Graph neural networks for molecular property prediction

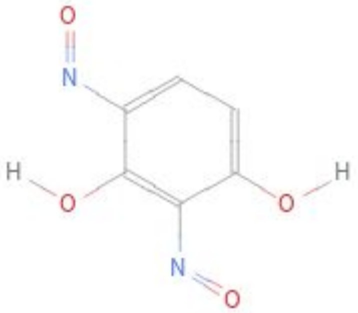
HIV replication inhibition

(And early work on
attempting to predict ~~West Nile Virus NS2bNS3 inhibition (dataset too small)~~
Flaviviral Genomic Capping Enzyme Inhibition)

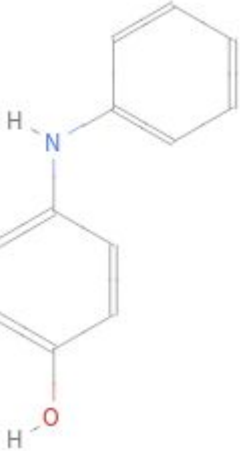
William Bruns
Stanford XCS224W student

Which molecule inhibits HIV replication?

(I'll make it easy by giving you a choice between 2 molecules, guess and you will be right 50% of the time)



O=Nc1ccc(O)c(N=O)c1O

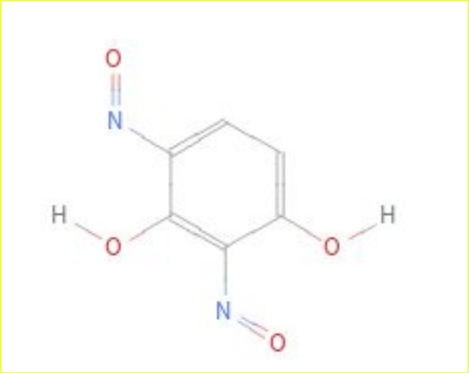


Oc1ccc(Nc2ccccc2)cc1

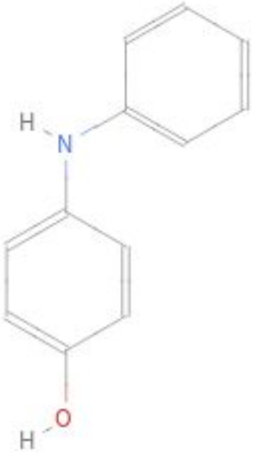
Molecule graphics from wolframalpha.com generated from Simplified Molecular Input Line Entry System (SMILES) strings from OGBG molhiv dataset
Example SMILES from https://snap.stanford.edu/ogb/data/graphproppred/csv_mol_download/hiv.zip
One of the first SMILES two class adjacent examples in hiv/mapping/mol.csv.gz (mapped to train split using train.csv, unmarked in this file) (during training OGB loader and official splits are used instead).
The very first example pair I was going to use, O=C(O)Cc1ccc(SSc2ccc(CC(=O)O)cc2)cc1 vs O=C(O)c1ccccc1SSc1ccccc1C(=O)O, was actually ambiguous, see https://pubchem.ncbi.nlm.nih.gov/compound/8409#section=Biological-Test-Results and search "aids"

Which molecule inhibits HIV replication?

(I’ll make it easy by giving you a choice between 2 molecules, guess and you will be right 50% of the time)



O=Nc1ccc(O)c(N=O)c1O
https://pubchem.ncbi.nlm.nih.gov/bioassay/179#sid=68320

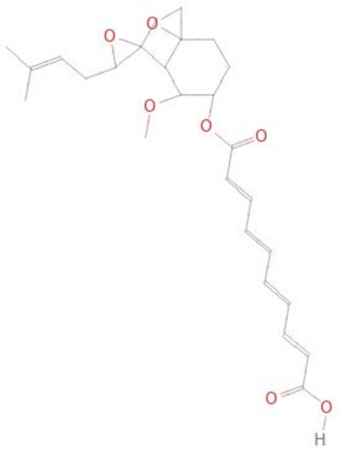


Oc1ccc(Nc2ccccc2)cc1
https://pubchem.ncbi.nlm.nih.gov/bioassay/179#sid=68322

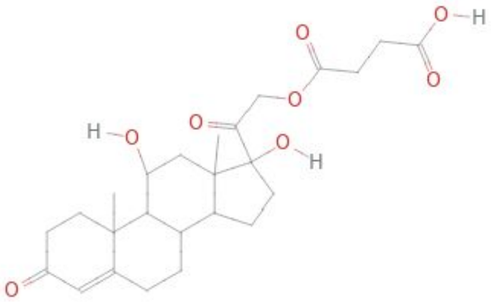
Molecule graphics from wolframalpha.com generated from Simplified Molecular Input Line Entry System (SMILES) strings from OGBG molhiv dataset
Example SMILES from https://snap.stanford.edu/ogb/data/graphproppred/csv_mol_download/hiv.zip
One of the first SMILES two class adjacent (original: CM vs CI) examples in hiv/mapping/mol.csv.gz (mapped to train split using train.csv, unmarked in this file) (during training OGB loader and official splits are used instead).
The very first example pair I was going to use, O=C(O)Cc1ccc(SSc2ccc(CC(=O)O)cc2)cc1 vs O=C(O)c1ccccc1SSc1ccccc1C(=O)O, was actually ambiguous, see https://pubchem.ncbi.nlm.nih.gov/compound/8409#section=Biological-Test-Results and search “aids”

Which molecule inhibits HIV replication?

(I’ll make it easy by giving you a choice between 2 molecules, guess and you will be right 50% of the time)



COC1C(OC(=O)C=CC=CC=CC=CC(=O)O)CCC2(CO2)C1C1(C)OC1CC=C(C)C

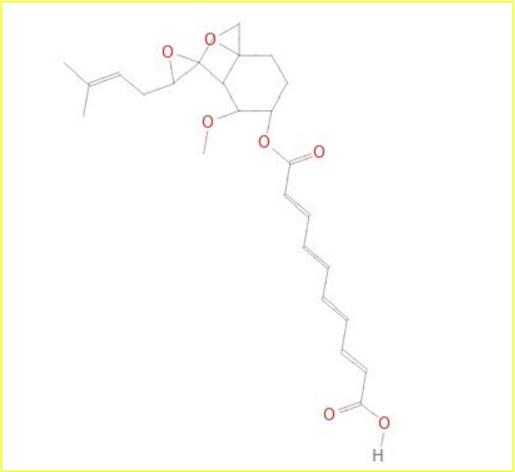


CC12CCC(=O)C=C1CCC1C2C(O)CC2(C)C1CCC2(O)C(=O)COC(=O)CCC(=O)O

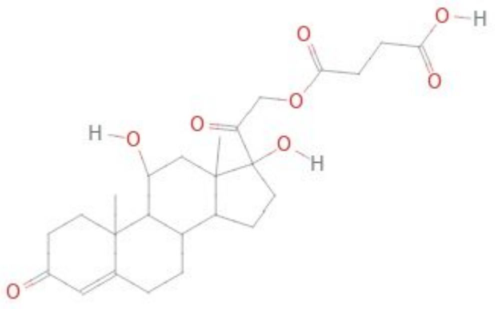
Molecule graphics from wolframalpha.com generated from SMILES strings from OGBG molhiv dataset
Example SMILES from https://snap.stanford.edu/ogb/data/graphproppred/csv_mol_download/hiv.zip
2 random SMILES, 1 from each class from hiv/mapping/mol.csv.gz (mapped to train split using train.csv, unmarked in this file) (during training OGB loader and official splits are used instead)

Which molecule inhibits HIV replication?

(I’ll make it easy by giving you a choice between 2 molecules, guess and you will be right 50% of the time)



Aka Fumagillin, which is not only a HIV replication inhibitor, btw: “Originally isolated from the fungus Aspergillus fumigatus, it is used for the control of Nosema infection in honey bees. It has a role as an angiogenesis inhibitor, an antibacterial drug, an antiprotozoal drug, a methionine aminopeptidase 2 inhibitor, an antimicrobial agent and a fungal metabolite.”



COC1C(OC(=O)C=CC=CC=CC=CC(=O)O)CCC2(CO2)C1C1(C)OC1CC=C(C)C
https://pubchem.ncbi.nlm.nih.gov/bioassay/179#sid=74694

CC12CCC(=O)C=C1CCC1C2C(O)CC2(C)C1CCC2(O)C(=O)COC(=O)CCC(=O)O
https://pubchem.ncbi.nlm.nih.gov/bioassay/179#sid=539584

Molecule graphics from wolframalpha.com generated from SMILES strings from OGBG molhiv dataset
Example SMILES from https://snap.stanford.edu/ogb/data/graphproppred/csv_mol_download/hiv.zip
2 random SMILES, 1 from each class (original CA vs CI) from hiv/mapping/mol.csv.gz (mapped to train split using train.csv, unmarked in this file) (during training OGB loader and official splits are used instead)

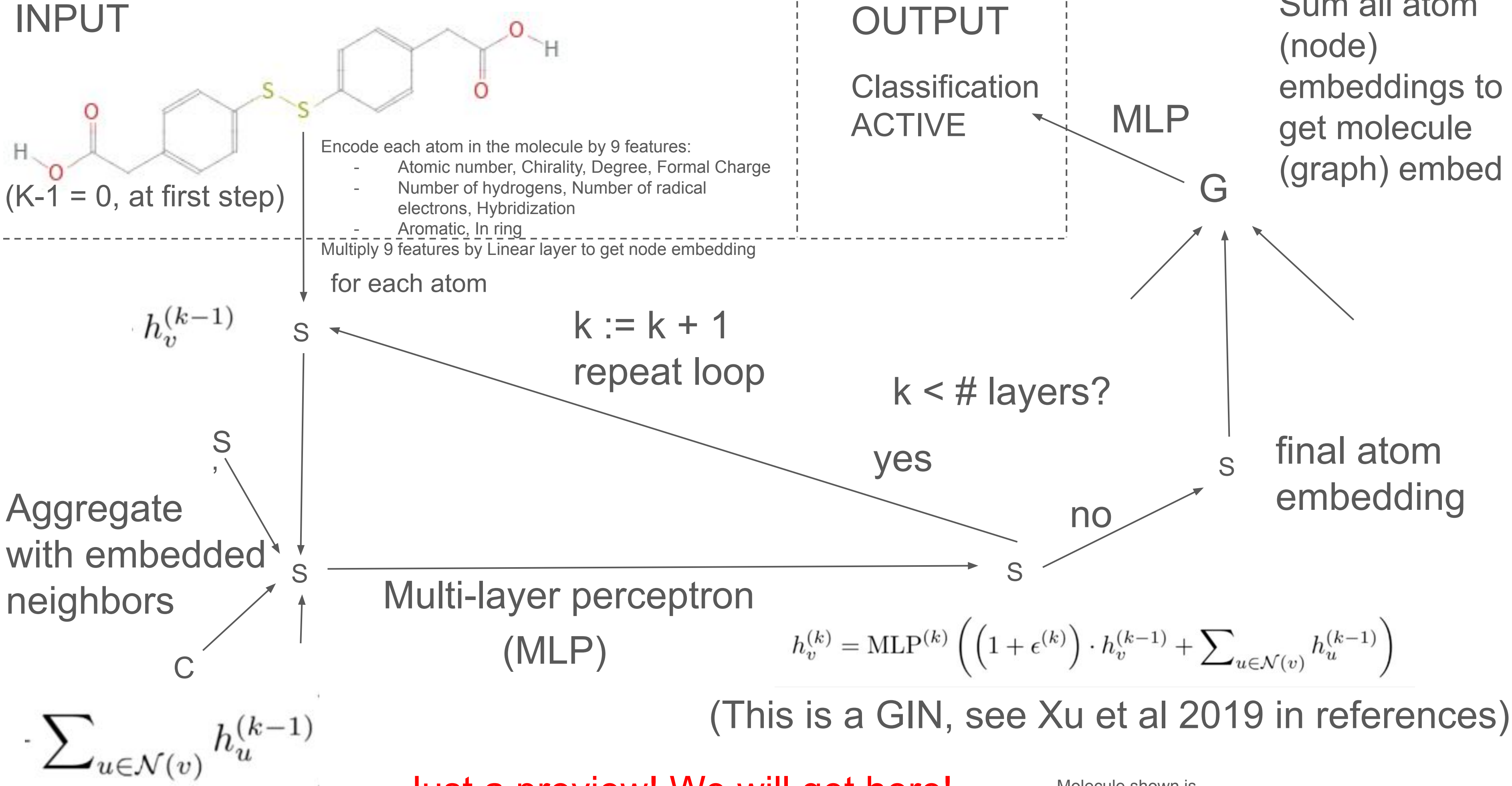
Can a computer predict this?
Why do we care?

“Time and money are precious resources when the vast majority of compounds fail to reach FDA approval and those that do cost $1.2 billion on average to research and develop.
When searching for lead molecules, it **costs about $100 to purchase a single compound in a commercially available library;** in the lead optimization phase, it costs about $2500 to synthesize a proposed derivative; up to another $2500 for functional assays of candidate ligands; and the subsequent mouse-model and human studies that follow a successful lead optimization campaign cost exponentially more.
A simple back-of-the-envelope calculation shows that experimentally testing all 100 million purchasable compounds in the ZINC small molecule database is financially intractable for even the best funded laboratories. Even then, the ZINC database is a small portion of the vast combinatorial expanse that is drug-like chemical space.”

- Evan Feinberg of Stanford’s Pande Lab 2018
  https://medium.com/@pandelab/ai-for-drug-discovery-in-two-stories-49d7b1f019f3

(The Pande Lab owns the MoleculeNet benchmark datasets (https://arxiv.org/abs/1703.00564) whose molHIV data is basis for OGB’s molhiv dataset shown earlier.
OGB/SNAP data is at https://snap.stanford.edu/ogb/data/graphproppred/csv_mol_download/hiv.zip (with train/valid/test splits),
derived from Pande Lab’s https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/HIV.csv ,
which is from a public dataset originally https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data , see also https://pubchem.ncbi.nlm.nih.gov/bioassay/179
. Note the predictions are made on “Active”/”Inactive” binary classification but moderately vs confirmed active are available in original dataset and numerical EC50 values are available in the AID 179 dataset.)

Tl;dr - We can predict: Use a Graph Neural Network



INPUT

(K-1 = 0, at first step)

Encode each atom in the molecule by 9 features:

- Atomic number, Chirality, Degree, Formal Charge
- Number of hydrogens, Number of radical electrons, Hybridization
- Aromatic, In ring

Multiply 9 features by Linear layer to get node embedding

for each atom

OUTPUT

Classification ACTIVE

MLP

Sum all atom (node) embeddings to get molecule (graph) embed

$h_v^{(k-1)}$

k := k + 1 repeat loop

k < # layers?

yes

no

final atom embedding

Aggregate with embedded neighbors

Multi-layer perceptron (MLP)

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

$$\sum_{u\in\mathcal{N}(v)} h_u^{(k-1)}$$

(This is a GIN, see Xu et al 2019 in references)

Just a preview! We will get here!

Molecule shown is https://pubchem.ncbi.nlm.nih.gov/bioassay/179#sid=67143

Inspiration: GNNs in the news

-

Healthcare:

- “Discovery of a structural class of antibiotics with explainable deep learning"
(Wong et al 2023, https://www.nature.com/articles/s41586-023-06887-8) (uses Chemprop, GNN library)
(molecular property prediction is NOT specific to antibiotics or antibacterials)

- "Massively Multitask Networks for Drug Discovery" (Ramsundar et al 2015, https://arxiv.org/abs/1502.02072 ; team that introduced MoleculeNet which is the basis of some OGB datasets including ogbg-molhiv)

- “Modeling Polypharmacy Side Effects with Graph Convolutional Networks” Zitnik et al 2018, https://arxiv.org/pdf/1802.00543 (via XCS224W)

- See also many references (separate from above) in Leskovec’s CS224W lecture 1.2 “Applications of Graph ML”

- Still machine learning on graphs but predicting protein structures:
AlphaFold, RoseTTAFold

- Designing amino acid sequences that fold to a specified structure: ProteinMPNN

SOTA Weather Forecasting:

- GraphCast uses GNN architecture Graph Isomorphism Network (GIN) to make global 10 day weather forecasts computable in 1 minute on a single machine that rival 6 hour national supercomputer forecasts
https://deepmind.google/discover/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/

Science generally:

- Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems
https://arxiv.org/abs/2307.08423

-

Objective: Stanford OGB benchmarks

+ Stretch goal of predicting PCBA-577-WNV (open data, not benchmark)

Open Graph Benchmark has multiple task types:

- Node attribute prediction
- Edge prediction
- Graph property prediction
    - Single-task
        - OGBG MoleculeNet MolHIV replication inhibition challenge
            - 41,127 molecules, 80/10/10 train/val/test splits, metric ROCAUC
            - Started with this
    - Multi-task
        - OGBG MoleculeNet PubChem BioAssay 128 multitask challenge
            - 437,929 molecules, metric AP
            - After single-task, doing this

MoleculeNet

A Benchmark for Molecular Machine Learning

A work by Pande Group at Stanford

uses

OGB

OPEN GRAPH BENCHMARK

uses

PubChem

NIH National Library of Medicine

National Center for Biotechnology Information

Noncompetition Unofficial Goal:

- non-OGB Single task -> Predict PubChem BioAssays for West Nile Virus or Flaviviruses
    - No current approved antivirals for West Nile Virus available!

      Dr. Fauci was sick with WNV this year (2024), which NIAID has funded research into for almost 25 years, with no treatments! They actually gave him antibiotics at first!

Tools

OPEN GRAPH BENCHMARK

- Data: OGB + MoleculeNet
    - Hu, Weihua and Fey, Matthias and Zitnik, Marinka and Dong, Yuxiao and Ren, Hongyu and Liu, Bowen and Catasta, Michele and Leskovec, Jure. Open Graph Benchmark: Datasets for Machine Learning on Graphs. arXiv preprint arXiv:2005.00687, 2020.
    - Wu, Zhenqin and Ramsundar, Bharath and Feinberg, Evan N and Gomes, Joseph and Geniesse, Caleb and SPappu, Aneesh and Leswing, Karl and Pande, Vijay. Moleculenet: a benchmark for molecular machine learning. Chemical Science, 9(2):513–530, 2018.

PyG

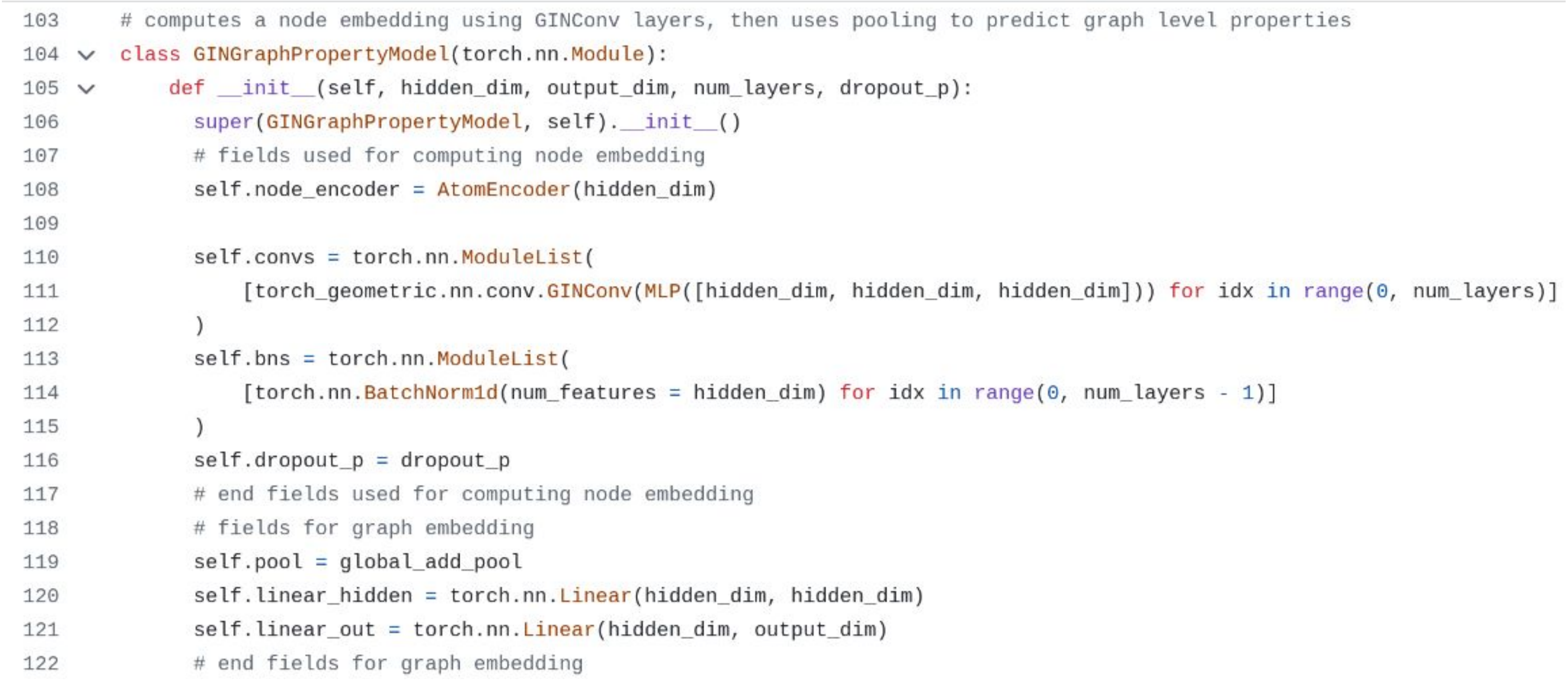
- Modeling: PyG + GIN
    - PyTorch Geometric (https://pyg.org/):
      Fey, Matthias and Lenssen, Jan E. Fast Graph Representation Learning with PyTorch Geometric. ICLR Workshop on Representation Learning on Graphs and Manifolds, 2019. (Graph Isomorphism Network (GIN) implementation used)
    - Graph Isomorphism Network:
      Xu, Keyulu and Hu, Weihua and Leskovec, Jure and Jegelka, Stefanie. How Powerful Are Graph Neural Networks? International Conference on Learning Representations, 2019. https://openreview.net/forum?id=ryGs6iA5Km , https://arxiv.org/pdf/1810.00826 . (Graph Isomorphism Network (GIN) original paper)

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

Approach: Tiny GIN

(32K parameters vs OGB team 1.8M parameter model)

https://github.com/willy-b/tiny-GIN-for-ogbg-molhiv



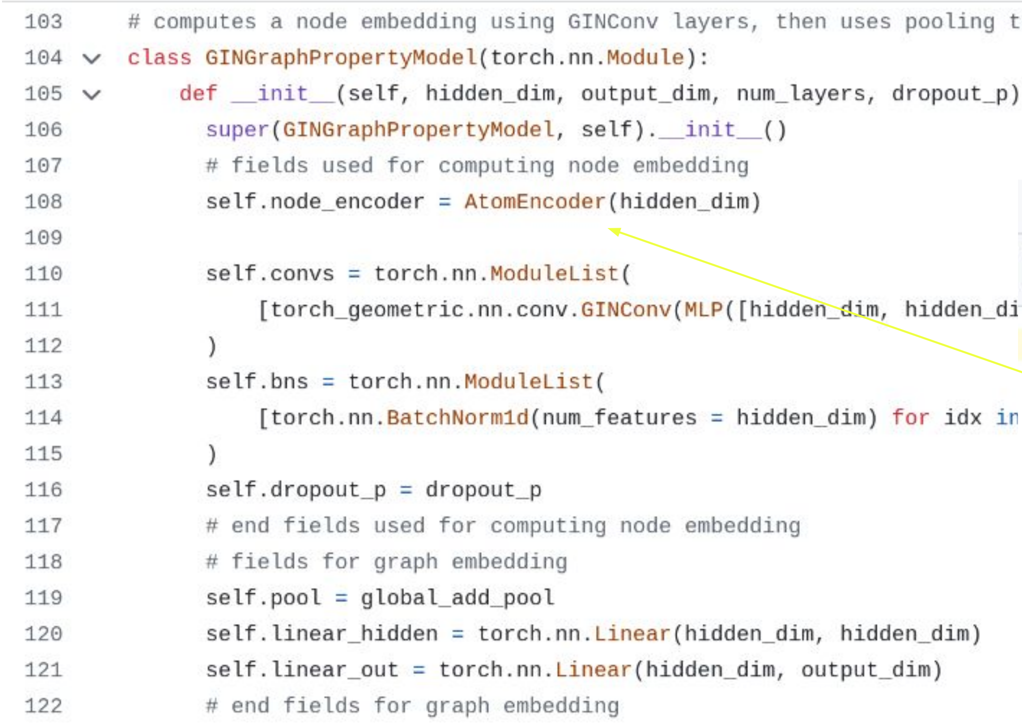
```
# computes a node embedding using GINConv layers, then uses pooling to predict graph level properties
class GINGraphPropertyModel(torch.nn.Module):
    def __init__(self, hidden_dim, output_dim, num_layers, dropout_p):
      super(GINGraphPropertyModel, self).__init__()
      # fields used for computing node embedding
      self.node_encoder = AtomEncoder(hidden_dim)

      self.convs = torch.nn.ModuleList(
          [torch_geometric.nn.conv.GINConv(MLP([hidden_dim, hidden_dim, hidden_dim])) for idx in range(0, num_layers)]
      )
      self.bns = torch.nn.ModuleList(
          [torch.nn.BatchNorm1d(num_features = hidden_dim) for idx in range(0, num_layers - 1)]
      )
      self.dropout_p = dropout_p
      # end fields used for computing node embedding
      # fields for graph embedding
      self.pool = global_add_pool
      self.linear_hidden = torch.nn.Linear(hidden_dim, hidden_dim)
      self.linear_out = torch.nn.Linear(hidden_dim, output_dim)
      # end fields for graph embedding
```

Approach: Tiny GIN

(32K parameters vs OGB team 1.8M parameter model)

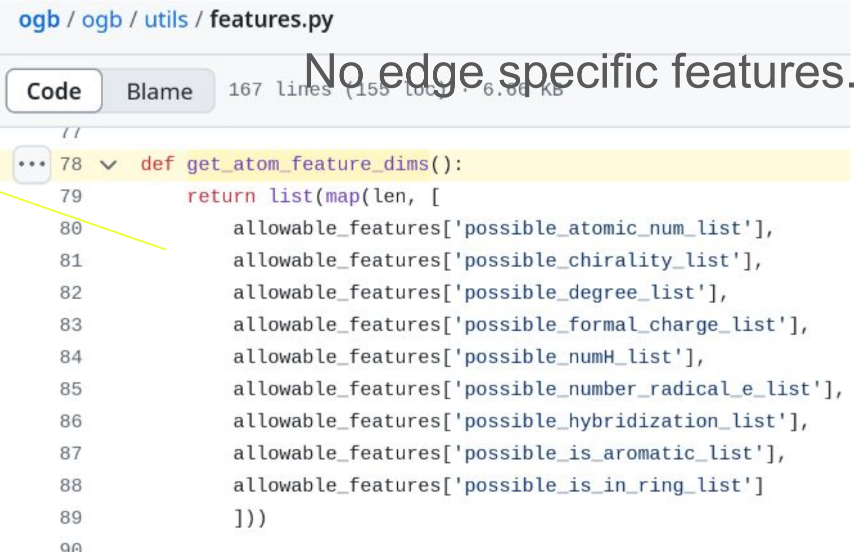
https://github.com/willy-b/tiny-GIN-for-ogbg-molhiv



```
# computes a node embedding using GINConv layers, then uses pooling to predict graph level properties
class GINGraphPropertyModel(torch.nn.Module):
    def __init__(self, hidden_dim, output_dim, num_layers, dropout_p):
        super(GINGraphPropertyModel, self).__init__()
        # fields used for computing node embedding
        self.node_encoder = AtomEncoder(hidden_dim)

        self.convs = torch.nn.ModuleList(
            [torch_geometric.nn.conv.GINConv(MLP([hidden_dim, hidden_di
        )
        self.bns = torch.nn.ModuleList(
            [torch.nn.BatchNorm1d(num_features = hidden_dim) for idx in
        )
        self.dropout_p = dropout_p
        # end fields used for computing node embedding
        # fields for graph embedding
        self.pool = global_add_pool
        self.linear_hidden = torch.nn.Linear(hidden_dim, hidden_dim)
        self.linear_out = torch.nn.Linear(hidden_dim, output_dim)
        # end fields for graph embedding
```

Using OGB AtomEncoder 9 feature Atom representation.



ogb / ogb / utils / features.py

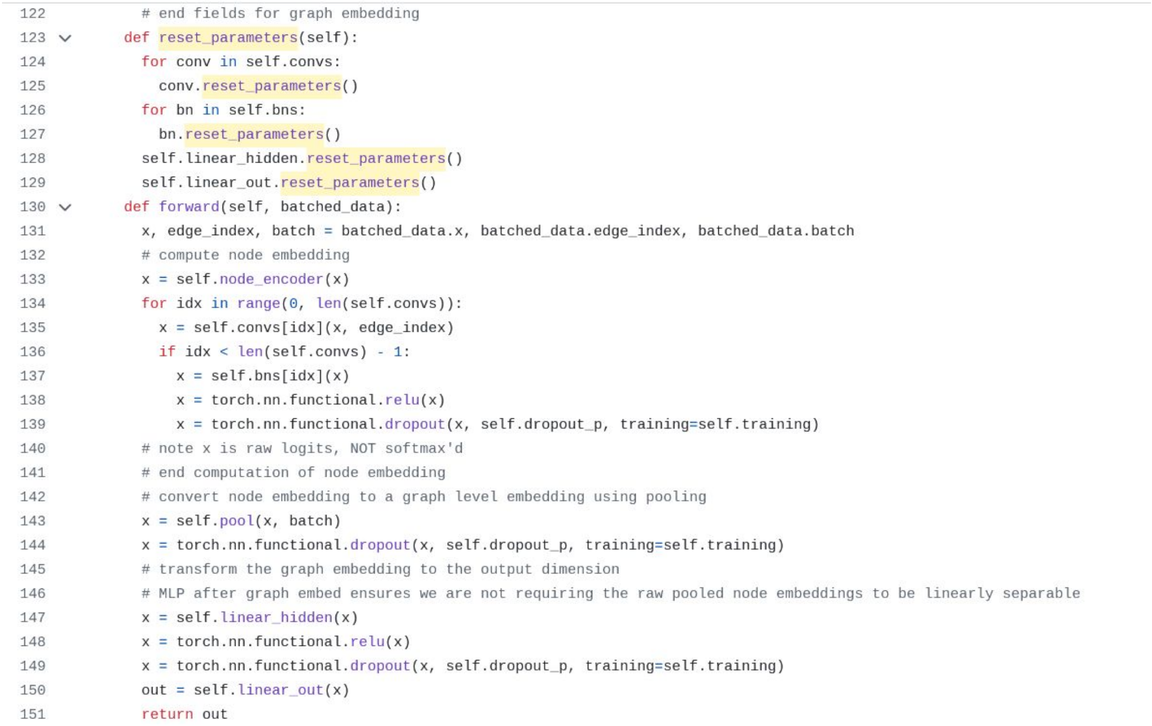
No edge specific features.

```
def get_atom_feature_dims():
    return list(map(len, [
        allowable_features['possible_atomic_num_list'],
        allowable_features['possible_chirality_list'],
        allowable_features['possible_degree_list'],
        allowable_features['possible_formal_charge_list'],
        allowable_features['possible_numH_list'],
        allowable_features['possible_number_radical_e_list'],
        allowable_features['possible_hybridization_list'],
        allowable_features['possible_is_aromatic_list'],
        allowable_features['possible_is_in_ring_list']
        ]))
```

Approach: Tiny GIN

(32K parameters vs OGB team 1.8M parameter model)

https://github.com/willy-b/tiny-GIN-for-ogbg-molhiv

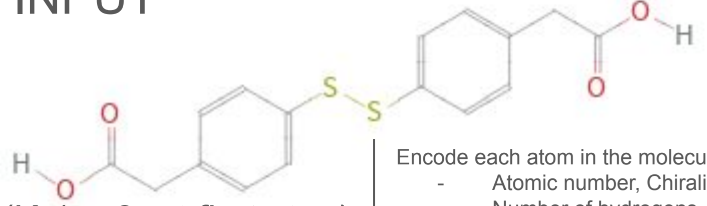


```python
    # end fields for graph embedding
  def reset_parameters(self):
    for conv in self.convs:
      conv.reset_parameters()
    for bn in self.bns:
      bn.reset_parameters()
    self.linear_hidden.reset_parameters()
    self.linear_out.reset_parameters()
  def forward(self, batched_data):
    x, edge_index, batch = batched_data.x, batched_data.edge_index, batched_data.batch
    # compute node embedding
    x = self.node_encoder(x)
    for idx in range(0, len(self.convs)):
      x = self.convs[idx](x, edge_index)
      if idx < len(self.convs) - 1:
        x = self.bns[idx](x)
        x = torch.nn.functional.relu(x)
        x = torch.nn.functional.dropout(x, self.dropout_p, training=self.training)
    # note x is raw logits, NOT softmax'd
    # end computation of node embedding
    # convert node embedding to a graph level embedding using pooling
    x = self.pool(x, batch)
    x = torch.nn.functional.dropout(x, self.dropout_p, training=self.training)
    # transform the graph embedding to the output dimension
    # MLP after graph embed ensures we are not requiring the raw pooled node embeddings to be linearly separable
    x = self.linear_hidden(x)
    x = torch.nn.functional.relu(x)
    x = torch.nn.functional.dropout(x, self.dropout_p, training=self.training)
    out = self.linear_out(x)
    return out
```

(continued from last slide)

Tl;dr - We can predict: Use a Graph Neural Network

INPUT



(K-1 = 0, at first step)

Encode each atom in the molecule by 9 features:

- Atomic number, Chirality, Degree, Formal Charge
- Number of hydrogens, Number of radical electrons, Hybridization
- Aromatic, In ring

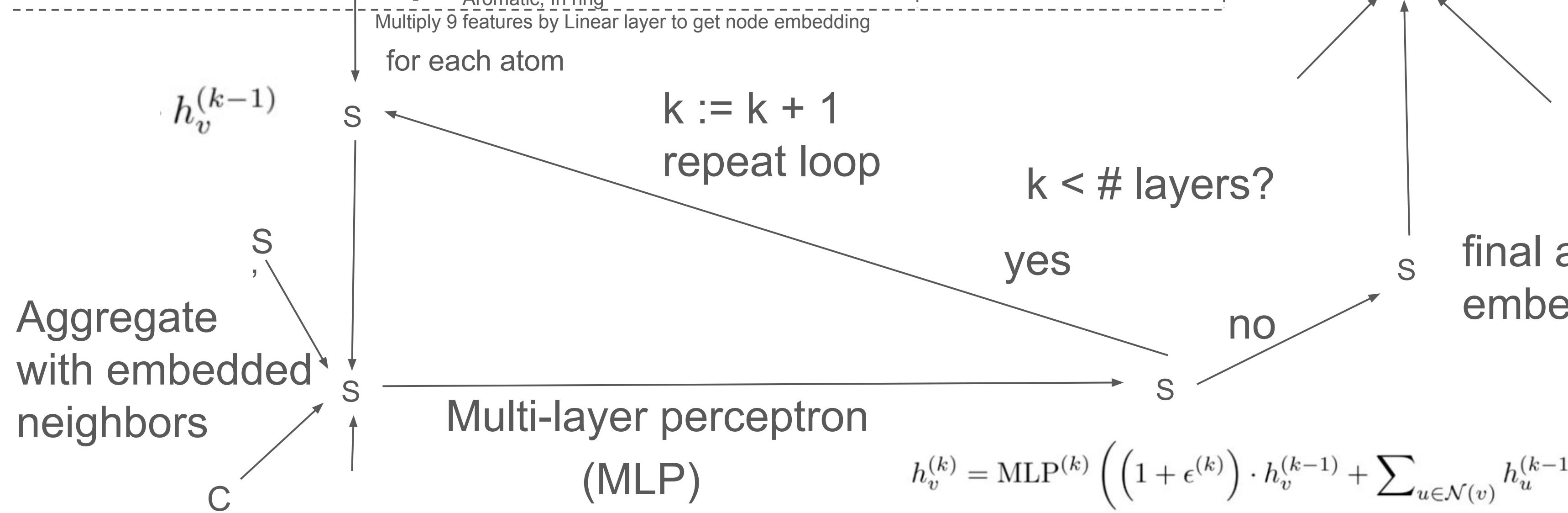
Multiply 9 features by Linear layer to get node embedding

OUTPUT

Classification ACTIVE

MLP

Sum all atom (node) embeddings to get molecule (graph) embed



for each atom

$h_v^{(k-1)}$

k := k + 1
repeat loop

k < # layers?

yes

no

final atom embedding

Aggregate with embedded neighbors

Multi-layer perceptron (MLP)

$$\sum_{u \in \mathcal{N}(v)} h_u^{(k-1)}$$

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

(This is a GIN, see Xu et al 2019 in references)

Molecule shown is
https://pubchem.ncbi.nlm.nih.gov/bioassay/179#sid=67143

Approach: Tiny GIN

(32K parameters vs OGB team 1.8M parameter model)

https://github.com/willy-b/tiny-GIN-for-ogbg-molhiv

Hyperparameter values used:

(results in 32,385 model parameters per `sum(p.numel() for p in model.parameters())`, the advised way to count model parameters per https://web.archive.org/web/20240324175343/https://ogb.stanford.edu/docs/leader_overview/)

num_layers: 2 (vs 5 layers in OGB team solution)

hidden_dim: 64 (vs 300 in the OGB team solution)

dropout: 0.5

learning_rate: 0.001

epochs: 50

batch_size: 32

weight_decay: 1e-6

per e.g. "Keeping Neural Networks Simple by Minimizing the Description Length of the Weights" (Hinton et al 1993, https://www.cs.toronto.edu/~fritz/absps/colt93.pdf)

add/sum pooling

MLP after node->graph embed pooling

9 atom features used, all edge features ignored

Choice of 2 layers is based on experiment and justified by e.g. GCN GNN layers/hops discussion in Xu et al 2018 “Representation Learning on Graphs with Jumping Knowledge Networks” https://arxiv.org/pdf/1806.03536 . Avoids over-smoothing.

Noting that the depth of network for GNN is not the same as depth of network for non-GNN deep neural networks, as it also controls the number of hops in the graph considered for the embedding of each node; one could also make the network used to compute node embedding based on each hop deeper without changing the number of GNN layers (hops)).

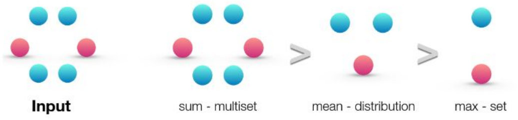
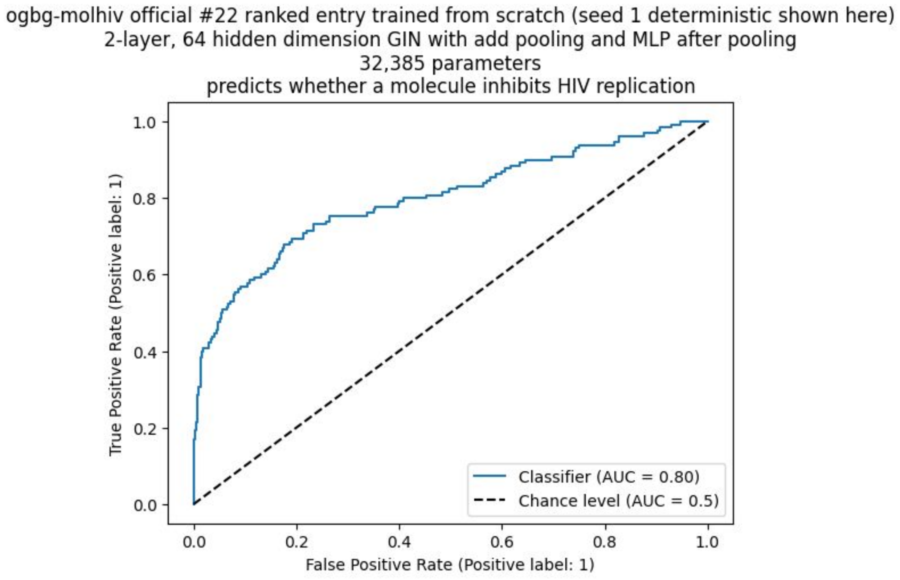
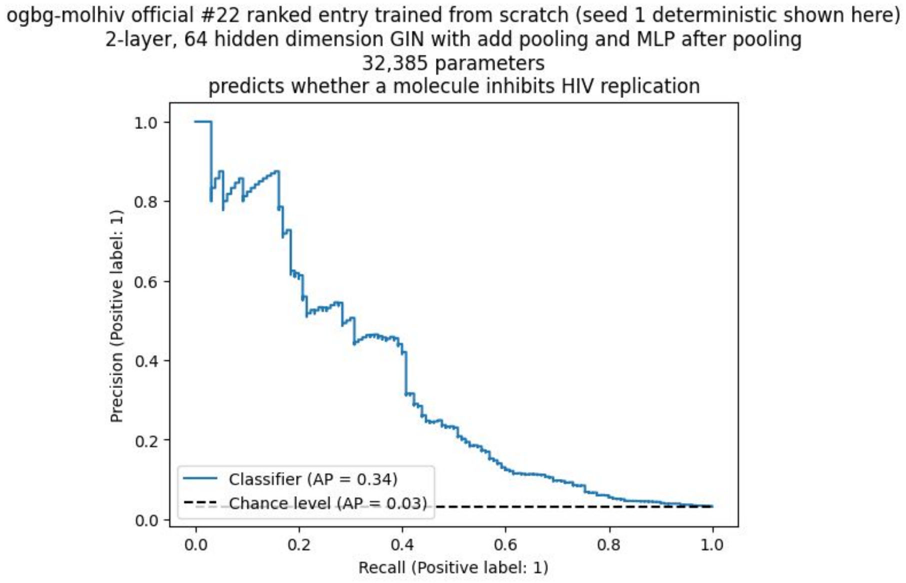


Figure 2: **Ranking by expressive power for sum, mean and max aggregators over a multiset.** Left panel shows the input multiset, *i.e.*, the network neighborhood to be aggregated. The next three panels illustrate the aspects of the multiset a given aggregator is able to capture: sum captures the full multiset, mean captures the proportion/distribution of elements of a given type, and the max aggregator ignores multiplicities (reduces the multiset to a simple set).

Results:
Receiver Operating Characteristic Curve
and Precision-Recall Curve


ogbg-molhiv official #22 ranked entry trained from scratch (seed 1 deterministic shown here)
2-layer, 64 hidden dimension GIN with add pooling and MLP after pooling
32,385 parameters
predicts whether a molecule inhibits HIV replication


ogbg-molhiv official #22 ranked entry trained from scratch (seed 1 deterministic shown here)
2-layer, 64 hidden dimension GIN with add pooling and MLP after pooling
32,385 parameters
predicts whether a molecule inhibits HIV replication

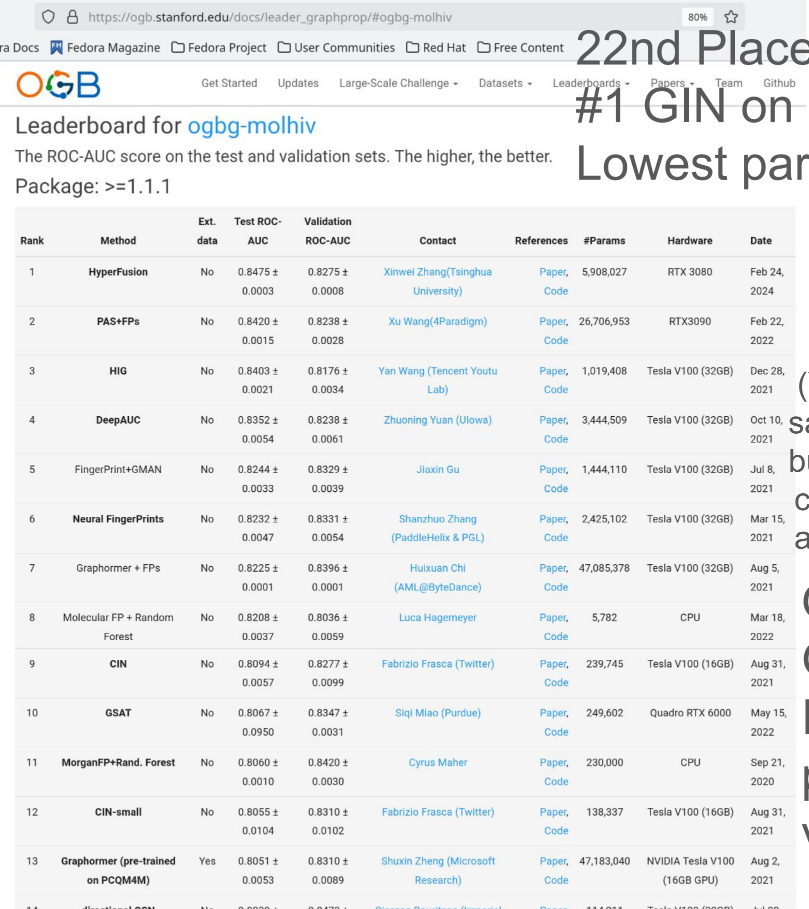
Note, there is variability.
This is seed 1, reported values were over 10 seeds and are similar but slightly worse than this on average, such that mean ROCAUC was 0.7835 +/- 0.0125 (mean +/- sample std, n=10) not 0.80 as shown above in detail.

Results (leaderboard)

22nd Place overall

#1 GIN on leaderboard

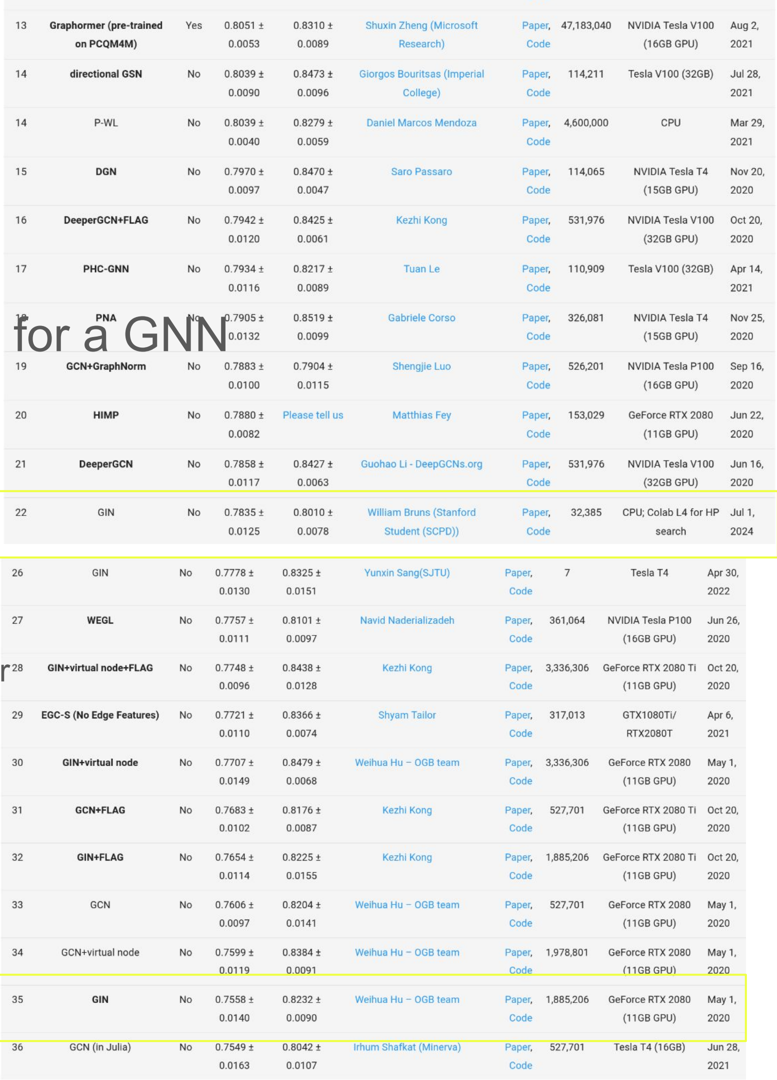
Lowest parameter count for a GNN



Leaderboard for ogbg-molhiv

The ROC-AUC score on the test and validation sets. The higher, the better.

Package: >=1.1.1



Rank	Method	Ext. data	Test ROC-AUC	Validation ROC-AUC	Contact	References	#Params	Hardware	Date
1	**HyperFusion**	No	0.8475 ± 0.0003	0.8275 ± 0.0008	Xinwei Zhang(Tsinghua University)	Paper, Code	5,908,027	RTX 3080	Feb 24, 2024
2	**PAS+FPs**	No	0.8420 ± 0.0015	0.8238 ± 0.0028	Xu Wang(4Paradigm)	Paper, Code	26,706,953	RTX3090	Feb 22, 2022
3	**HIG**	No	0.8403 ± 0.0021	0.8176 ± 0.0034	Yan Wang (Tencent Youtu Lab)	Paper, Code	1,019,408	Tesla V100 (32GB)	Dec 28, 2021
4	**DeepAUC**	No	0.8352 ± 0.0054	0.8238 ± 0.0061	Zhuoning Yuan (UIowa)	Paper, Code	3,444,509	Tesla V100 (32GB)	Oct 10, 2021
5	FingerPrint+GMAN	No	0.8244 ± 0.0033	0.8329 ± 0.0039	Jiaxin Gu	Paper, Code	1,444,110	Tesla V100 (32GB)	Jul 8, 2021
6	**Neural FingerPrints**	No	0.8232 ± 0.0047	0.8331 ± 0.0054	Shanzhuo Zhang (PaddleHelix & PGL)	Paper, Code	2,425,102	Tesla V100 (32GB)	Mar 15, 2021
7	Graphormer + FPs	No	0.8225 ± 0.0001	0.8396 ± 0.0001	Huixuan Chi (AML@ByteDance)	Paper, Code	47,085,378	Tesla V100 (32GB)	Aug 5, 2021
8	Molecular FP + Random Forest	No	0.8208 ± 0.0037	0.8036 ± 0.0059	Luca Hagemeyer	Paper, Code	5,782	CPU	Mar 18, 2022
9	**CIN**	No	0.8094 ± 0.0057	0.8277 ± 0.0099	Fabrizio Frasca (Twitter)	Paper, Code	239,745	Tesla V100 (16GB)	Aug 31, 2021
10	**GSAT**	No	0.8067 ± 0.0950	0.8347 ± 0.0031	Siqi Miao (Purdue)	Paper, Code	249,602	Quadro RTX 6000	May 15, 2022
11	**MorganFP+Rand. Forest**	No	0.8060 ± 0.0010	0.8420 ± 0.0030	Cyrus Maher	Paper, Code	230,000	CPU	Sep 21, 2020
12	**CIN-small**	No	0.8055 ± 0.0104	0.8310 ± 0.0102	Fabrizio Frasca (Twitter)	Paper, Code	138,337	Tesla V100 (16GB)	Aug 31, 2021
13	**Graphormer (pre-trained on PCQM4M)**	Yes	0.8051 ± 0.0053	0.8310 ± 0.0089	Shuxin Zheng (Microsoft Research)	Paper, Code	47,183,040	NVIDIA Tesla V100 (16GB GPU)	Aug 2, 2021
14	**directional GSN**	No	0.8039 ± 0.0090	0.8473 ± 0.0096	Giorgos Bouritsas (Imperial College)	Paper, Code	114,211	Tesla V100 (32GB)	Jul 28, 2021
14	P-WL	No	0.8039 ± 0.0040	0.8279 ± 0.0059	Daniel Marcos Mendoza	Paper, Code	4,600,000	CPU	Mar 29, 2021
15	**DGN**	No	0.7970 ± 0.0097	0.8470 ± 0.0047	Saro Passaro	Paper, Code	114,065	NVIDIA Tesla T4 (15GB GPU)	Nov 20, 2020
16	**DeeperGCN+FLAG**	No	0.7942 ± 0.0120	0.8425 ± 0.0061	Kezhi Kong	Paper, Code	531,976	NVIDIA Tesla V100 (32GB GPU)	Oct 20, 2020
17	**PHC-GNN**	No	0.7934 ± 0.0116	0.8217 ± 0.0089	Tuan Le	Paper, Code	110,909	Tesla V100 (32GB)	Apr 14, 2021
18	**PNA**	No	0.7905 ± 0.0132	0.8519 ± 0.0099	Gabriele Corso	Paper, Code	326,081	NVIDIA Tesla T4 (15GB GPU)	Nov 25, 2020
19	**GCN+GraphNorm**	No	0.7883 ± 0.0100	0.7904 ± 0.0115	Shengjie Luo	Paper, Code	526,201	NVIDIA Tesla P100 (16GB GPU)	Sep 16, 2020
20	**HIMP**	No	0.7880 ± 0.0082	Please tell us	Matthias Fey	Paper, Code	153,029	GeForce RTX 2080 (11GB GPU)	Jun 22, 2020
21	**DeeperGCN**	No	0.7858 ± 0.0117	0.8427 ± 0.0063	Guohao Li - DeepGCNs.org	Paper, Code	531,976	NVIDIA Tesla V100 (32GB GPU)	Jun 16, 2020
22	GIN	No	0.7835 ± 0.0125	0.8010 ± 0.0078	William Bruns (Stanford Student (SCPD))	Paper, Code	32,385	CPU; Colab L4 for HP search	Jul 1, 2024
26	GIN	No	0.7778 ± 0.0130	0.8325 ± 0.0151	Yunxin Sang(SJTU)	Paper, Code	7	Tesla T4	Apr 30, 2022
27	**WEGL**	No	0.7757 ± 0.0111	0.8101 ± 0.0097	Navid Naderializadeh	Paper, Code	361,064	NVIDIA Tesla P100 (16GB GPU)	Jun 26, 2020
28	**GIN+virtual node+FLAG**	No	0.7748 ± 0.0096	0.8438 ± 0.0128	Kezhi Kong	Paper, Code	3,336,306	GeForce RTX 2080 Ti (11GB GPU)	Oct 20, 2020
29	**EGC-S (No Edge Features)**	No	0.7721 ± 0.0110	0.8366 ± 0.0074	Shyam Tailor	Paper, Code	317,013	GTX1080Ti/ RTX2080T	Apr 6, 2021
30	**GIN+virtual node**	No	0.7707 ± 0.0149	0.8479 ± 0.0068	Weihua Hu – OGB team	Paper, Code	3,336,306	GeForce RTX 2080 (11GB GPU)	May 1, 2020
31	**GCN+FLAG**	No	0.7683 ± 0.0102	0.8176 ± 0.0087	Kezhi Kong	Paper, Code	527,701	GeForce RTX 2080 Ti (11GB GPU)	Oct 20, 2020
32	**GIN+FLAG**	No	0.7654 ± 0.0114	0.8225 ± 0.0155	Kezhi Kong	Paper, Code	1,885,206	GeForce RTX 2080 Ti (11GB GPU)	Oct 20, 2020
33	GCN	No	0.7606 ± 0.0097	0.8204 ± 0.0141	Weihua Hu – OGB team	Paper, Code	527,701	GeForce RTX 2080 (11GB GPU)	May 1, 2020
34	GCN+virtual node	No	0.7599 ± 0.0119	0.8384 ± 0.0091	Weihua Hu – OGB team	Paper, Code	1,978,801	GeForce RTX 2080 (11GB GPU)	May 1, 2020
35	**GIN**	No	0.7558 ± 0.0140	0.8232 ± 0.0090	Weihua Hu – OGB team	Paper, Code	1,885,206	GeForce RTX 2080 (11GB GPU)	May 1, 2020
36	GCN (in Julia)	No	0.7549 ± 0.0163	0.8042 ± 0.0107	Irhum Shafkat (Minerva)	Paper, Code	527,701	Tesla T4 (16GB)	Jun 28, 2021

(Yunxin Sang's says 7 parameters but is >50K confirmed with author and reported)

OGB team GIN Is 1.8M parameters vs our 32K

Future directions

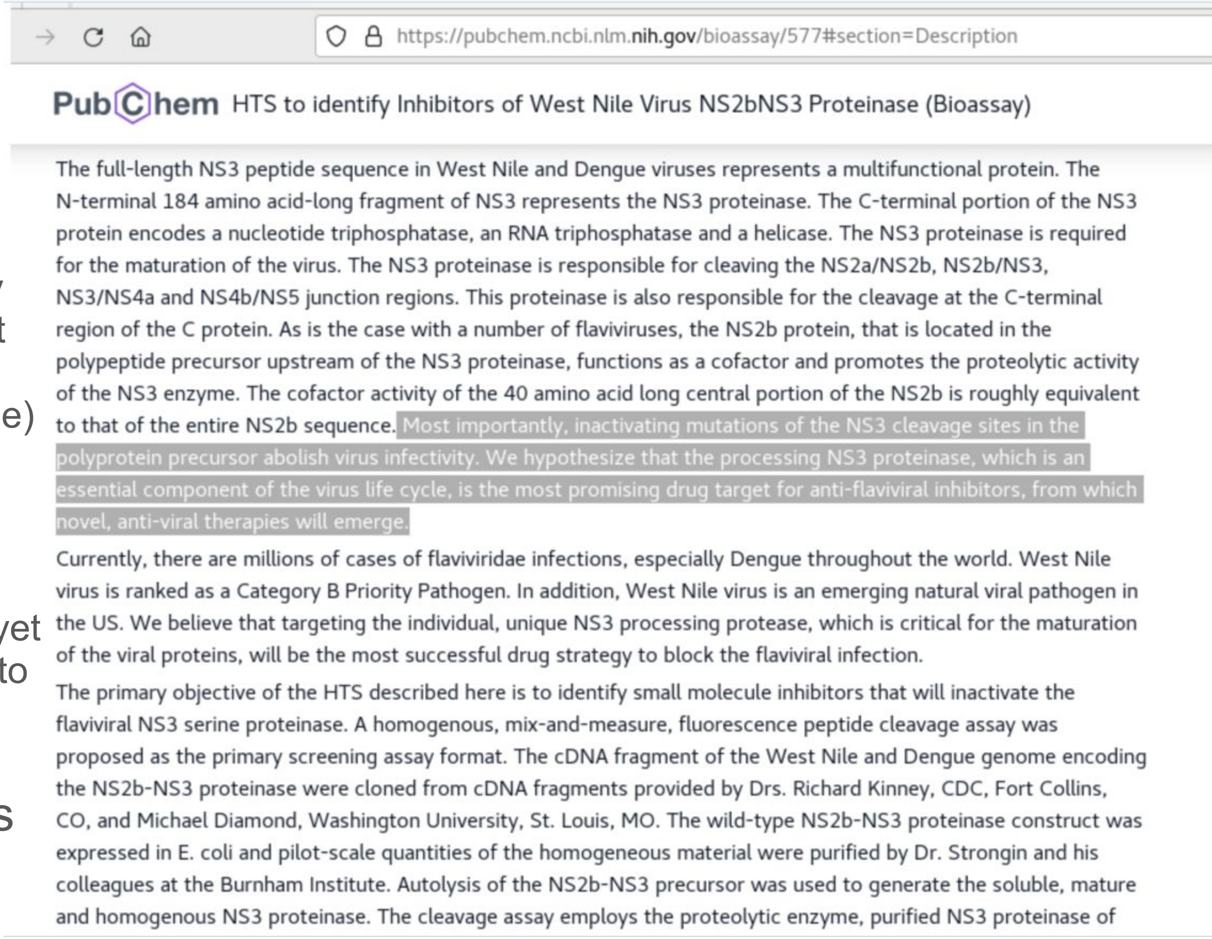
What I’m excited about:

West Nile Virus doesn’t have any approved antivirals! (unlike HIV which has many)

Could we speed up antiviral discovery by training a graph neural network to predict molecules that hit targets expected to inhibit the virus (e.g. NS2bNS3 proteinase) and then screen millions of molecules in e.g. the ZINC database for candidates?

I converted some PCBA data available (AID 577) into OGB format and started testing (not ready to release any results yet but gets some traction not SO dissimilar to say ogbg-molhiv benchmark).

If you are interested in collaborating on these problems please contact me at adde.animulis@gmail.com or https://github.com/willy-b



https://pubchem.ncbi.nlm.nih.gov/bioassay/577#section=Description
PubChem

HTS to identify Inhibitors of West Nile Virus NS2bNS3 Proteinase (Bioassay)

The full-length NS3 peptide sequence in West Nile and Dengue viruses represents a multifunctional protein. The N-terminal 184 amino acid-long fragment of NS3 represents the NS3 proteinase. The C-terminal portion of the NS3 protein encodes a nucleotide triphosphatase, an RNA triphosphatase and a helicase. The NS3 proteinase is required for the maturation of the virus. The NS3 proteinase is responsible for cleaving the NS2a/NS2b, NS2b/NS3, NS3/NS4a and NS4b/NS5 junction regions. This proteinase is also responsible for the cleavage at the C-terminal region of the C protein. As is the case with a number of flaviviruses, the NS2b protein, that is located in the polypeptide precursor upstream of the NS3 proteinase, functions as a cofactor and promotes the proteolytic activity of the NS3 enzyme. The cofactor activity of the 40 amino acid long central portion of the NS2b is roughly equivalent to that of the entire NS2b sequence. Most importantly, inactivating mutations of the NS3 cleavage sites in the polyprotein precursor abolish virus infectivity. We hypothesize that the processing NS3 proteinase, which is an essential component of the virus life cycle, is the most promising drug target for anti-flaviviral inhibitors, from which novel, anti-viral therapies will emerge.

Currently, there are millions of cases of flaviviridae infections, especially Dengue throughout the world. West Nile virus is ranked as a Category B Priority Pathogen. In addition, West Nile virus is an emerging natural viral pathogen in the US. We believe that targeting the individual, unique NS3 processing protease, which is critical for the maturation of the viral proteins, will be the most successful drug strategy to block the flaviviral infection.

The primary objective of the HTS described here is to identify small molecule inhibitors that will inactivate the flaviviral NS3 serine proteinase. A homogenous, mix-and-measure, fluorescence peptide cleavage assay was proposed as the primary screening assay format. The cDNA fragment of the West Nile and Dengue genome encoding the NS2b-NS3 proteinase were cloned from cDNA fragments provided by Drs. Richard Kinney, CDC, Fort Collins, CO, and Michael Diamond, Washington University, St. Louis, MO. The wild-type NS2b-NS3 proteinase construct was expressed in E. coli and pilot-scale quantities of the homogeneous material were purified by Dr. Strongin and his colleagues at the Burnham Institute. Autolysis of the NS2b-NS3 precursor was used to generate the soluble, mature and homogenous NS3 proteinase. The cleavage assay employs the proteolytic enzyme, purified NS3 proteinase of

Future directions

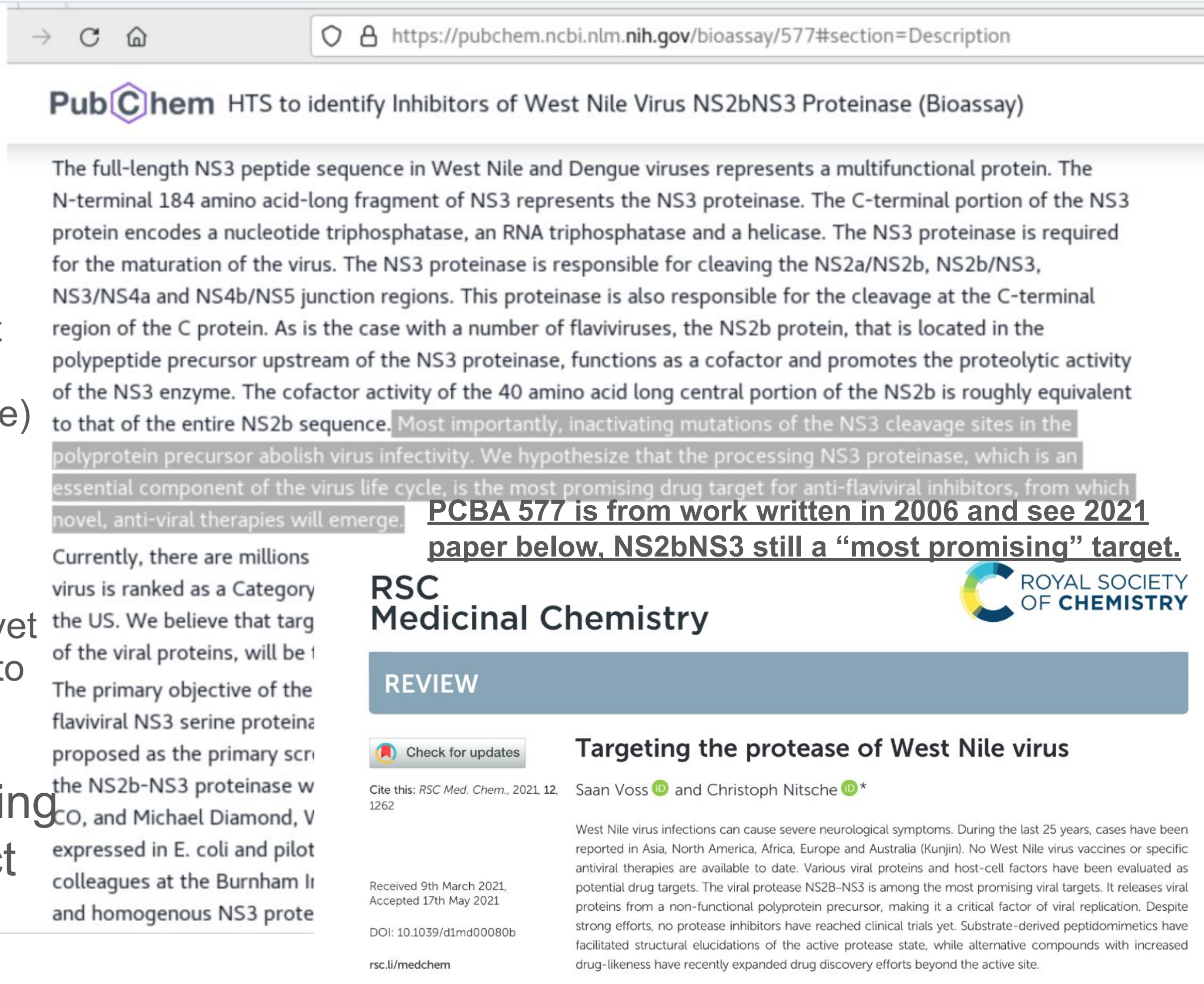
What I’m excited about:

West Nile Virus doesn’t have any approved antivirals! (unlike HIV which has many)

Could we speed up antiviral discovery by training a graph neural network to predict molecules that hit targets expected to inhibit the virus (e.g. NS2bNS3 proteinase) and then screen millions of molecules in e.g. the ZINC database for candidates?

I converted some PCBA data available (AID 577) into OGB format and started testing (not ready to release any results yet but gets some traction not SO dissimilar to say ogbg-molhiv benchmark).

If you are interested in collaborating on these problems please contact me at adde.animulis@gmail.com or https://github.com/willy-b



https://pubchem.ncbi.nlm.nih.gov/bioassay/577#section=Description

PubChem HTS to identify Inhibitors of West Nile Virus NS2bNS3 Proteinase (Bioassay)

The full-length NS3 peptide sequence in West Nile and Dengue viruses represents a multifunctional protein. The N-terminal 184 amino acid-long fragment of NS3 represents the NS3 proteinase. The C-terminal portion of the NS3 protein encodes a nucleotide triphosphatase, an RNA triphosphatase and a helicase. The NS3 proteinase is required for the maturation of the virus. The NS3 proteinase is responsible for cleaving the NS2a/NS2b, NS2b/NS3, NS3/NS4a and NS4b/NS5 junction regions. This proteinase is also responsible for the cleavage at the C-terminal region of the C protein. As is the case with a number of flaviviruses, the NS2b protein, that is located in the polypeptide precursor upstream of the NS3 proteinase, functions as a cofactor and promotes the proteolytic activity of the NS3 enzyme. The cofactor activity of the 40 amino acid long central portion of the NS2b is roughly equivalent to that of the entire NS2b sequence. Most importantly, inactivating mutations of the NS3 cleavage sites in the polyprotein precursor abolish virus infectivity. We hypothesize that the processing NS3 proteinase, which is an essential component of the virus life cycle, is the most promising drug target for anti-flaviviral inhibitors, from which novel, anti-viral therapies will emerge.

Currently, there are millions
virus is ranked as a Category
the US. We believe that targ
of the viral proteins, will be t

The primary objective of the
flaviviral NS3 serine proteina
proposed as the primary scr
the NS2b-NS3 proteinase w
CO, and Michael Diamond, V
expressed in E. coli and pilot
colleagues at the Burnham Ir
and homogenous NS3 prote

PCBA 577 is from work written in 2006 and see 2021 paper below, NS2bNS3 still a “most promising” target.

RSC Medicinal Chemistry

ROYAL SOCIETY OF CHEMISTRY

REVIEW

Check for updates

Cite this: *RSC Med. Chem.*, 2021, 12, 1262

Received 9th March 2021,
Accepted 17th May 2021

DOI: 10.1039/d1md00080b

rsc.li/medchem

Targeting the protease of West Nile virus

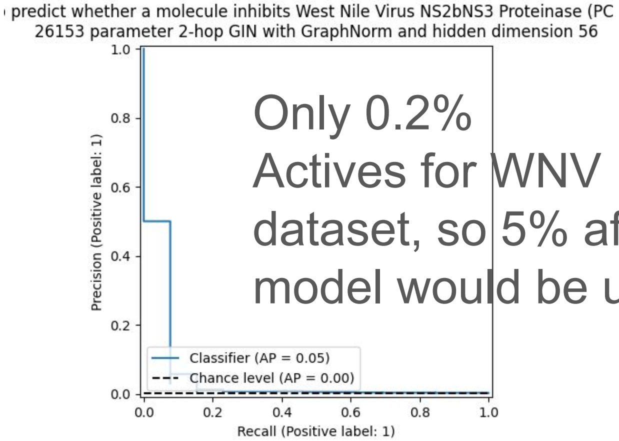
Saan Voss and Christoph Nitsche*

West Nile virus infections can cause severe neurological symptoms. During the last 25 years, cases have been reported in Asia, North America, Africa, Europe and Australia (Kunjin). No West Nile virus vaccines or specific antiviral therapies are available to date. Various viral proteins and host-cell factors have been evaluated as potential drug targets. The viral protease NS2B–NS3 is among the most promising viral targets. It releases viral proteins from a non-functional polyprotein precursor, making it a critical factor of viral replication. Despite strong efforts, no protease inhibitors have reached clinical trials yet. Substrate-derived peptidomimetics have facilitated structural elucidations of the active protease state, while alternative compounds with increased drug-likeness have recently expanded drug discovery efforts beyond the active site.

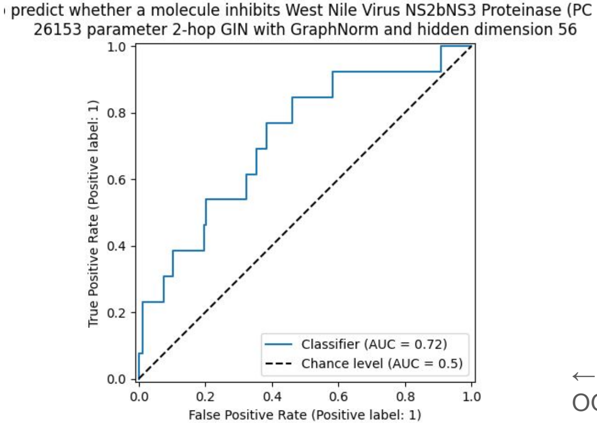
WNV NS2bNS3

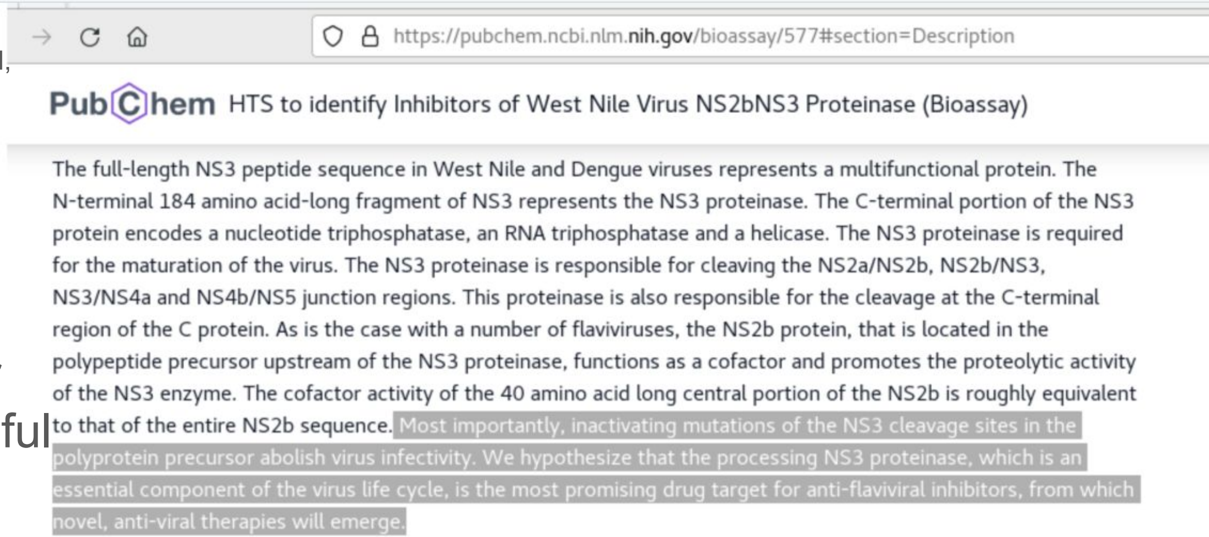
EARLY RESULTS! VALID SET USED FOR EARLY STOPPING! LOW SAMPLE SIZE!

Sneak peek! On validation looking pretty good, **but we used validation for early stopping!**



Only 0.2% Actives for WNV dataset, so 5% after model would be useful




https://pubchem.ncbi.nlm.nih.gov/bioassay/577#section=Description
PubChem

HTS to identify Inhibitors of West Nile Virus NS2bNS3 Proteinase (Bioassay)

The full-length NS3 peptide sequence in West Nile and Dengue viruses represents a multifunctional protein. The N-terminal 184 amino acid-long fragment of NS3 represents the NS3 proteinase. The C-terminal portion of the NS3 protein encodes a nucleotide triphosphatase, an RNA triphosphatase and a helicase. The NS3 proteinase is required for the maturation of the virus. The NS3 proteinase is responsible for cleaving the NS2a/NS2b, NS2b/NS3, NS3/NS4a and NS4b/NS5 junction regions. This proteinase is also responsible for the cleavage at the C-terminal region of the C protein. As is the case with a number of flaviviruses, the NS2b protein, that is located in the polypeptide precursor upstream of the NS3 proteinase, functions as a cofactor and promotes the proteolytic activity of the NS3 enzyme. The cofactor activity of the 40 amino acid long central portion of the NS2b is roughly equivalent to that of the entire NS2b sequence. Most importantly, inactivating mutations of the NS3 cleavage sites in the polyprotein precursor abolish virus infectivity. We hypothesize that the processing NS3 proteinase, which is an essential component of the virus life cycle, is the most promising drug target for anti-flaviviral inhibitors, from which novel, anti-viral therapies will emerge.

> "The first principle is that you must not fool yourself--and you are the easiest person to fool."
> "I would like to add something that's not essential to the science, but something I kind of believe, which is that you should not fool the layman when you're talking as a scientist....bending over backwards to show how you are maybe wrong, that you ought to have when acting as a scientist. And this is our responsibility as scientists, certainly to other scientists, and I think to laymen."
>
> - Richard P. Feynman (https://speakola.com/grad/richard-feynman-caltech-1974)

← Typical validation results with current hyperparameters. (Test may vary and will probably be lower. Earlier results for OGB HIV reported were holdout test performance - that had already finished this stage.)

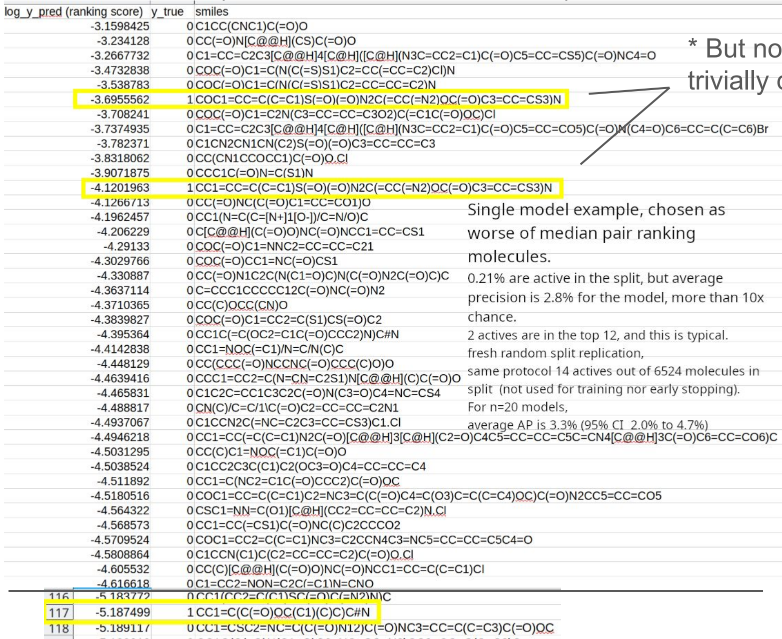
WNV NS2bNS3

EARLY RESULTS! LOW SAMPLE SIZE, INTERPRET WITH CAUTION

Sneak peek!

Avoid fooling ourselves, use additional holdout data, separate from validation used for early stopping.
First run with holdout from training split gave 0.58 ROCAUC with 4% AP (both statistically significant p<0.05),
But does it replicate? With twenty more random holdouts, same protocol, it replicates, within 1% ROCAUC and 1% AP, see below.
(Is it an artifact of our split? Unlikely for AP, but still TBD on test split and after that also random splits)

Showing median model detail from replication with n=20 models
Holdout had 14/6524 active molecules (0.2%) (separate from train/valid)
2 actives in top 12 as ranked by model (p<0.03 to get >0 by binomial)
3 actives by rank 117 as rank by median model (p<0.003 to get >= 3 actives by this rank)
2.8% AP overall (3.3% mean AP n=20 models)



	log_y_pred (ranking score)	y_true	smiles
	-3.1598425	0	C1CC(CNC1)C(=O)O
	-3.234128	0	CC(=O)N[C@@H](CS)C(=O)O
	-3.2667732	0	C1=CC=C2C3[C@@H]4[C@H]([C@H](N3C=CC2=C1)C(=O)C5=CC=CS5)C(=O)NC4=O
	-3.4732838	0	COC(=O)C1=C(N(C(=S)S1)C2=CC(=CC=C2)Cl)N
	-3.538783	0	COC(=O)C1=C(N(C(=S)S1)C2=CC=CC=C2)N
	-3.6955562	1	COC1=CC=C(C=C1)S(=O)(=O)N2C(=CC(=N2)OC(=O)C3=CC=CS3)N
	-3.708241	0	COC(=O)C1=C2N(C3=CC=CC=C3O2)C(=C1C(=O)OC)Cl
	-3.7374935	0	C1=CC=C2C3[C@@H]4[C@H]([C@H](N3C=CC2=C1)C(=O)C5=CC=CO5)C(=O)N(C4=O)C6=CC=C(C=C6)Br
	-3.782371	0	C1CN2CN1CN(C2)S(=O)(=O)C3=CC=CC=C3
	-3.8318062	0	CC(CN1CCOCC1)C(=O)O.Cl
	-3.9071875	0	CCC1C(=O)N=C(S1)N
	-4.1201963	1	CC1=CC=C(C=C1)S(=O)(=O)N2C(=CC(=N2)OC(=O)C3=CC=CS3)N
	-4.1266713	0	CC(=O)NC(C(=O)C1=CC=CO1)O
	-4.1962457	0	CC1(N=C(C=[N+]1[O-])/C=N/O)C
	-4.206229	0	C[C@@H](C(=O)O)NC(=O)NCC1=CC=CS1
	-4.29133	0	COC(=O)C1=NNC2=CC=CC=C21
	-4.3029766	0	COC(=O)CC1=NC(=O)CS1
	-4.330887	0	CC(=O)N1C2C(N(C1=O)C)N(C(=O)N2C(=O)C)C
	-4.3637114	0	C=CCC1CCCCC12C(=O)NC(=O)N2
	-4.3710365	0	CC(C)OCC(CN)O
	-4.3839827	0	COC(=O)C1=CC2=C(S1)CS(=O)C2
	-4.395364	0	CC1C(=C(OC2=C1C(=O)CCC2)N)C#N
	-4.4142838	0	CC1=NOC(=C1)/N=C/N(C)C
	-4.448129	0	CC(CCC(=O)NCCNC(=O)CCC(C)O)O
	-4.4639416	0	CCC1=CC2=C(N=CN=C2S1)N[C@@H](C)C(=O)O
	-4.465831	0	C1C2C=CC1C3C2C(=O)N(C3=O)C4=NC=CS4
	-4.488817	0	CN(C)/C=C/1\C(=O)C2=CC=CC=C2N1
	-4.4937067	0	C1CCN2C(=NC=C2C3=CC=CS3)C1.Cl
	-4.4946218	0	CC1=CC(=C(C=C1)N2C(=O)[C@@H]3[C@H](C2=O)C4C5=CC=CC=C5C=CN4[C@@H]3C(=O)C6=CC=CO6)C
	-4.5031295	0	CC(C)C1=NOC(=C1)C(=O)O
	-4.5038524	0	C1CC2C3C(C1)C2(OC3=O)C4=CC=CC=C4
	-4.511892	0	CC1=C(NC2=C1C(=O)CCC2)C(=O)OC
	-4.5180516	0	COC1=CC=C(C=C1)C2=NC3=C(C(=O)C4=C(O3)C=C(C=C4)OC)C(=O)N2CC5=CC=CO5
	-4.564322	0	CSC1=NN=C(O1)[C@H](CC2=CC=CC=C2)N.Cl
	-4.568573	0	CC1=CC(=CS1)C(=O)NC(C)C2CCCO2
	-4.5709524	0	COC1=CC2=C(C=C1)NC3=C2CCN4C3=NC5=CC=CC=C5C4=O
	-4.5808864	0	C1CCN(C1)C(C2=CC=CC=C2)C(=O)O.Cl
	-4.605532	0	CC(C)[C@@H](C(=O)O)NC(=O)NCC1=CC=C(C=C1)Cl
	-4.616618	0	C1=CC2=NON=C2C(=C1)N=CNO
116	-5.183772	0	CC1(CC2=C(C1)SC(=O)C(=N2)N)C
117	-5.187499	1	CC1=C(C(=O)OC(C1)(C)C)C#N
118	-5.189117	0	CC1=CSC2=NC=C(C(=O)N12)C(=O)NC3=CC=C(C=C3)C(=O)OC

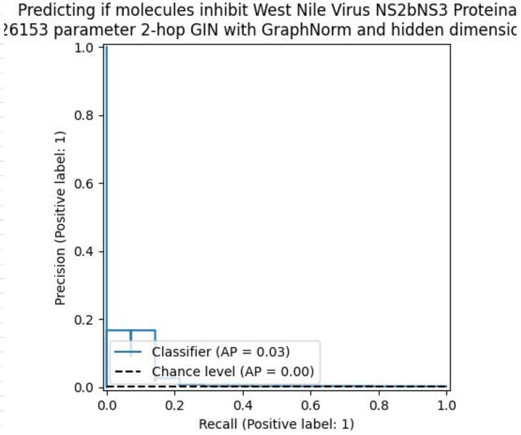
* But note the first two actives are trivially different from one another

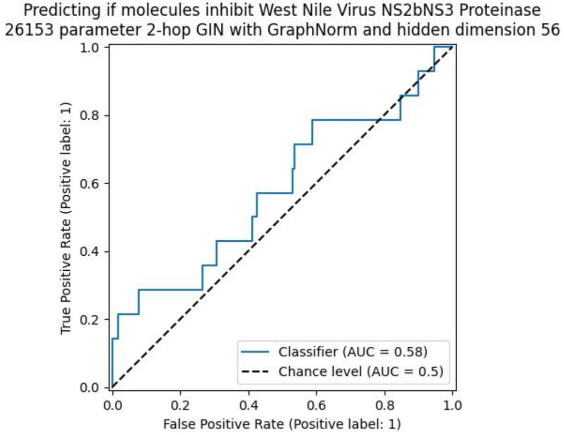
Single model example, chosen as worse of median pair ranking molecules.

0.21% are active in the split, but average precision is 2.8% for the model, more than 10x chance.

2 actives are in the top 12, and this is typical.
fresh random split replication,
same protocol 14 actives out of 6524 molecules in split (not used for training nor early stopping).
For n=20 models,
average AP is 3.3% (95% CI 2.0% to 4.7%)

First holdout run (n=20 seeds, 20 random holdouts from fixed training split)
ROCAUC 58% (95% CI 55% to 62%), 4.1% AP (95% CI 2.7% to 5.6% AP).
Second holdout replication (n=20) with fresh random split (random holdout from train) is consistent
ROCAUC 57% (95% CI 54% to 60%), 3.3% AP (95% CI 2.0% to 4.7%)
Different random holdout from train split entirely

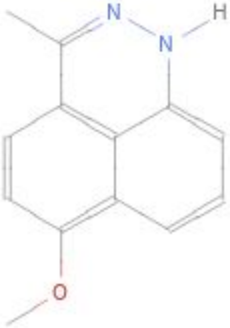


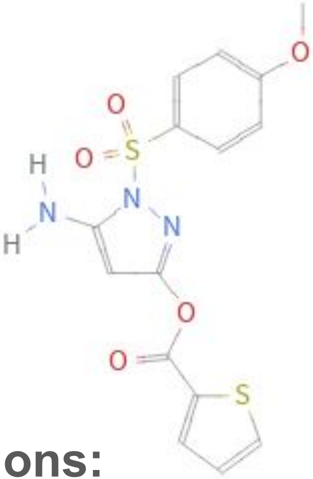


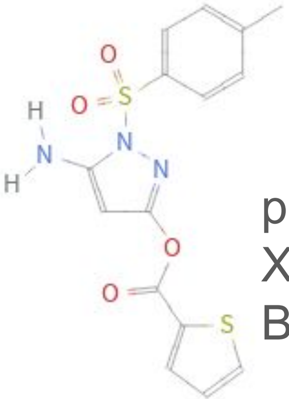
WNV NS2bNS3

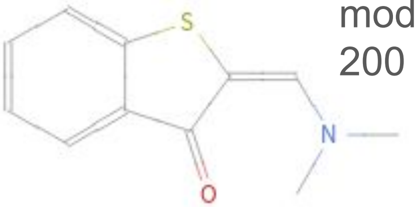
EARLY RESULTS! LOW SAMPLE SIZE, INTERPRET WITH CAUTION

Sneak peek!





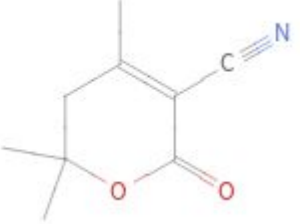




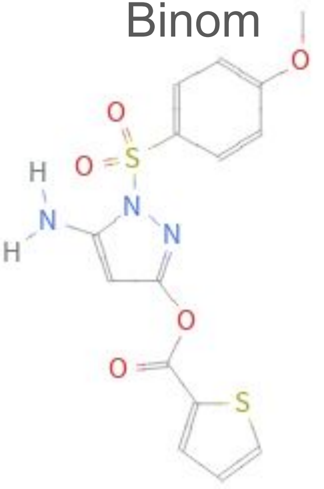
Upper median model finds in first 200 results

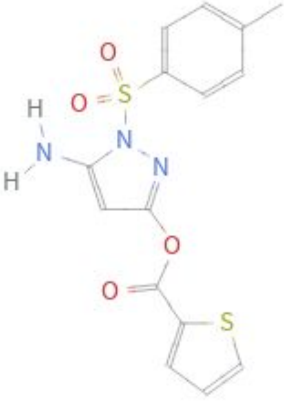
p<0.0013
X >= 4
Binom

Lower median Model (last slide) finds in first 200 results



p<0.0095
X >= 3
Binom





Some Limitations:

- Only 119 actives to split (train/val/gen/test)
- Not a scaffold split, some molecules trivially different
- OGBG-MOLHIV used a scaffold split to require generalization

Would be good to pair with a domain expert to get a scaffold split of larger input dataset for WNV

Low data WNV (a Flavivirus) -> more data for Flaviviruses

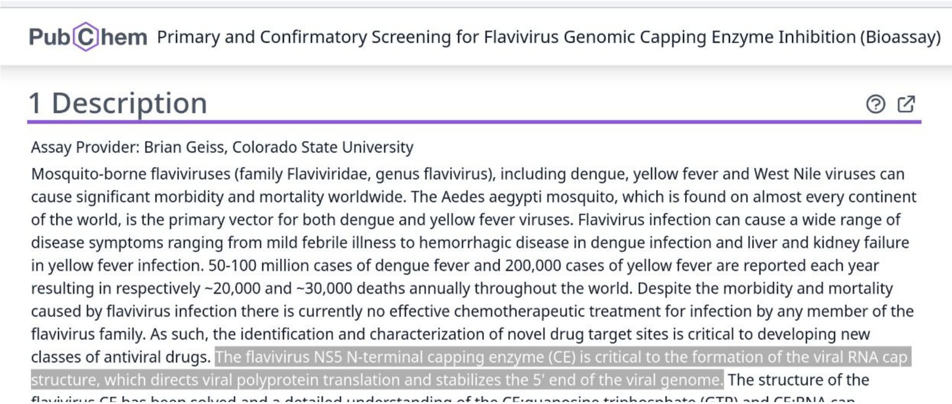
Problem:
Could not scaffold split WNV PCBA 577 due to insufficient data size and not enough data to train and test GNN.

Solution:
Pivot to Flavivirus level data instead of WNV specific: PCBA 588689 (targeting common enzyme to WNV, Dengue, Yellow Fever, and other Flaviviruses).

338.9K molecules, ~0.3% active (1013 actives), enough to use Bemis-Murcko scaffold split and actually train a GNN like we did for ogbg-molhiv.

Report at:
https://raw.githubusercontent.com/willy-b/tiny-GIN-for-WNV/c470235b30f3e840e70ed9af126b78879be47b3d/gnns-to-predict-flaviviral-genomic-capping-enzyme-inhibition.pdf

Convert
To
OGB
Format
And
Scaffold
Split

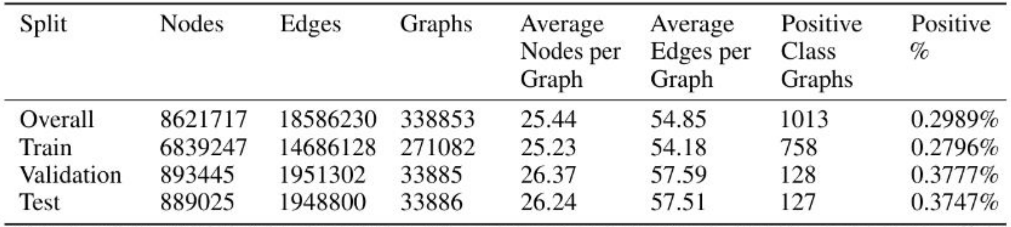


PubChem Primary and Confirmatory Screening for Flavivirus Genomic Capping Enzyme Inhibition (Bioassay)

1 Description

Assay Provider: Brian Geiss, Colorado State University

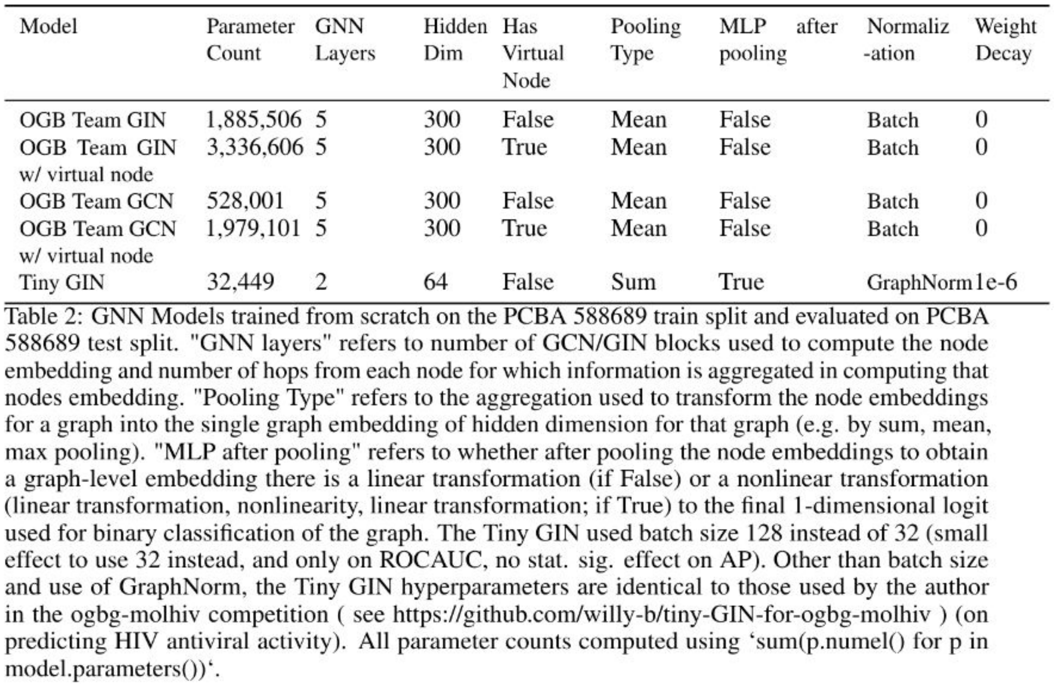
Mosquito-borne flaviviruses (family Flaviviridae, genus flavivirus), including dengue, yellow fever and West Nile viruses can cause significant morbidity and mortality worldwide. The Aedes aegypti mosquito, which is found on almost every continent of the world, is the primary vector for both dengue and yellow fever viruses. Flavivirus infection can cause a wide range of disease symptoms ranging from mild febrile illness to hemorrhagic disease in dengue infection and liver and kidney failure in yellow fever infection. 50-100 million cases of dengue fever and 200,000 cases of yellow fever are reported each year resulting in respectively ~20,000 and ~30,000 deaths annually throughout the world. Despite the morbidity and mortality caused by flavivirus infection there is currently no effective chemotherapeutic treatment for infection by any member of the flavivirus family. As such, the identification and characterization of novel drug target sites is critical to developing new classes of antiviral drugs. The flavivirus NS5 N-terminal capping enzyme (CE) is critical to the formation of the viral RNA cap structure, which directs viral polyprotein translation and stabilizes the 5' end of the viral genome. The structure of the flavivirus CE has been solved and a detailed understanding of the CE:guanosine triphosphate (GTP) and CE:RNA cap



Split	Nodes	Edges	Graphs	Average Nodes per Graph	Average Edges per Graph	Positive Class Graphs	Positive %
Overall	8621717	18586230	338853	25.44	54.85	1013	0.2989%
Train	6839247	14686128	271082	25.23	54.18	758	0.2796%
Validation	893445	1951302	33885	26.37	57.59	128	0.3777%
Test	889025	1948800	33886	26.24	57.51	127	0.3747%

Table 1: PCBA 588689 Dataset statistics. The graphs are split by Bemis-Murcko molecular scaffold split, sorted by descending scaffold cardinality (most common scaffolds in train, then validation, then test) with ties broken by random ordering. The test set is all unique scaffolds. "Positive class graphs" in this binary classification problem refer to the Active Molecules, i.e. those that would inhibit the Flaviviral Genomic Capping Enzyme. The scaffold splitting ensures that the Active Molecules in the validation and test splits are not structurally similar to any in the training set.

OGB baseline models and Tiny GIN for Flaviviral dataset

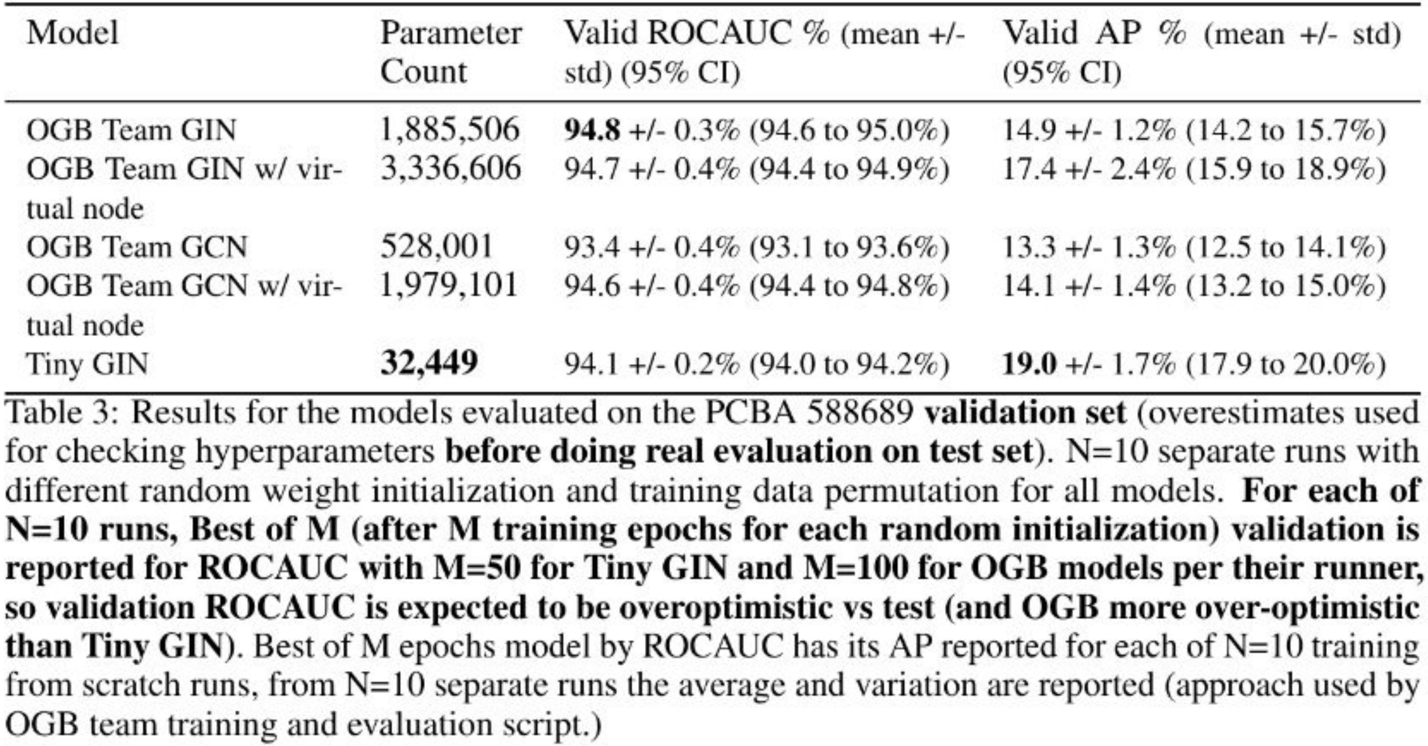


Model	Parameter Count	GNN Layers	Hidden Dim	Has Virtual Node	Pooling Type	MLP after pooling	Normaliz-ation	Weight Decay
OGB Team GIN	1,885,506	5	300	False	Mean	False	Batch	0
OGB Team GIN w/ virtual node	3,336,606	5	300	True	Mean	False	Batch	0
OGB Team GCN	528,001	5	300	False	Mean	False	Batch	0
OGB Team GCN w/ virtual node	1,979,101	5	300	True	Mean	False	Batch	0
Tiny GIN	32,449	2	64	False	Sum	True	GraphNorm	1e-6

Table 2: GNN Models trained from scratch on the PCBA 588689 train split and evaluated on PCBA 588689 test split. "GNN layers" refers to number of GCN/GIN blocks used to compute the node embedding and number of hops from each node for which information is aggregated in computing that nodes embedding. "Pooling Type" refers to the aggregation used to transform the node embeddings for a graph into the single graph embedding of hidden dimension for that graph (e.g. by sum, mean, max pooling). "MLP after pooling" refers to whether after pooling the node embeddings to obtain a graph-level embedding there is a linear transformation (if False) or a nonlinear transformation (linear transformation, nonlinearity, linear transformation; if True) to the final 1-dimensional logit used for binary classification of the graph. The Tiny GIN used batch size 128 instead of 32 (small effect to use 32 instead, and only on ROCAUC, no stat. sig. effect on AP). Other than batch size and use of GraphNorm, the Tiny GIN hyperparameters are identical to those used by the author in the ogbg-molhiv competition (see https://github.com/willy-b/tiny-GIN-for-ogbg-molhiv) (on predicting HIV antiviral activity). All parameter counts computed using ‘sum(p.numel() for p in model.parameters())‘.

Same Tiny GIN as used for ogbg-molhiv compared to OGB baseline GIN/GCNs

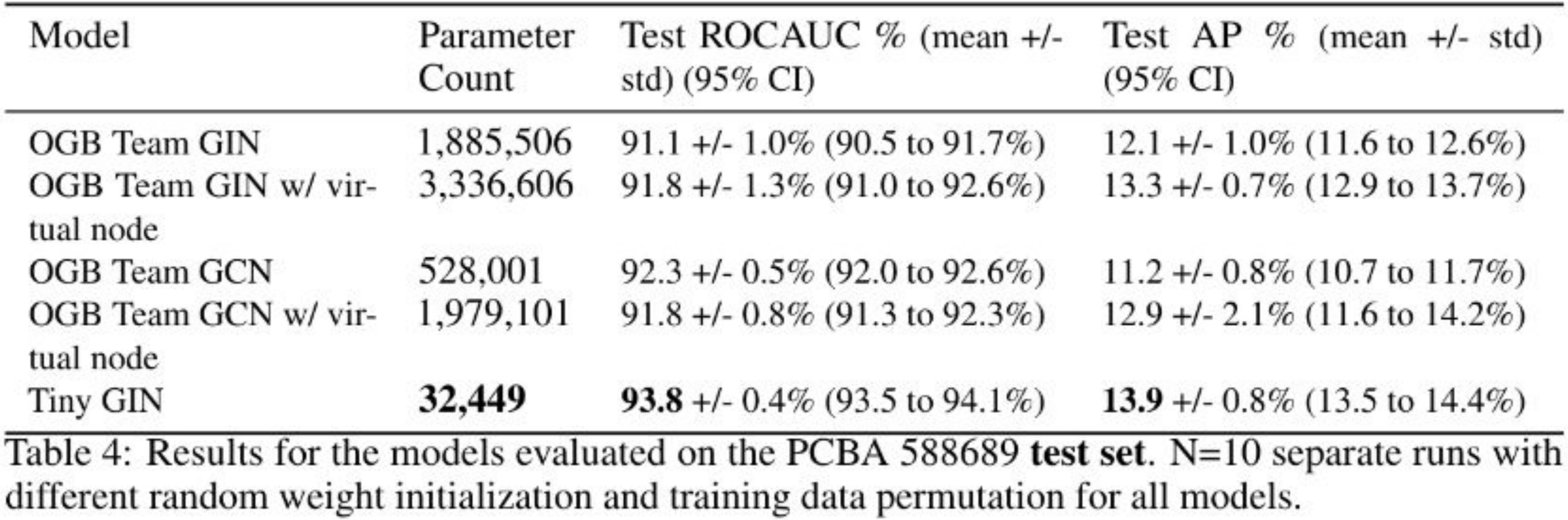
Flaviviral Genomic Capping Enzyme inhibition (checking hyperparams on valid set)



Model	Parameter Count	Valid ROCAUC % (mean +/- std) (95% CI)	Valid AP % (mean +/- std) (95% CI)
OGB Team GIN	1,885,506	**94.8** +/- 0.3% (94.6 to 95.0%)	14.9 +/- 1.2% (14.2 to 15.7%)
OGB Team GIN w/ virtual node	3,336,606	94.7 +/- 0.4% (94.4 to 94.9%)	17.4 +/- 2.4% (15.9 to 18.9%)
OGB Team GCN	528,001	93.4 +/- 0.4% (93.1 to 93.6%)	13.3 +/- 1.3% (12.5 to 14.1%)
OGB Team GCN w/ virtual node	1,979,101	94.6 +/- 0.4% (94.4 to 94.8%)	14.1 +/- 1.4% (13.2 to 15.0%)
Tiny GIN	**32,449**	94.1 +/- 0.2% (94.0 to 94.2%)	**19.0** +/- 1.7% (17.9 to 20.0%)

Table 3: Results for the models evaluated on the PCBA 588689 **validation set** (overestimates used for checking hyperparameters **before doing real evaluation on test set**). N=10 separate runs with different random weight initialization and training data permutation for all models. **For each of N=10 runs, Best of M (after M training epochs for each random initialization) validation is reported for ROCAUC with M=50 for Tiny GIN and M=100 for OGB models per their runner, so validation ROCAUC is expected to be overoptimistic vs test (and OGB more over-optimistic than Tiny GIN**). Best of M epochs model by ROCAUC has its AP reported for each of N=10 training from scratch runs, from N=10 separate runs the average and variation are reported (approach used by OGB team training and evaluation script.)

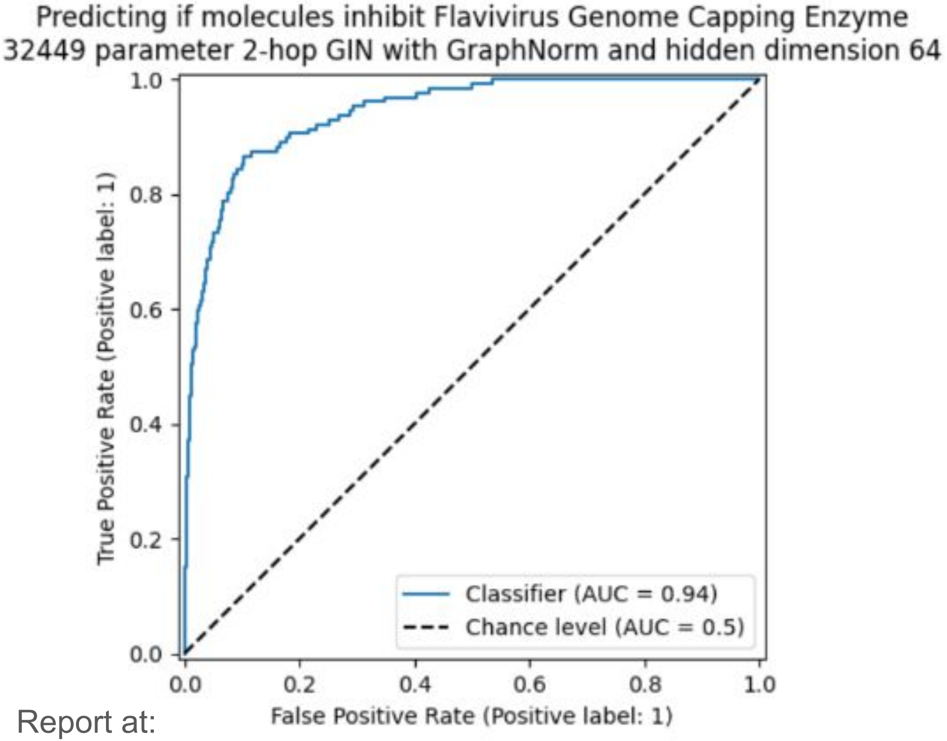
Flaviviral Genomic Capping Enzyme inhibition prediction **test results**



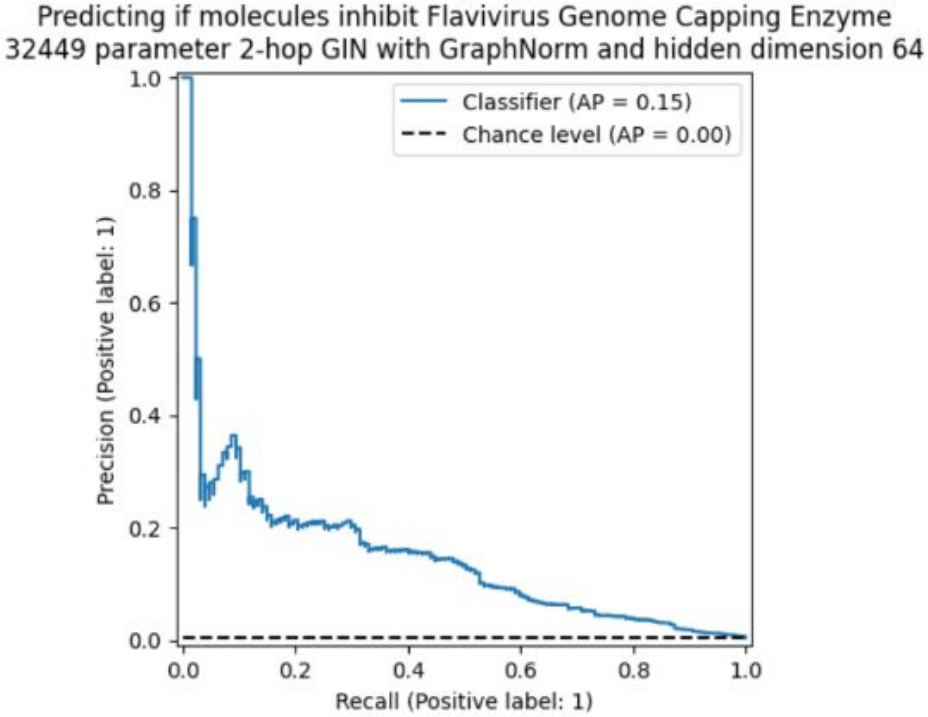
Model	Parameter Count	Test ROCAUC % (mean +/- std) (95% CI)	Test AP % (mean +/- std) (95% CI)
OGB Team GIN	1,885,506	91.1 +/- 1.0% (90.5 to 91.7%)	12.1 +/- 1.0% (11.6 to 12.6%)
OGB Team GIN w/ virtual node	3,336,606	91.8 +/- 1.3% (91.0 to 92.6%)	13.3 +/- 0.7% (12.9 to 13.7%)
OGB Team GCN	528,001	92.3 +/- 0.5% (92.0 to 92.6%)	11.2 +/- 0.8% (10.7 to 11.7%)
OGB Team GCN w/ virtual node	1,979,101	91.8 +/- 0.8% (91.3 to 92.3%)	12.9 +/- 2.1% (11.6 to 14.2%)
Tiny GIN	**32,449**	**93.8** +/- 0.4% (93.5 to 94.1%)	**13.9** +/- 0.8% (13.5 to 14.4%)

Table 4: Results for the models evaluated on the PCBA 588689 **test set**. N=10 separate runs with different random weight initialization and training data permutation for all models.

Test ROC and PRC for FGCE by Tiny GIN (seed 1, of 10)



Predicting if molecules inhibit Flavivirus Genome Capping Enzyme
32449 parameter 2-hop GIN with GraphNorm and hidden dimension 64



Predicting if molecules inhibit Flavivirus Genome Capping Enzyme
32449 parameter 2-hop GIN with GraphNorm and hidden dimension 64

Report at:
https://raw.githubusercontent.com/willy-b/tiny-GIN-for-WNV/c470235b30f3e840e70ed9af126b78879be47b3d/gnns-to-predict-flaviviral-genomic-capping-enzyme-inhibition.pdf

References

Hu, Weihua and Fey, Matthias and Zitnik, Marinka and Dong, Yuxiao and Ren, Hongyu and Liu, Bowen and Catasta, Michele and Leskovec, Jure. Open Graph Benchmark: Datasets for Machine Learning on Graphs. arXiv preprint arXiv:2005.00687, 2020.

Wu, Zhenqin and Ramsundar, Bharath and Feinberg, Evan N and Gomes, Joseph and Geniesse, Caleb and SPappu, Aneesh and Leswing, Karl and Pande, Vijay. Moleculenet: a benchmark for molecular machine learning. Chemical Science, 9(2):513–530, 2018.

Fey, Matthias and Lenssen, Jan E. Fast Graph Representation Learning with PyTorch Geometric. ICLR Workshop on Representation Learning on Graphs and Manifolds, 2019. (Graph Isomorphism Network (GIN) implementation used)

Xu, Keyulu and Hu, Weihua and Leskovec, Jure and Jegelka, Stefanie. How Powerful Are Graph Neural Networks? International Conference on Learning Representations, 2019. https://openreview.net/forum?id=ryGs6iA5Km , https://arxiv.org/pdf/1810.00826 . (Graph Isomorphism Network (GIN) original paper)

PubChem [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004-. PubChem Bioassay Record for AID 577, HTS to identify Inhibitors of West Nile Virus NS2bNS3 Proteinase , Source: University of Pittsburgh Molecular Library Screening Center; [cited 2024 Nov. 23]. Available from: https://pubchem.ncbi.nlm.nih.gov/bioassay/577

PubChem [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004-. PubChem Bioassay Record for AID 588689, Primary and Confirmatory Screening for Flavivirus Genomic Capping Enzyme Inhibition, Source: Southern Research Specialized Biocontainment Screening Center; [cited 2025 Mar. 21]. Available from: https://pubchem.ncbi.nlm.nih.gov/bioassay/588689