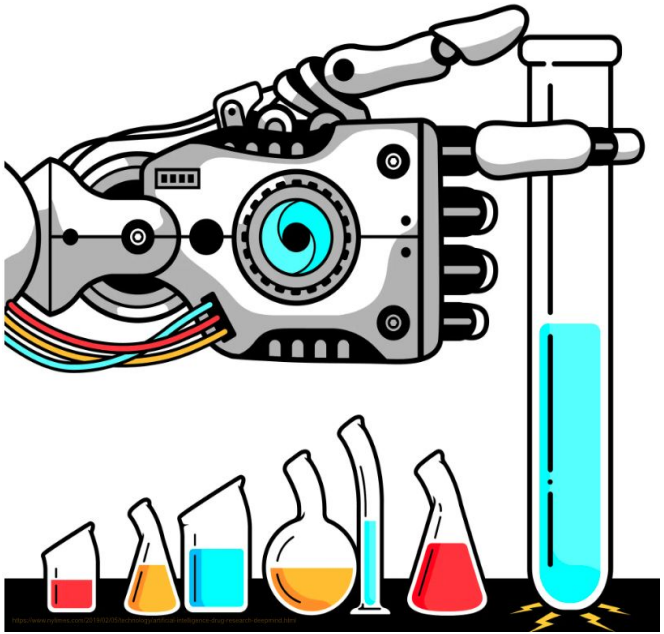
28.11.2022

Dr. Nasim Abdollahi [1,2] & Dr. Farnoosh Khodakarami [2]

1 University of Toronto

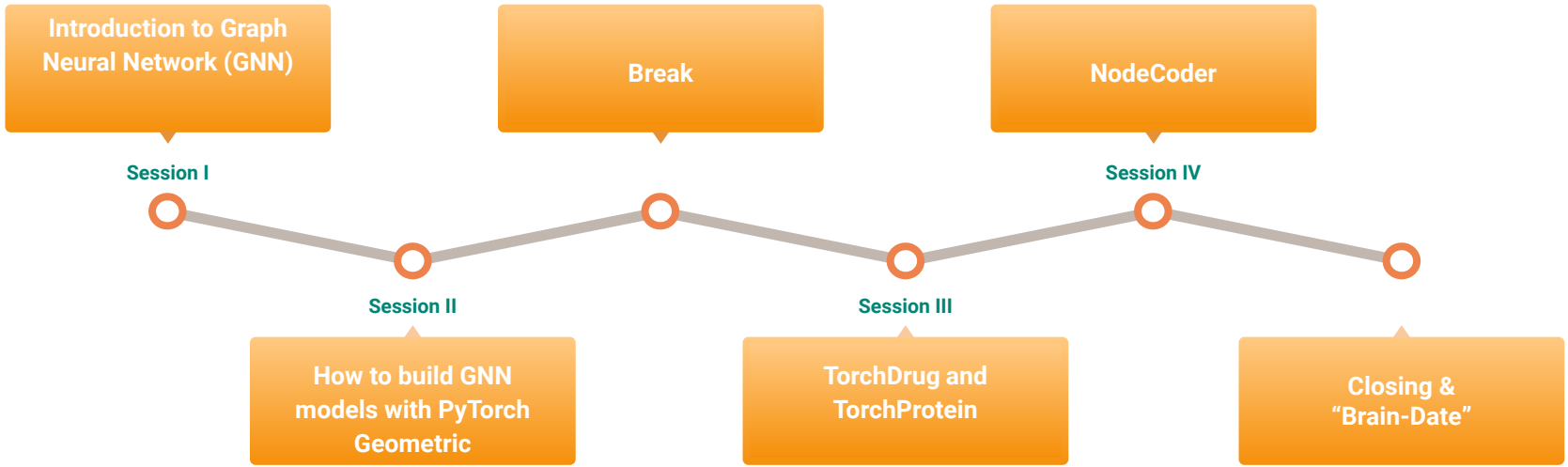
2 Cyclica

Goals



1. Graph Neural Network (GNN) in drug discovery
2. How to build GNN with PyTorch Geometric
3. TorchDrug - ML platform for drug discovery
4. TorchProtein - ML library for protein science
5. NodeCoder - Graph-based ML platform for predicting active sites of modeled proteins

Workshop Roadmap

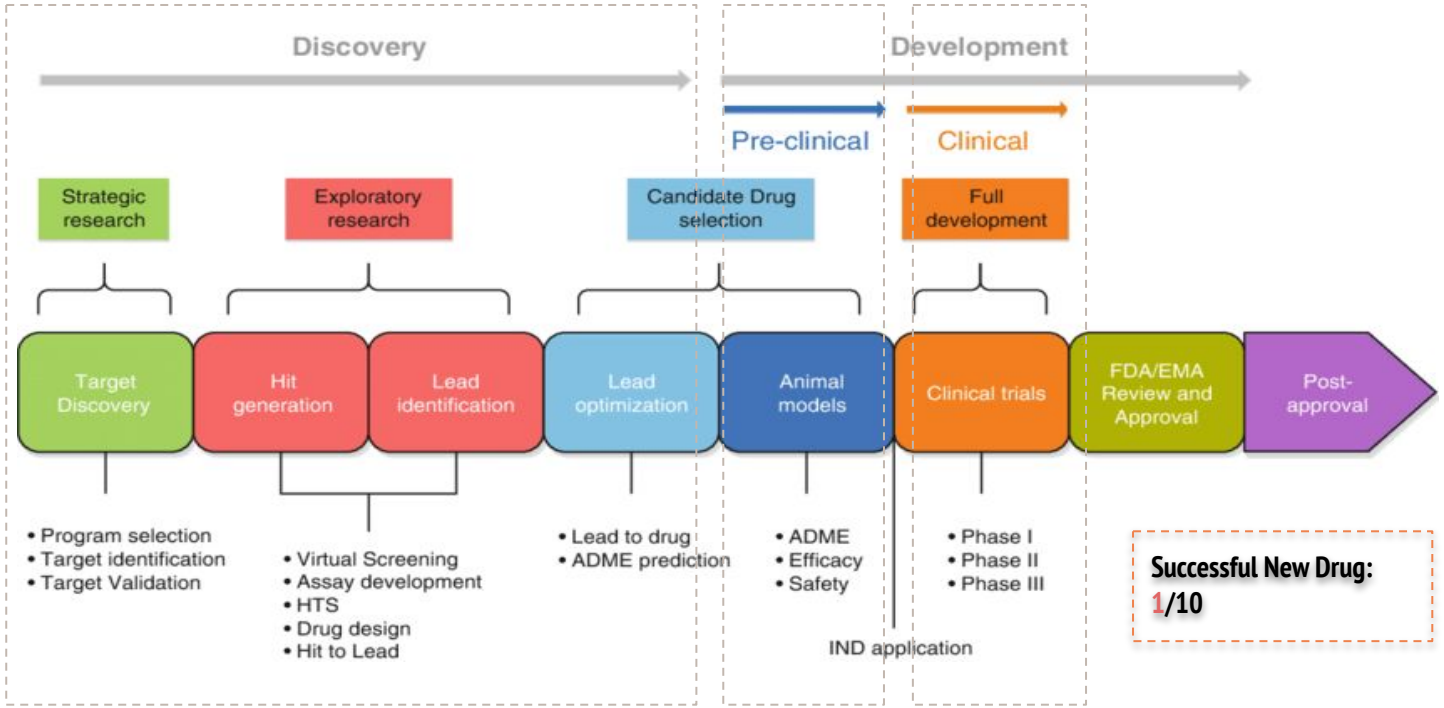


Section I:

Introduction to Graph Neural Network (GNN)

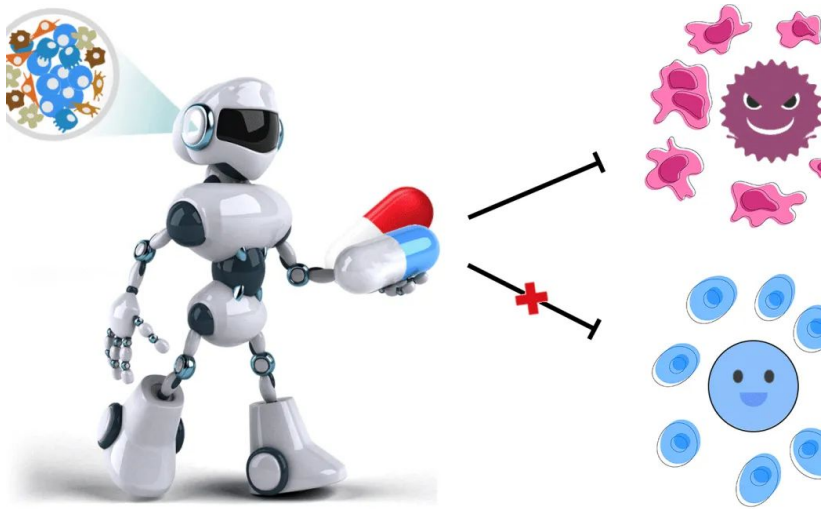
GNNs have shown big sparks in graph-based ML modeling for drug discovery and protein science!

AI and Drug Discovery - Why Computational Methods?



Medicinal Biotechnology for Disease Modeling, Clinical Therapy, and Drug Discovery and Development

Future Trends in Drug Design and Ultimate Goal



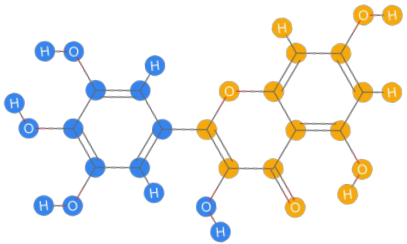
https://www.helsinki.fi/en/hilife-helsinki-institute-life-science/news/ai-guides-accurate-design-patient-tailored-combinatorial-strategies-aml-treatment

- Accelerating research process and reducing risk and expenditure in clinical trials
- Developing AI methods capable of gathering and analysing large amounts of data in a short time, to select appropriate targets and complimentary ligands
- **Ultimate Goal**: design and develop specific, non-toxic, effective and **patient-tailored drug** over short period of time (several hours!)
- Tech Giants making bets in the space!

Graph Neural Networks (GNNs) in Drug Discovery

- ML models used for drug discovery (based on literature):
  - Descriptor-based models
  - Graph-based models

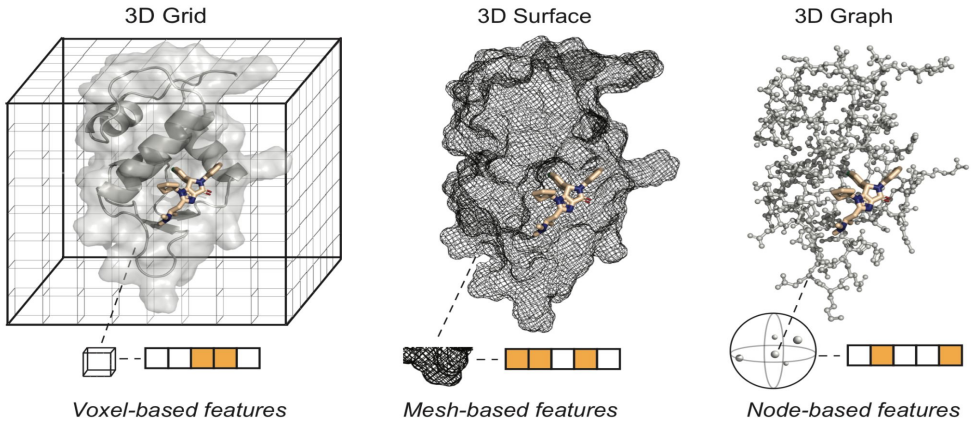
- As graph is a natural representation for **protein** and **molecule**, GNNs have shown big sparks in graph-based ML modeling for drug discovery and protein science



- Some popular applications in drug discovery and protein science:
  - Modeling topology of a protein structure
  - Predicting protein function
  - Predicting Protein–Protein Interaction
  - Predicting Drug–Target Interaction
  - Drug Response Prediction
  - Automated Drug Discovery

Structure-based Drug Design - De Novo Molecular Design and Property Prediction

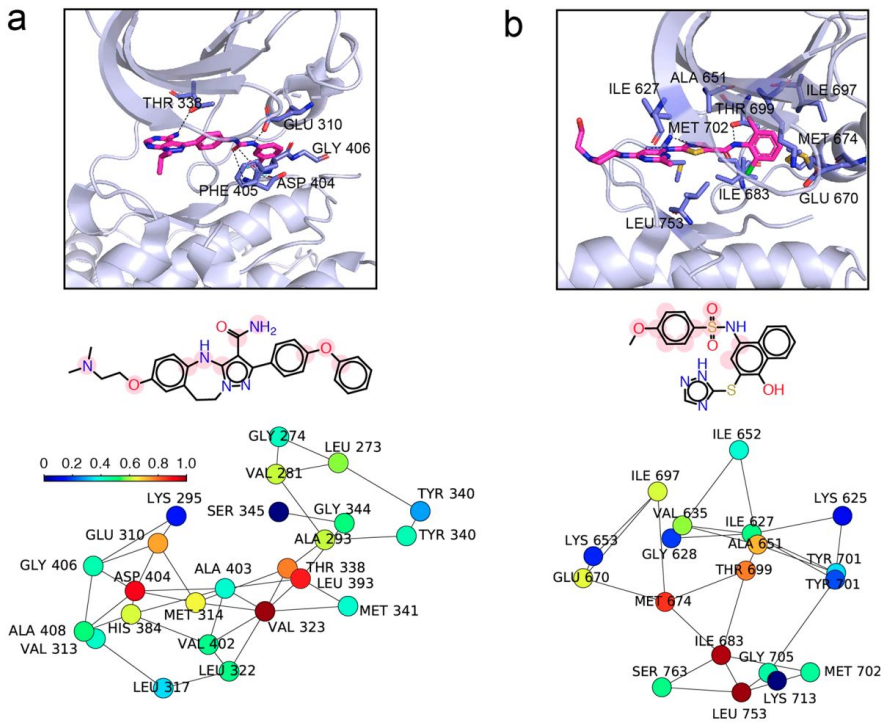
- Structure-based drug design - methods that leverage 3D structures of macromolecular targets, such as proteins or nucleic acids, to identify suitable ligands
- Geometric deep learning and different representation of macromolecular 3D structures: grid, surface, graph
- Graph - natural representation of protein 3D structure, graph-based neural networks have shown big potential for structure-based drug discovery and design, with emphasis on:
    - Molecular property prediction
    - De novo molecular design



https://arxiv.org/pdf/2210.11250.pdf

Virtual Screening for Prioritizing Compounds

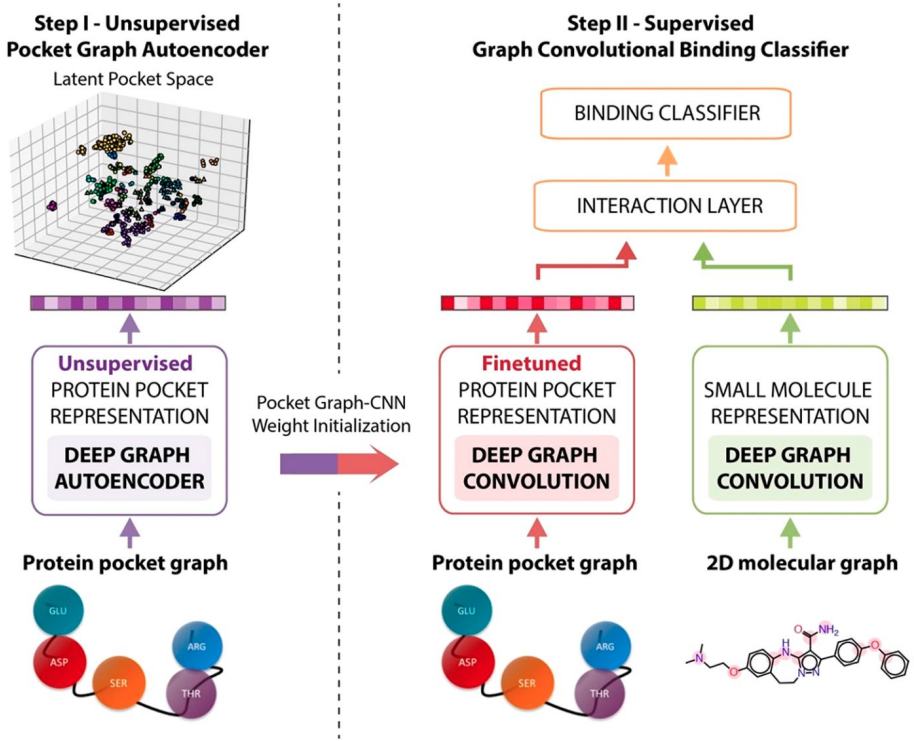
- For therapeutically relevant target (protein receptor or enzyme), prioritizing the most potent compounds is often the first step toward new drug development
- Experimental approach - time-consuming and labor intensive, cannot easily scale up to explore diverse chemical and structural space
- Computational methods for virtual screening - search libraries of small molecules to identify structures most likely to bind to a drug target, typically a protein receptor or enzyme



https://pubs.acs.org/doi/10.1021/acs.jcim.9b00628

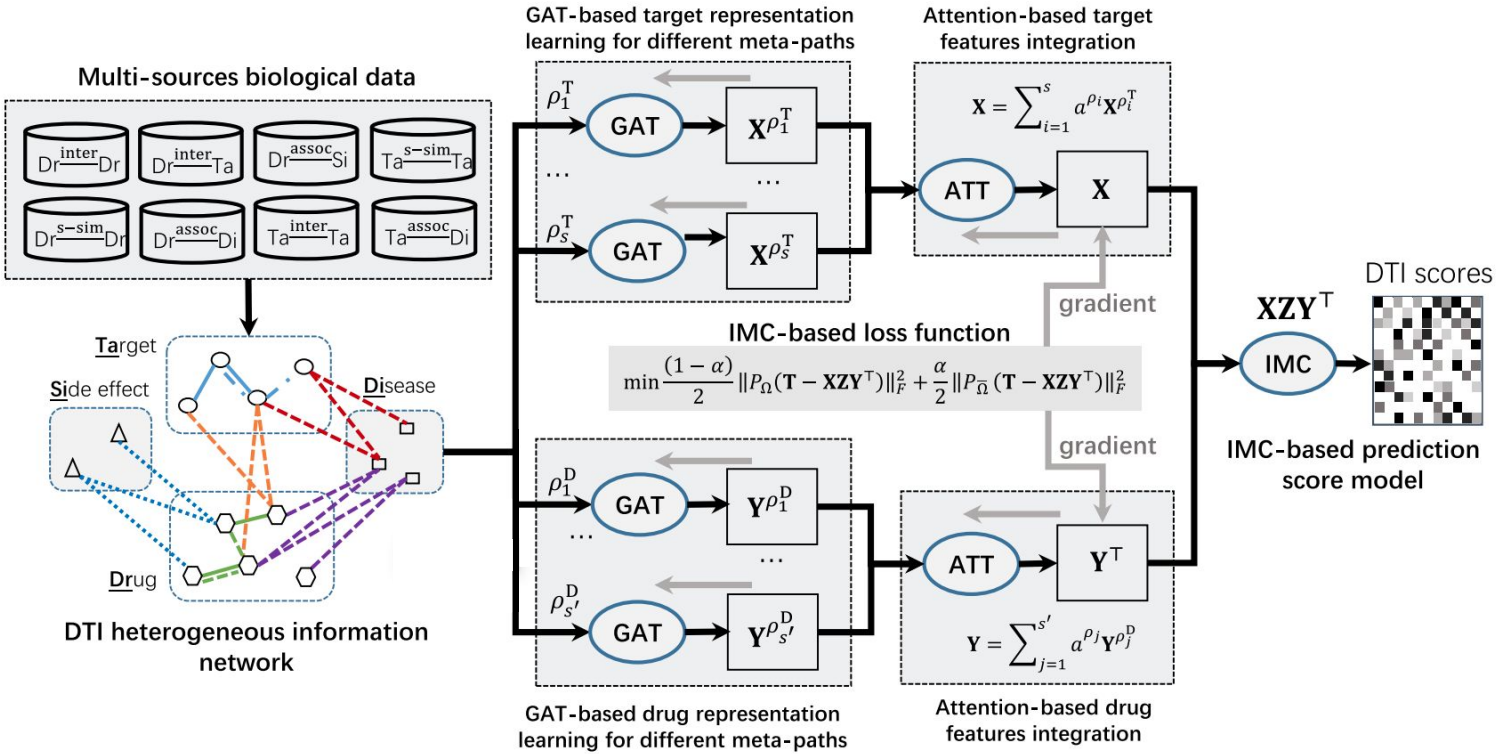
Virtual Screening for Ligand-Target Interactions Prediction

- For disease target to identify ligand binding sites
- For approved drugs to identifying unexpected off-targets - predicting and explaining observed side-effects and drug repurposing
- Most accurate method for determining target-ligand interactions, experimental characterization of target-ligand binding affinities
- Time-consuming and labor intensive
- Computational methods for virtual screening to predict binding affinities between target-ligand pairs



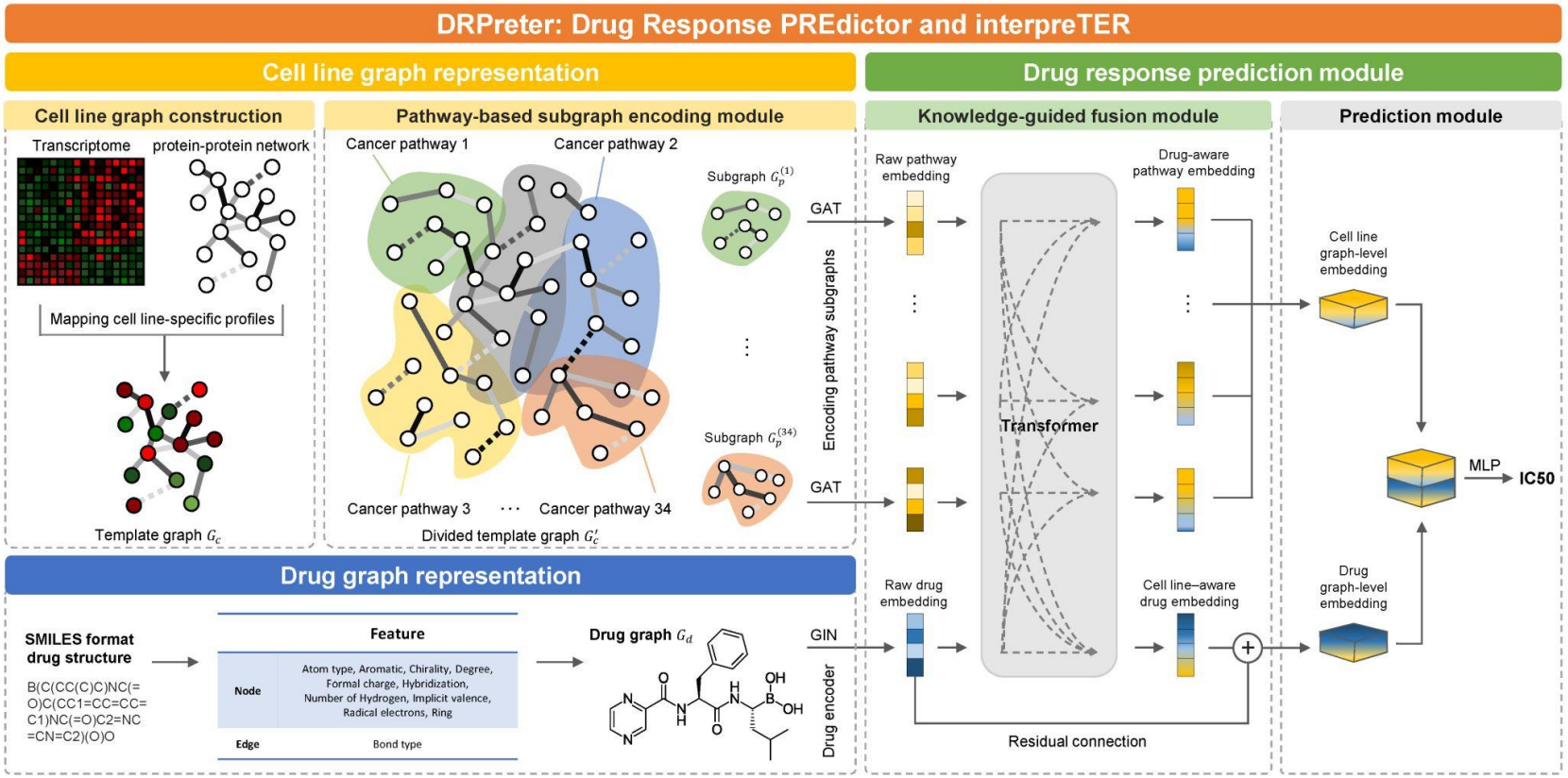
https://pubs.acs.org/doi/10.1021/acs.jcim.9b00628

Drug Repurposing and Discovery - Protein–Protein and Drug–Target Interaction Prediction



IMCHGAN: Inductive Matrix Completion With Heterogeneous Graph Attention Networks for Drug-Target Interactions Prediction

Drug Response Predictor and Interpreter - DRPreter

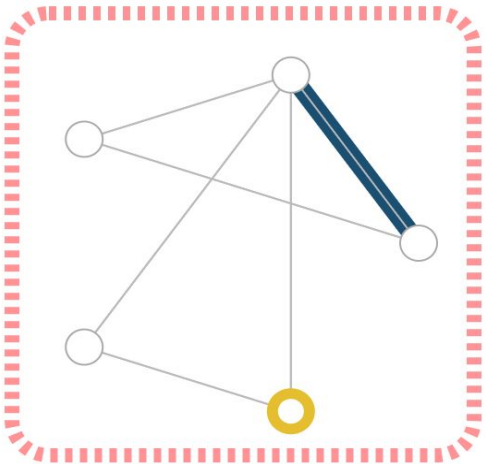


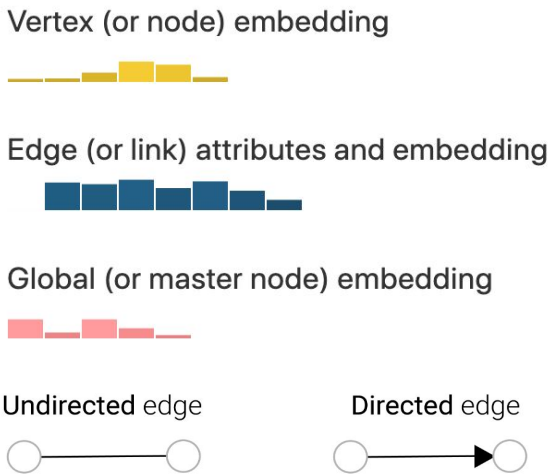
[DRPreter: Interpretable Anticancer Drug Response Prediction Using Knowledge-Guided Graph Neural Networks and Transformer]
[GitHub]

Graph Neural Network

A class of neural network for processing data represented by graph data structures!

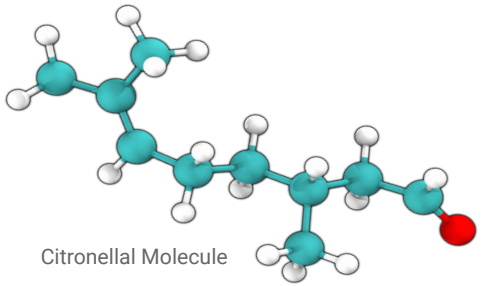
Graph Representation

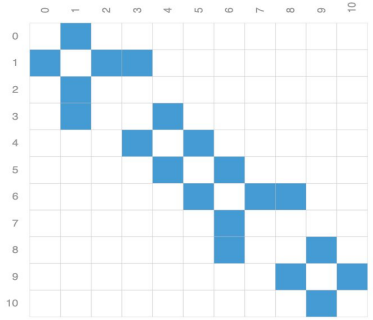


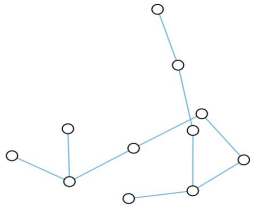


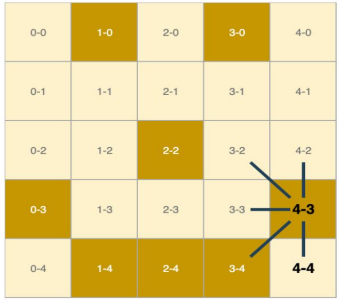
https://distill.pub/2021/gnn-intro/

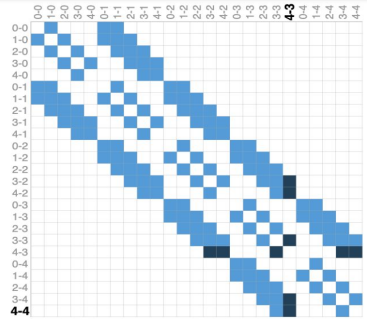
Graph Data

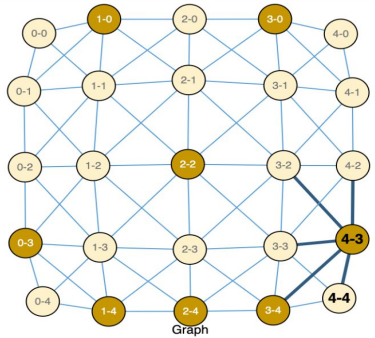


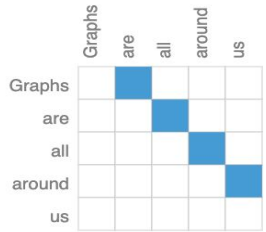












https://distill.pub/2021/gnn-intro/

Prediction Tasks on Graphs

General types of prediction tasks on graphs:

- **Graph-level**: predict a single property for a whole graph
- **Node-level**: predict some property for each node in a graph
- **Edge-level**: predict the property or presence of edges/links in a graph

All three levels of prediction problems can be solved with a single model class, the GNN!

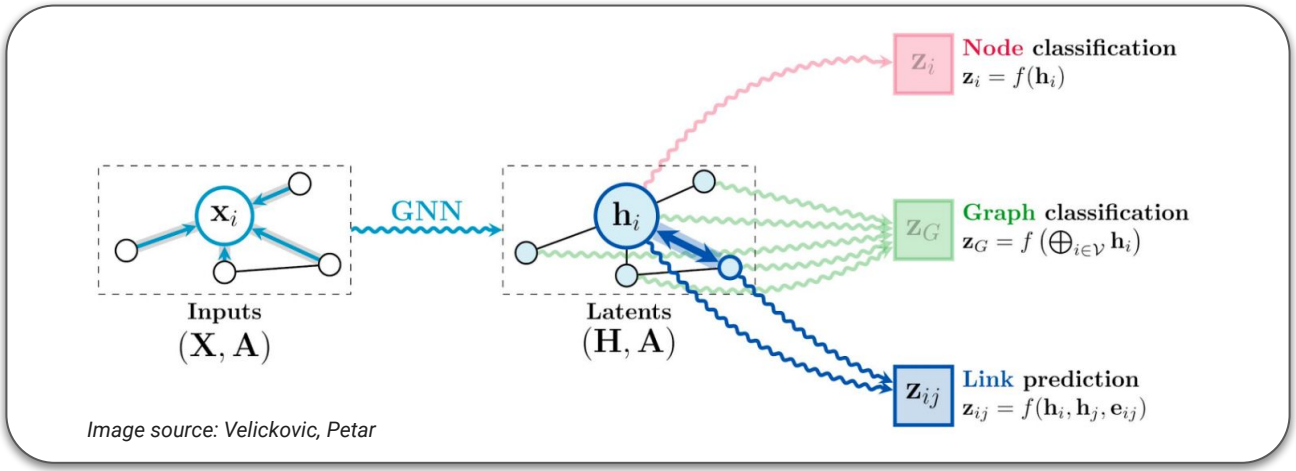
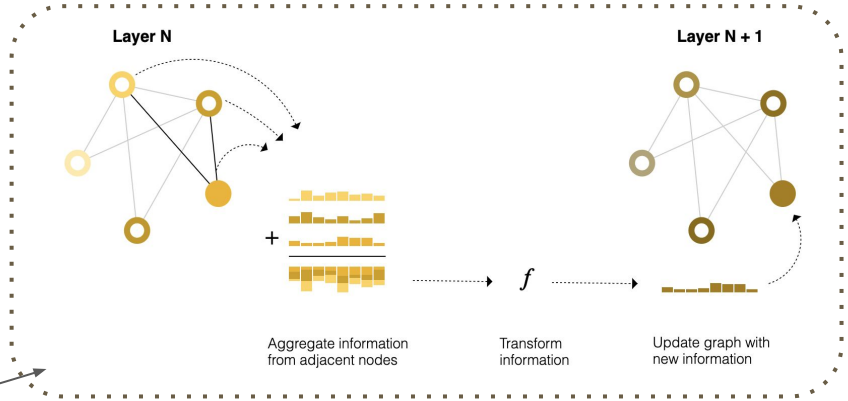
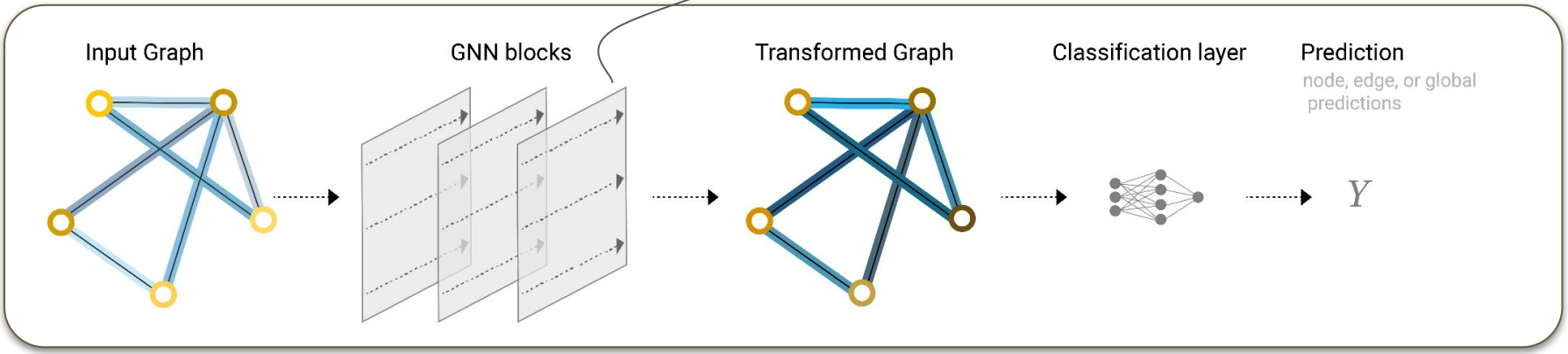


Image source: Velickovic, Petar

Overview of the Main GNN Components

- **Input layer**: defines initial representation of graph data, embed input feature to nodes and edges
- **GNN layer:** encodes information on graph structure, then, exploits information to update the initial representation of nodes and edges
- **Prediction layer (MLP)**: performs a specific learning task employing the encoded graph representation obtained from the GNN layer(s)

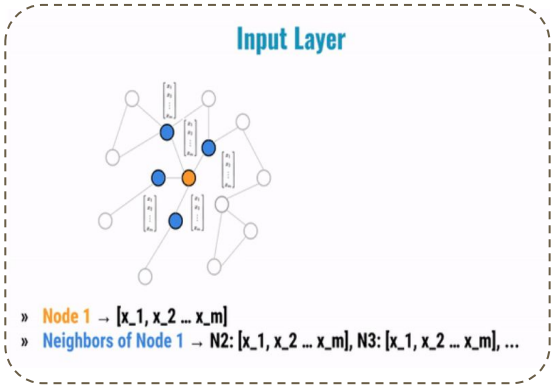


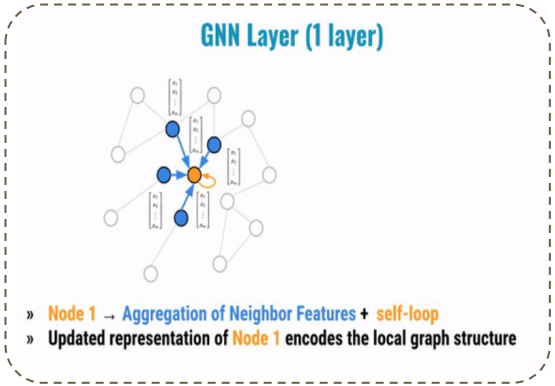


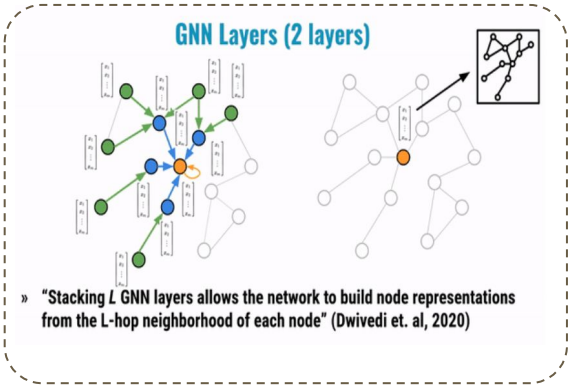
https://distill.pub/2021/gnn-intro/

GNN - Pairwise Message Passing

- GNN works by propagating features on the graph, exchanging information between adjacent nodes
- Initial features of nodes are updated by means of these steps:
  - Gather all the neighboring node embeddings (messages)
  - Aggregate all messages (via aggregate function)
  - All pooled messages are passed through an update function







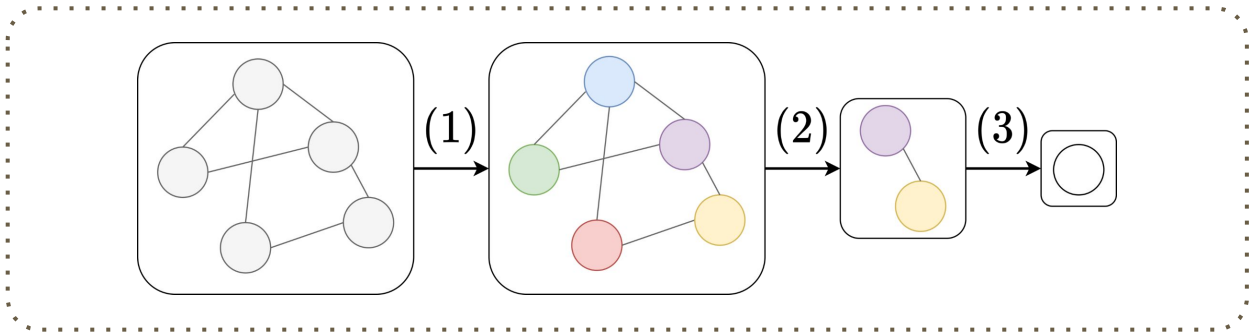
https://towardsdatascience.com/understanding-the-building-blocks-of-graph-neural-networks-intro-56627f0719d5

GNN - Pooling Layers

(1) - Pairwise message passing

(2) - Local pooling layer - coarsens graph via downsampling in a similar fashion to pooling layers in convolutional neural networks

(3) - Global pooling layer - provides fixed-size representation of the whole graph (must be permutation invariant)

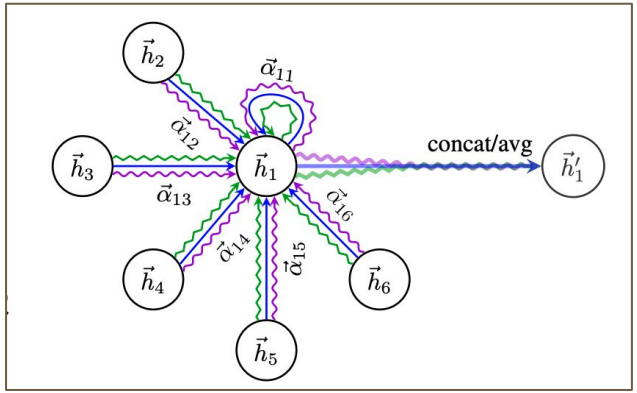


https://en.wikipedia.org/wiki/Graph_neural_network

Different GNN Models

The main differences between different GNN layers:

- Type of **aggregation**, which is performed by exploiting the local graph structure
- **Vanilla Graph Convolutional Networks** (GCNs) introduce in 2017, the aggregation/update is an isotropic operation, meaning features of neighbor nodes are considered in the same way.
- **Graph Attention Network** (GAT) introduced in 2018, aggregation is an anisotropic operation, in which contribution of each neighbor node in the aggregation is weighted according to its importance.



An illustration of multihead attention (with K = 3 heads) by node 1 on its neighborhood.
https://arxiv.org/pdf/1710.10903.pdf

Open Source Libraries for GNN Models

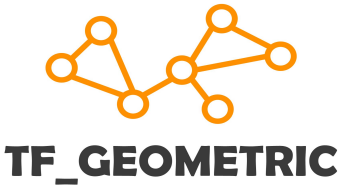
Several open source libraries implementing GNNs:

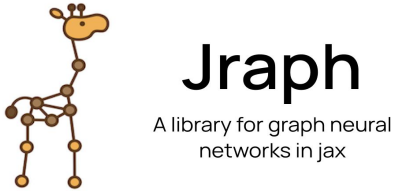
[Pytorch Geometric]

[TensorFlow GNN]

[Jraph] ([Colab])

[Deep Graph Library (DGL)]





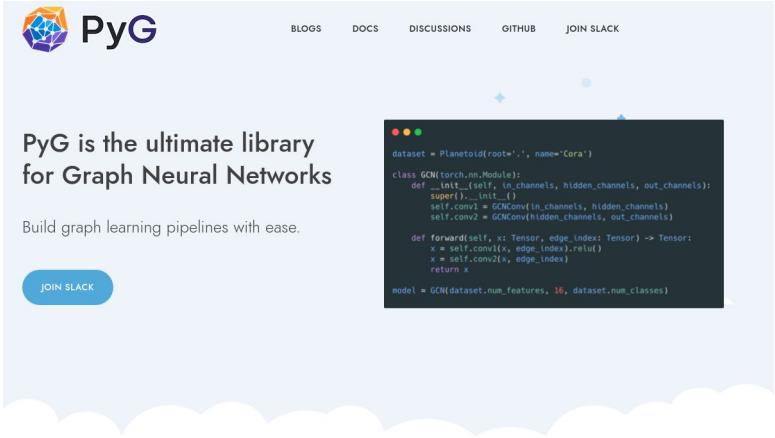
Section II:

How to build GNN models with PyTorch Geometric?

A popular open source library implementing graph neural networks.

PyG is a library built upon **PyTorch** to easily write and train **Graph Neural Networks** for a wide range of applications related to structured data.

PyG is both friendly to machine learning researchers and first-time users of machine learning toolkits.



PyTorch Geometric (PyG)

PyTorch Geometric (PyG) is a PyTorch library for deep learning on graphs, point clouds and manifolds

- simplifies implementing and working with Graph Neural Networks (GNNs)
- bundles fast implementations from published papers
- tries to be easily comprehensible and non-magical

PyG

- PyG consists of various methods for deep learning on graphs and other irregular structures
- PyG also known as geometric deep learning, from a variety of published papers.

Highly modularized pipeline for GNN:

Data: Data loading and data splitting

Model: Modularized GNN implementations

Tasks:

Node-level,

edge-level and

graph-level tasks

Evaluation: Accuracy, ROC AUC, ...

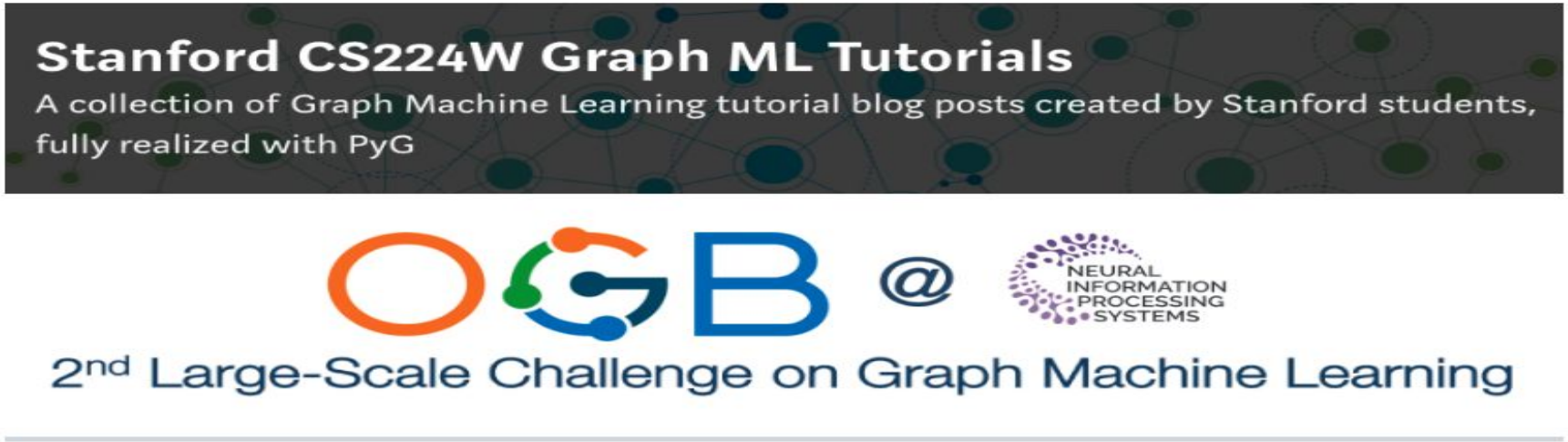
To learn more about PyG

https://www.pyg.org/

https://pytorch-geometric.readthedocs.io/en/latest/notes/introduction.html

https://pytorch-geometric.readthedocs.io/en/latest/notes/colabs.html

github : https://github.com/pyg-team/pytorch_geometric



Build GNN models with PyTorch Geometric

Link to Colab Tutorial for PyG

shorturl.at/fl036

See ya in 15 minutes!

Section III:

TorchDrug and TorchProtein

Powerful and flexible machine learning platforms for drug discovery.

TorchDrug is a PyTorch-based machine learning toolbox designed for several purposes.

- Easy implementation of graph operations in a PyTorchic style with GPU support
- Being friendly to practitioners with minimal knowledge about drug discovery
- Rapid prototyping of machine learning research

SMILES

SMILES (Simplified molecular-input line-entry system)

strings are a compact way of representing a molecule, instead of using systematic chemical names.

In SMILES, atoms are represented by their atomic symbols.

The notation allows you to quickly and easily generate novel chemical structures.

The basic unit of a SMILES string is the atomic symbol wrapped in brackets

TorchProtein

TorchProtein is built on top of TorchDrug to facilitate protein science with machine learning techniques from multiple aspects.

- Represent the protein sequence and structure with a unified data structure which supports GPU acceleration
- Define extensive and flexible building blocks for the rapid prototyping of machine learning solutions to diverse protein tasks
- Empower protein representation learning related applications with comprehensive benchmarks and a protein model

Links to learn more about TorchProtein

https://colab.research.google.com/drive/1sb2w3evdEWm-GYo28RksvzJ74p63xHMn?usp=sharing#scrollTo=TT31VHHdO7O7

https://github.com/DeepGraphLearning/torchdrug

https://torchdrug.ai/get_started

TorchDrog and TorchProtein Tutorial

Link to Colab Tutorial for TorchDrug

Link to Colab Tutorial for TorchProtein

TorchDrug: shorturl.at/jqHIO

TorchProtein: shorturl.at/FLRW5

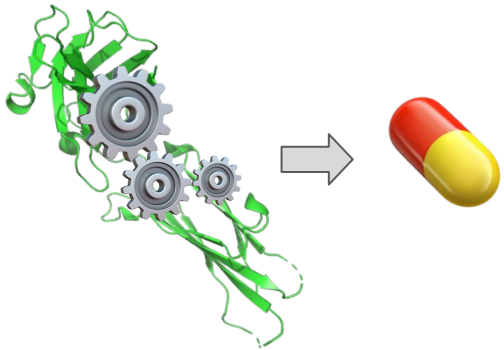
Section IV:

NodeCoder

A graph-based ML platform that predicts active sites of modeled proteins!

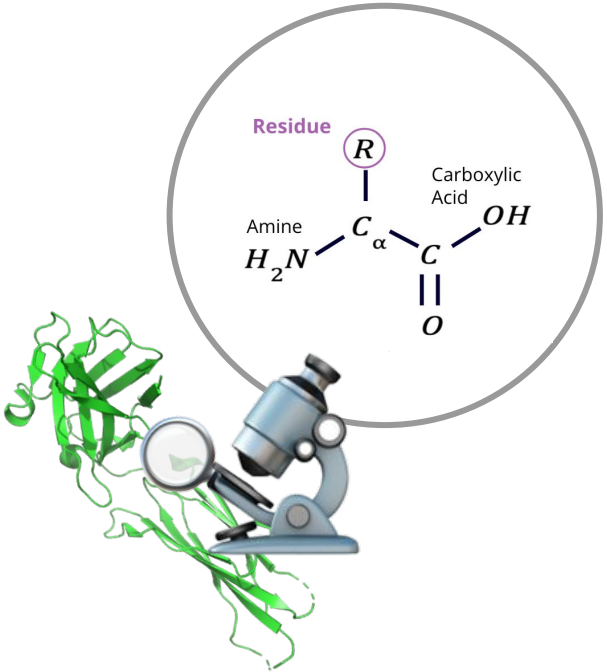
Importance of Unlocking Proteins' Functions in Drug Discovery/Design

- Proteins are building blocks of nature
- By unlocking their functions, biological scientists can better understand the natural world:
  - Understand disease more quickly
  - Opportunity to modify the functions for therapeutic uses
  - Design drugs for different diseases: Proteins can be stimulated or inhibited in terms of their normal functions



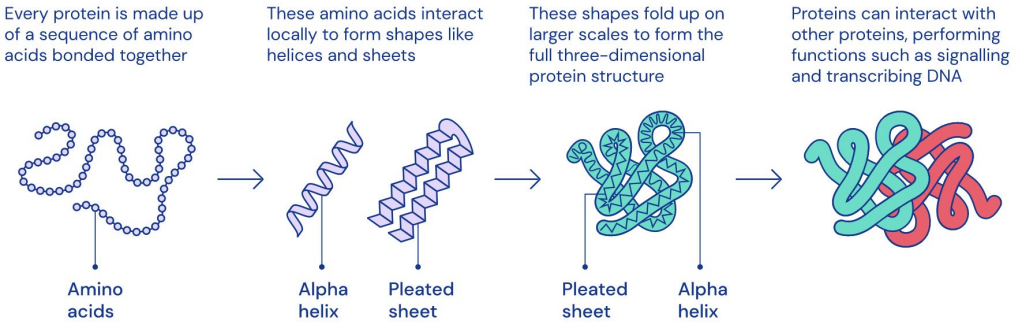
Residue Characterization to Achieve Protein Function Annotation

- Amino acids are building blocks that link together to form proteins
- There are 20 different amino acids for human proteins
- Amino acids contain (right figure)
  - Common elements between different amino acids
  - Specific elements called residues

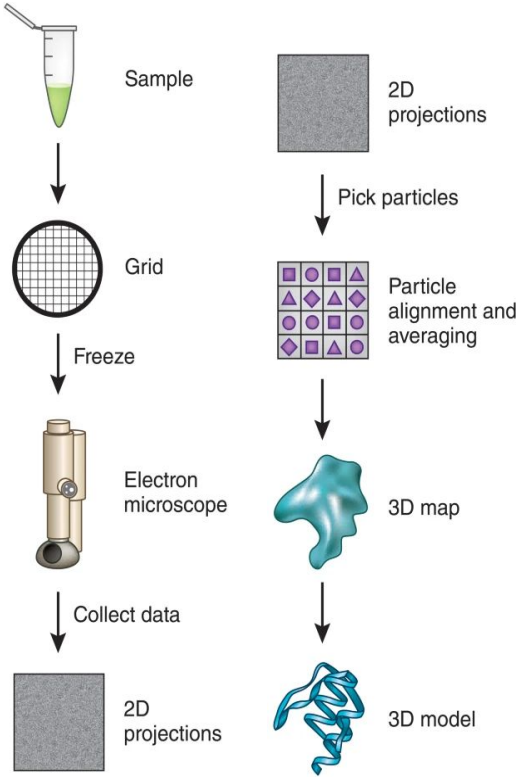


Protein Structure

- Protein - a sequence of amino acids that folds to form 3D structures
- Experimental methods such as X-ray crystallography, Cryo-electron microscopy are expensive and take months to years to solve a protein
- Protein structural coverage is bottlenecked by laborious and expensive process



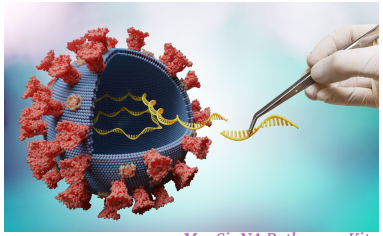
https://www.deepmind.com/blog/alphafold-using-ai-for-scientific-discovery-2020



The overall single-particle cryo-EM workflow, from protein sample to 3D model. Nature Publishing Group

Example

- Emerging pathogens, a target for pharmaceutical research!



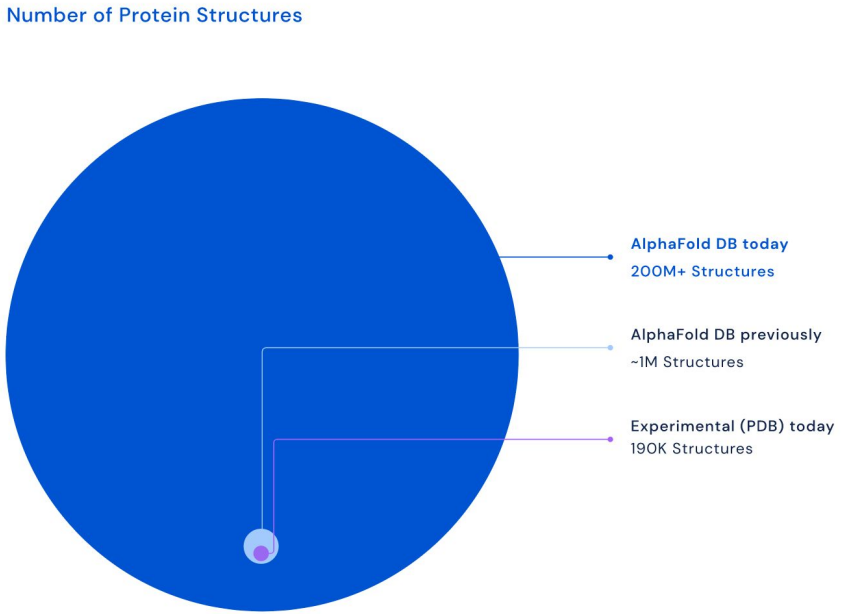
MagSi-NA Pathogens Kit

Genomic sequences of new pathogens → **Reliable predicted 3D structure models are rapidly generated** → **Predicted structure models can be used for functional characterization** → **Structure-based drug design and/or repurposing**

Following publication of novel coronavirus genome in Jan 2020, reliable 3D protein structure models for coronavirus proteins were automatically generated within days by: **SwissModel**, **ZhangLab**, **AlphaFold**, **Rosetta**

AlphaFold2 Proteome Coverage

- DeepMind and the EMBL-EBI released the structures of the protein universe
- July 2021, AlphaFold and EMBL-EBI launched the Protein Structure Database to accelerate scientific research (~1M structures, 48 species)
- July 28th, 2022, AlphaFold DB provides open access to
  - **200M+ protein structure** predictions
  - Predicted protein structures across **~1M species**
- **AlphaFold2 New Rival- Meta AI:** Nov 2022, predicted **600M+ protein structures** from metagenomics sources (language-model based structure prediction tool)



https://alphafold.ebi.ac.uk/
https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe
https://www.nature.com/articles/d41586-022-03539-1#:~:text=Researchers%20at%20Meta%20(formerly%20Facebook,that%20haven't%20been%20characterized.

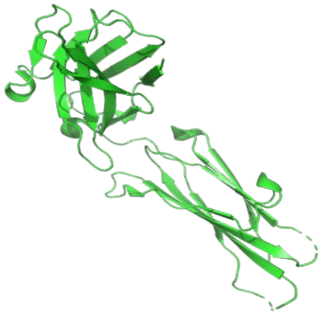
Limitations of Computational Structure Prediction Models

- Predicted structures lack much of the 'biological details' presented by experimental studies, such as **binding sites** and **macromolecular interfaces**!
- Biological details that provide scientists with important context required to interpret proteins' functions

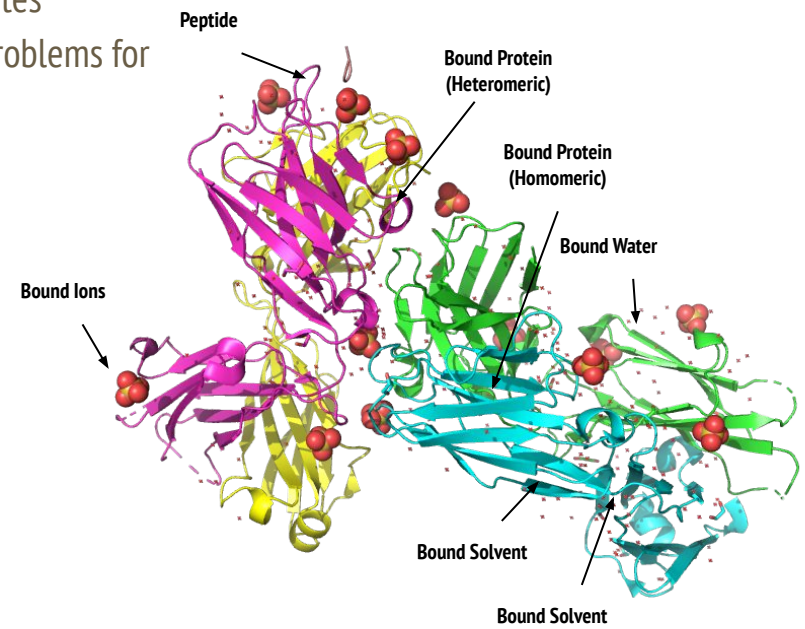
https://blog.cyclicarx.com/limitations-of-alphafold

NodeCoder

- A deep learning framework that annotates 3D protein structures with predicted ligand-, DNA-, RNA-, peptide-, or protein- binding sites
- With flexibility of being rapidly deployed to multiple new problems for predicting any residue-based annotations

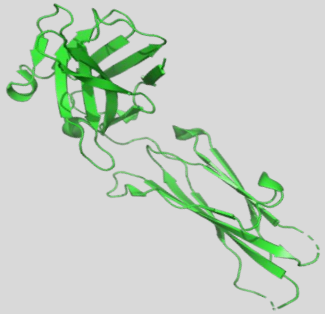


AlphaFold2 Modeled Structure (Single Chain)



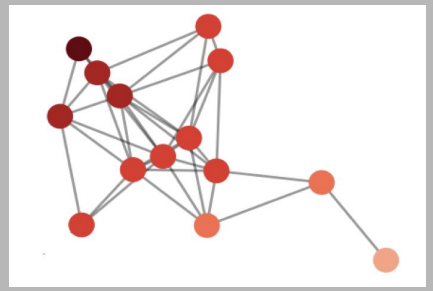
Protein graph for residue characterization

Protein Structure



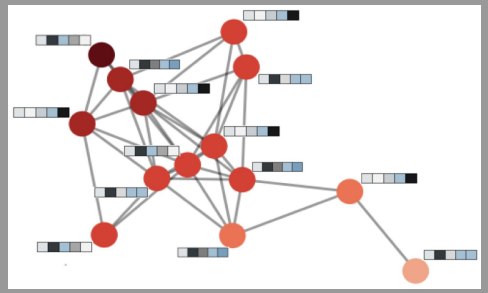
Contact Network

Each residue is a node in protein graph



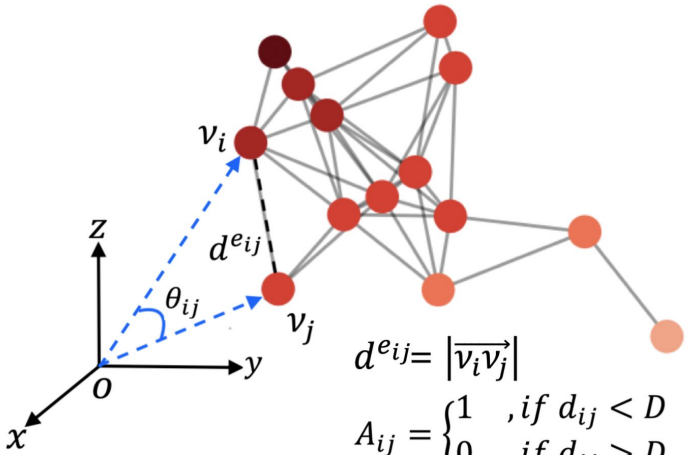
Node Features

Node features represent known properties of residue

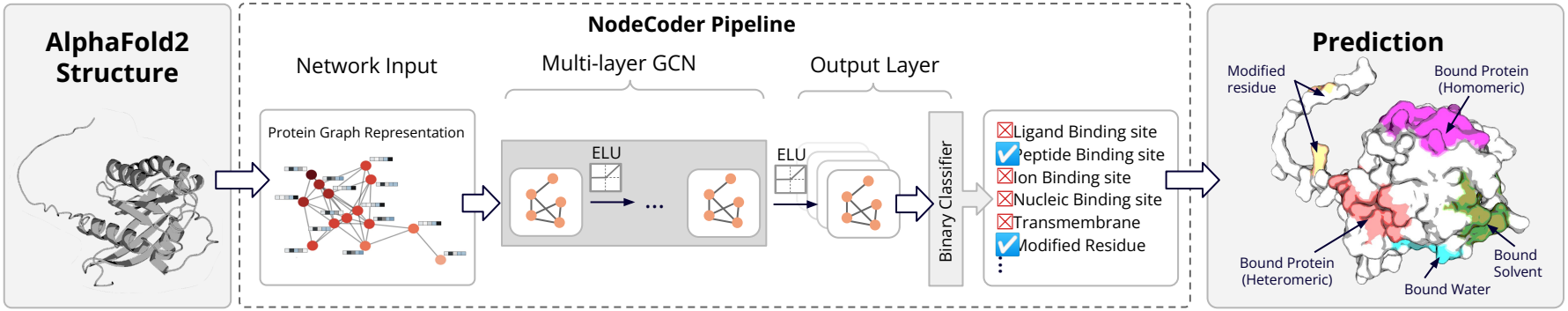


Protein graph construction

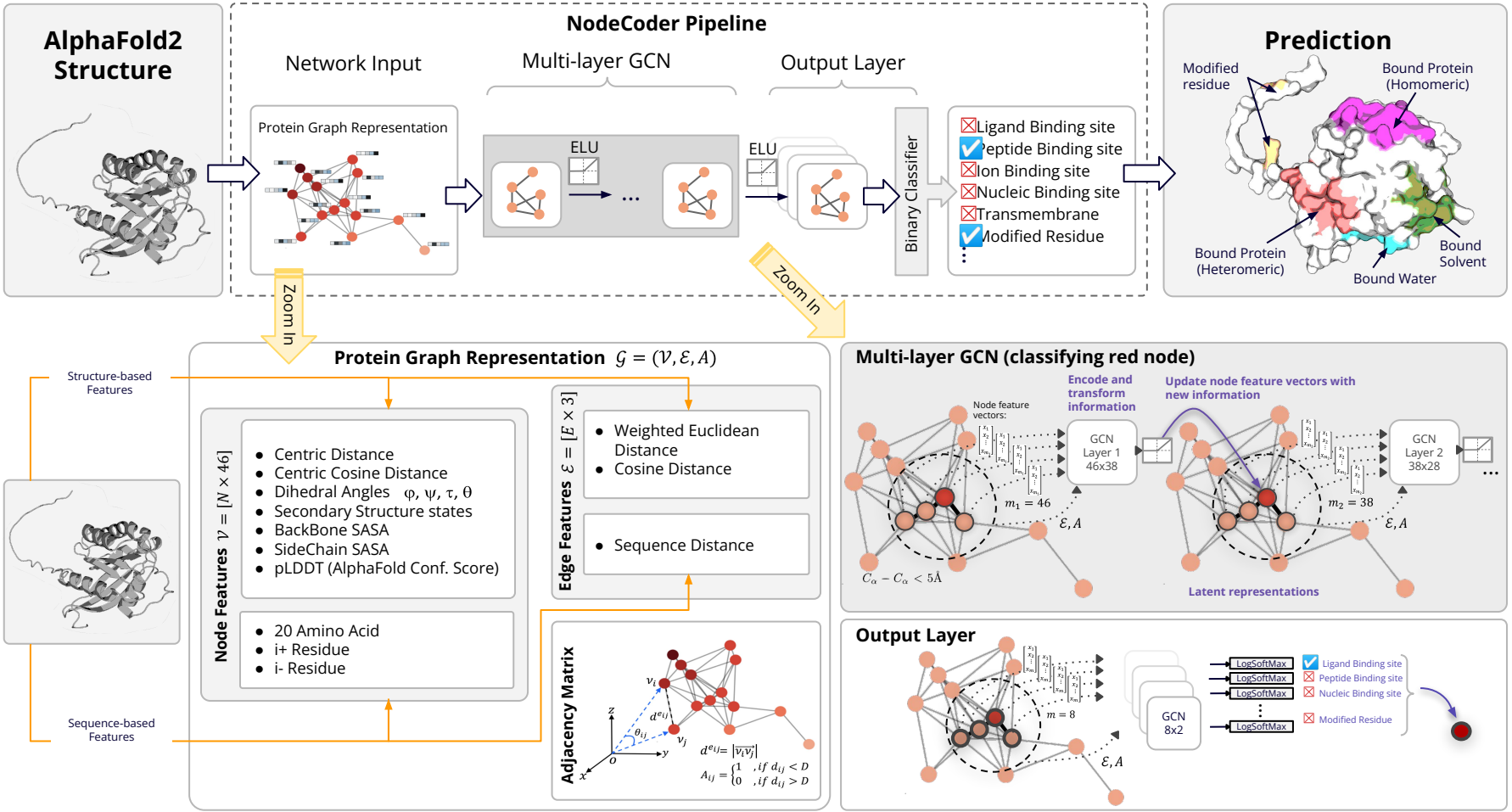
- Residue coordinates (C-alpha atoms)
- Edges correspond to inter-residue contacts within pre-set distance
- Edge features - spatial relation between residues



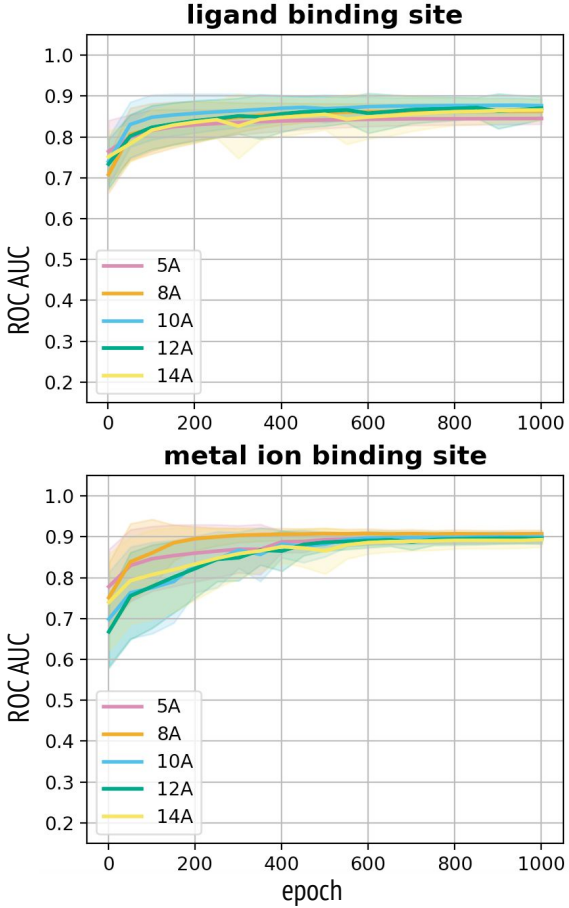
NodeCoder Pipeline

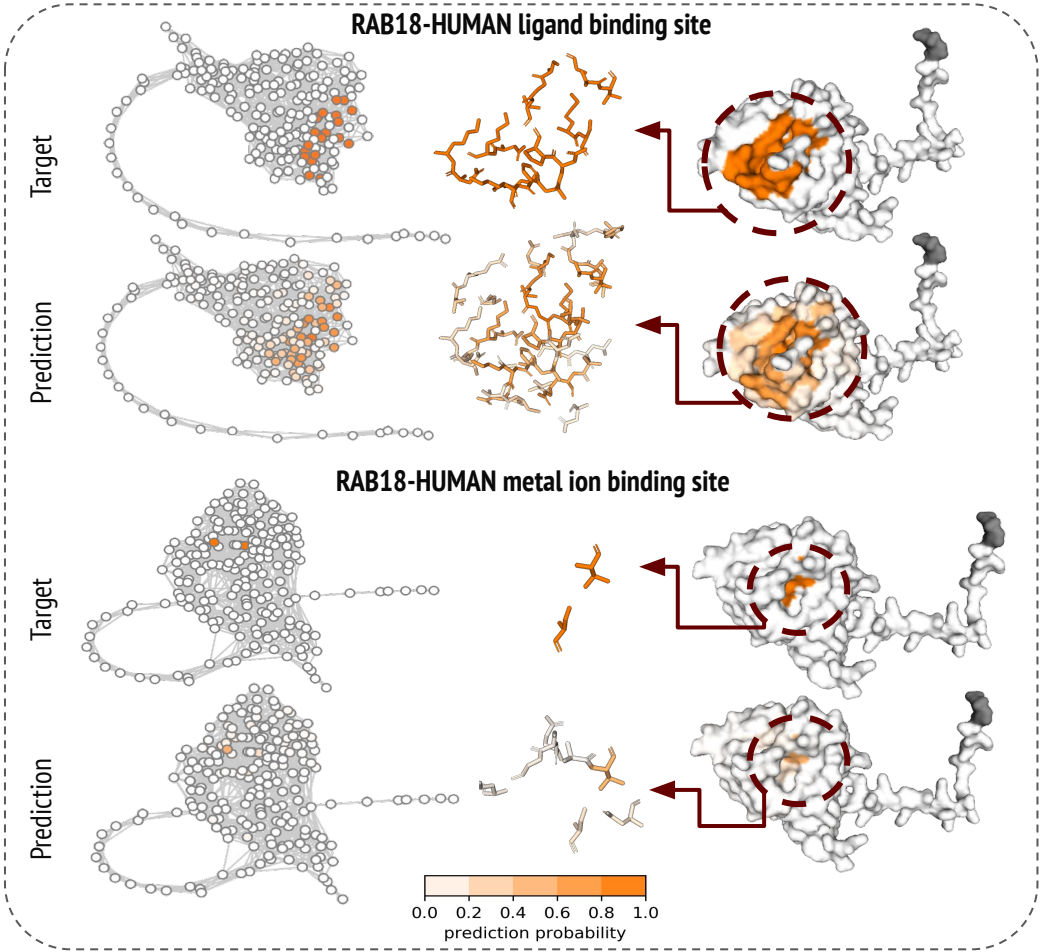


NodeCoder Pipeline

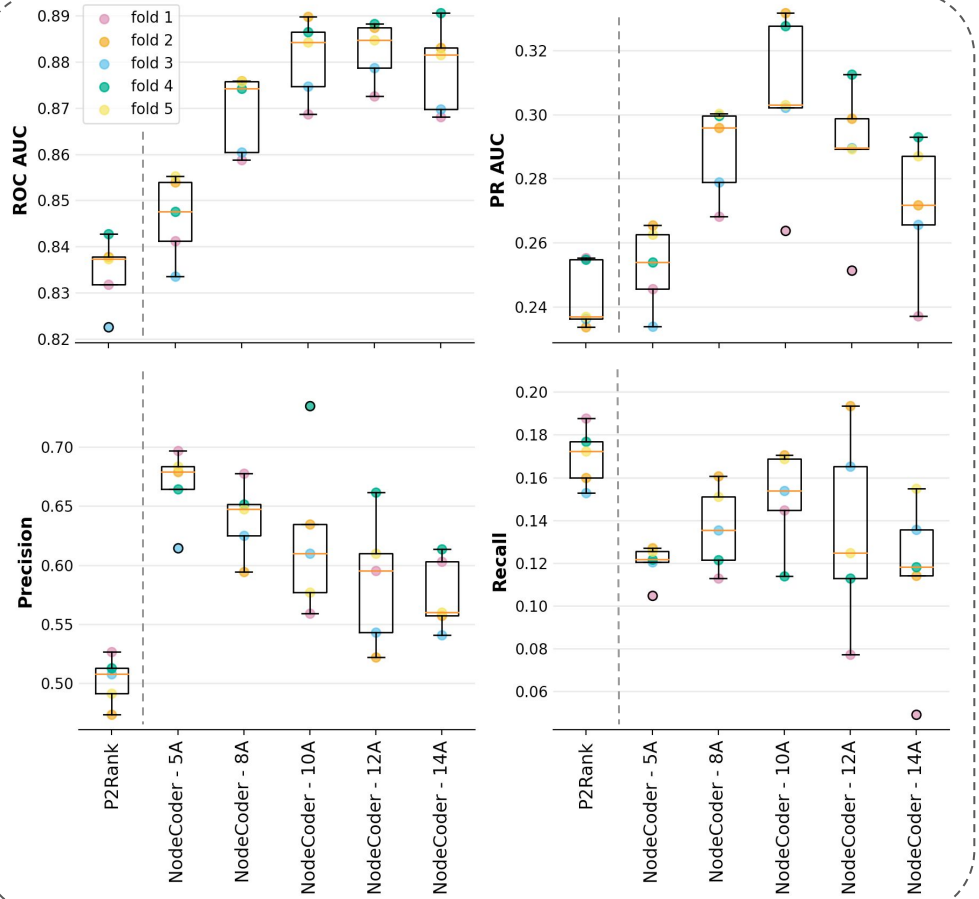


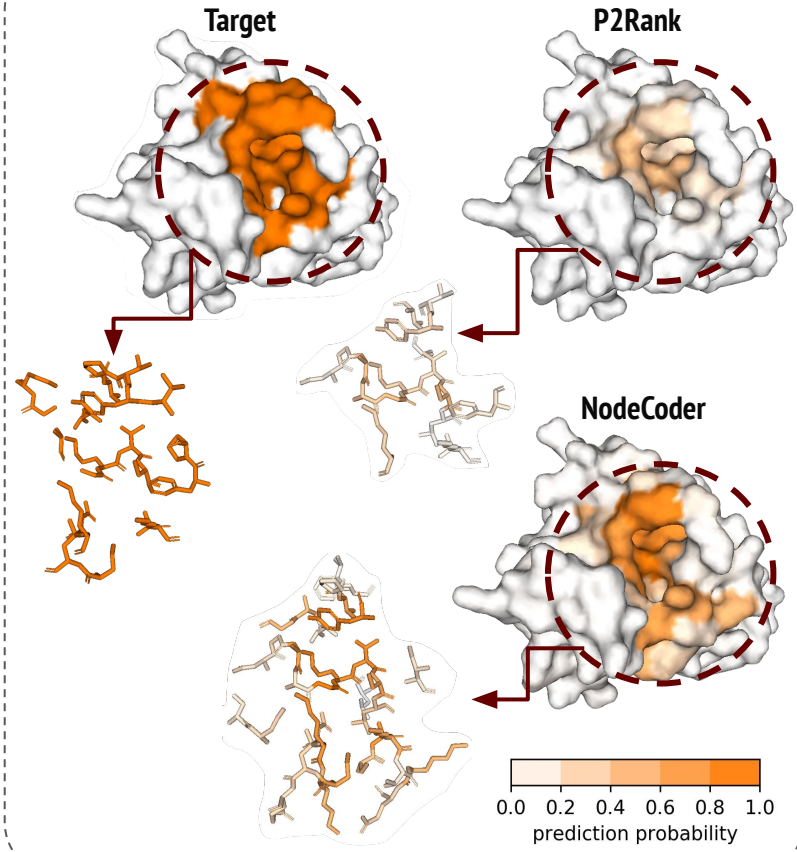
Experiments and Results





Experiments and Results



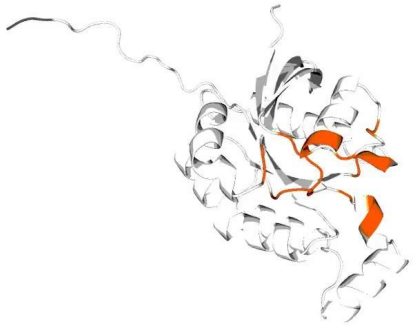


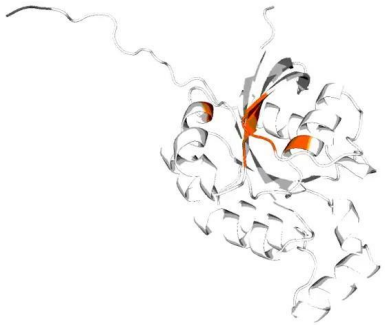
RHOB_HUMAN

Ligand Binding Site

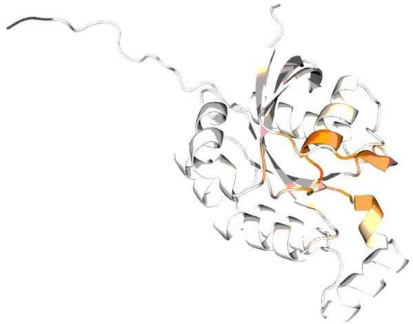
Inorganic Binding Site

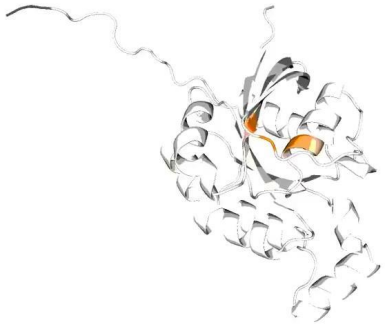
Target

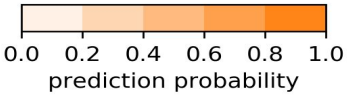




Prediction







Python Package

Installable Python Package and GitHub Repository

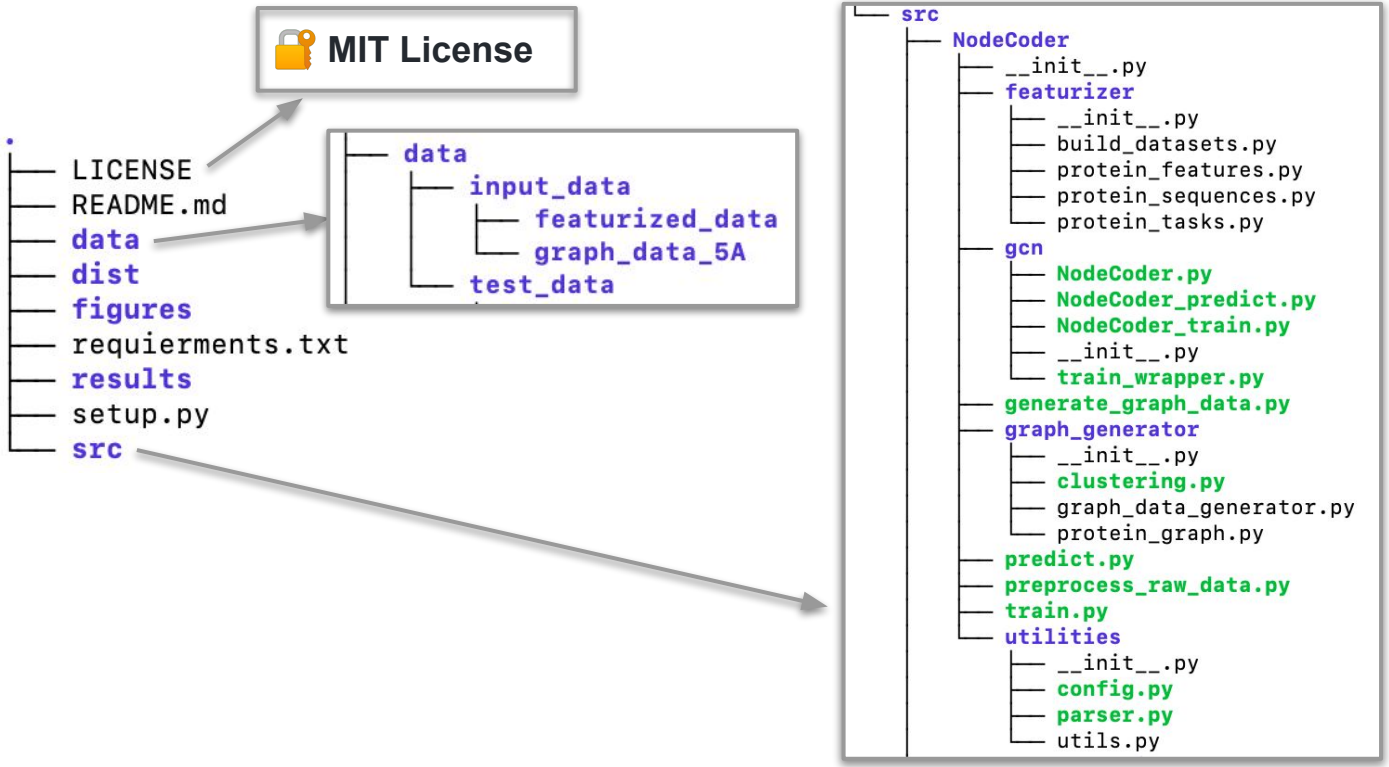
- Installable python package pypi.org
  https://pypi.org/project/NodeCoder/

```
$ pip install --extra-index-url https://pypi.org/simple/ NodeCoder
```

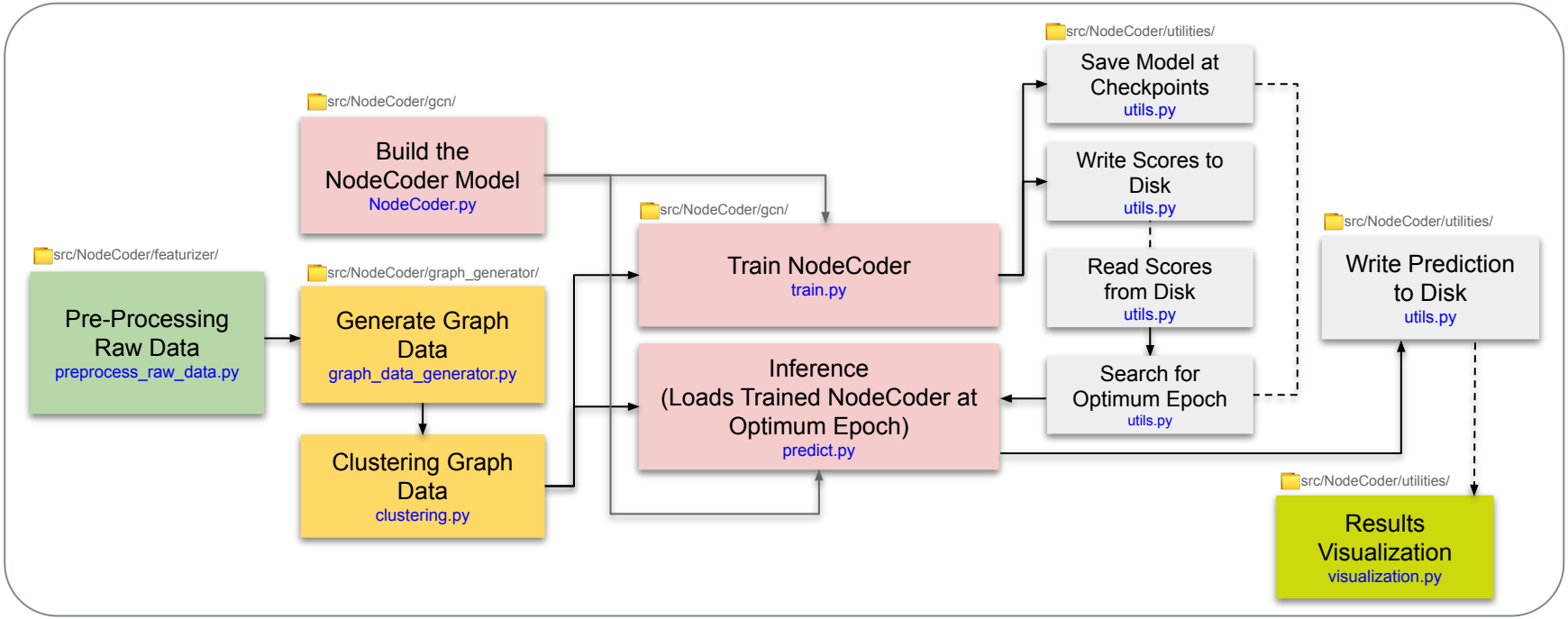
- Installable GitHub repo (still private)
  https://github.com/NasimAbdollahi/NodeCoder

```
$ git clone https://github.com/NasimAbdollahi/NodeCoder.git
$ pip install .
```

Installable Python Package and GitHub Repository



NodeCoder Pipeline Workflow



NodeCoder Usage

```
>>> from NodeCoder import preprocess_raw_data
>>> from NodeCoder import generate_graph_data
>>> from NodeCoder import train
>>> from NodeCoder import predict
```

```
>>> preprocess_raw_data.main(alphafold_data_path='.', uniprot_data_path='.', biolip_data_path='.',
biolip_data_skip_path='.', TAX_ID='9606', PROTEOME_ID='UP000005640')
```

```
>>> generate_graph_data.main(TAX_ID='9606', PROTEOME_ID='UP000005640', threshold_dist=5,
cross_validation_fold_number=5)
```

```
>>> train.main(threshold_dist=5, multi_task_learning=False, Task=['y_Ligand'], centrality_feature=True,
cross_validation_fold_number=5, epochs=1000)
```

```
>>> predict.main(protein_ID='KI3L1_HUMAN', threshold_dist=5, trained_model_fold_number=1,
multi_task_learning=False, Task=['y_Ligand'], centrality_feature=True, cross_validation_fold_number=5,
epochs=1000)
```

Thank you!

"For the things we have to learn before we can do them,
we learn by doing them."

— Aristotle

nasim.abdollahi@utoronto.ca/@cyclicarx.com

farnoosh.khodakarami@cyclicarx.com

References & Useful Links

- Graph convolutional networks: a comprehensive review
- A Gentle Introduction to Graph Neural Networks
- Understanding the Building Blocks of Graph Neural Networks
- https://en.wikipedia.org/wiki/Graph_neural_network
- The Graph Neural Network Model
- https://github.com/thunlp/GNNPapers
- Structure-based drug design with geometric deep learning
- Understanding the Building Blocks of Graph Neural Networks (Intro)
- Graph Attention Networks Under the Hood
- Graph Attention Networks - Annotated PyTorch Paper Implementations
- https://modelzoo.co/category/graph
- Medicinal Biotechnology for Disease Modeling, Clinical Therapy, and Drug Discovery and Development
- https://www.helsinki.fi/en/hilife-helsinki-institute-life-science/news/ai-guides-accurate-design-patient-tailored-combinatorial-strategies-aml-treatment
- https://www.deepmind.com/blog/alphafold-using-ai-for-scientific-discovery-2020
- https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
- https://alphafold.ebi.ac.uk/
- https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe
- Evolutionary-scale prediction of atomic level protein structure with a language model
- Structure-based drug design with geometric deep learning
- Graph Convolutional Neural Networks for Predicting Drug-Target Interactions
- IMCHGAN: Inductive Matrix Completion With Heterogeneous Graph Attention Networks for Drug-Target Interactions Prediction
- DRPreter: Interpretable Anticancer Drug Response Prediction Using Knowledge-Guided Graph Neural Networks and Transformer
- Deep Surrogate Docking: Accelerating Automated Drug Discovery with Graph Neural Networks
- GRAPH ATTENTION NETWORKS

- Pytorch Geometric
- torch_geometrics.nn
- GNN cheatsheet
- Deep Graph Library (DGL)
- TensorFlow GNN
- Jraph (Colab)
- TorchDrug
- TorchProtein
- TorchDrug Property Prediction Tutorial
- https://github.com/DeepGraphLearning/torchdrug
- https://torchdrug.ai/get_started

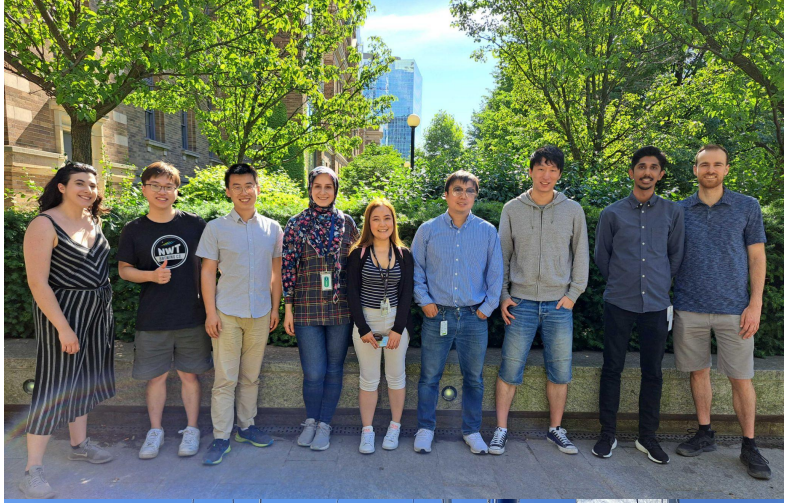
TorchDrug TorchProtein

PyG DGL DEEP GRAPH LIBRARY TF_GEOMETRIC Jraph

A library for graph neural networks in jax

Acknowledgements

CYCLICA
Mitacs
VECTOR INSTITUTE
UNIVERSITY OF TORONTO
compute canada | calcul canada
WangLab @UofT
CIFAR



Links:

- TorchGeometric: shorturl.at/fl036
- TorchDrug: shorturl.at/jqHIO
- TorchProtein: shorturl.at/FLRW5