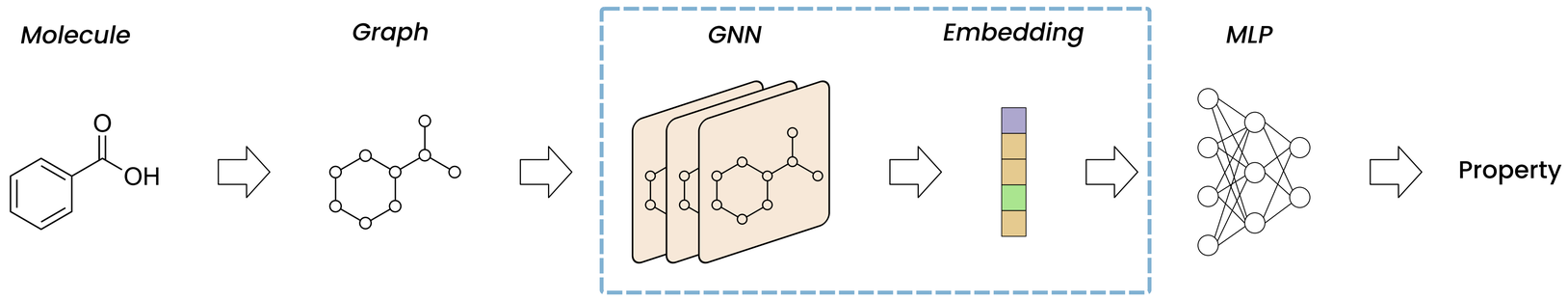
Bio-Isostere Guided Molecular Property Prediction

Anatol Ehrlich, Nils M. Kriege, Christoph Flamm

University of Vienna

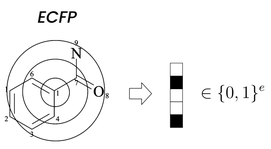
The Problem



We want to improve this!

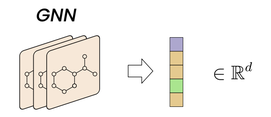
ECFP

- Generate good embeedings
- Are not trainable



GNN

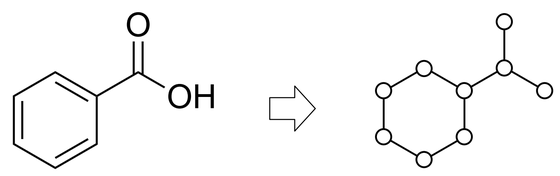
- Generate worse default embeddings than ECFPs
- Are end-to-end trainable



0. Motivation

- Improve **Molecular Property Predictions**
- Create better **embeddings** of molecules using **Graph Neural Networks (GNNs)**
- Make use of large amounts of **unlabeled data**

1. Molecules as Graphs



- Abstract **Molecules as Graphs**
- Enrich with **Node & Edge Features**

Node Features

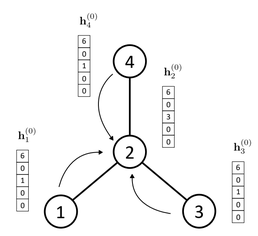
atomic_num,
chirality,
degree,
formal charge
num_hs
num_radical_e
hybridization,
is_aromatic,
is_in_ring

Edge Features

bond_type,
stereoinfo,
is_conjugated

The Approach

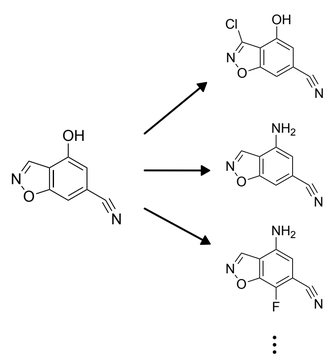
2. Graph Neural Networks(GNN)



Message Passing Framework

- $\mathbf{m}_u^{(i)} = \text{Aggregate}^{(i)}\left(\{\!\{\mathbf{h}_v^{(i)} \mid v \in \mathcal{N}(u)\}\!\}\right)$
- $\mathbf{h}_u^{(i+1)} = \text{Combine}^{(i)}\left(\mathbf{h}_u^{(i)}, \mathbf{m}_u^{(i)}\right)$

3. Data Augmentation



Bio-Isosteres

10M Molecules

label	smiles
0	N#Cc1cc(O)c2oncc2c1
1	Cc1nccc(C=CC(C)Cl)n1
2	O=PCOCC(COCP=O)OCP=O
3	COc1ccc(Br)cc1SC
4	CSC1C(=O)N=NC=C1Cl
5	COCCNc1nccc2occc12
6	O=C1CCCCCC(OCCO)C1
7	Fc1cccc(Oc2cccnc2)c1
8	CCOc1c(CC)cccc1OC
9	CCOCCCO[Si]C1CCCCO1
...	

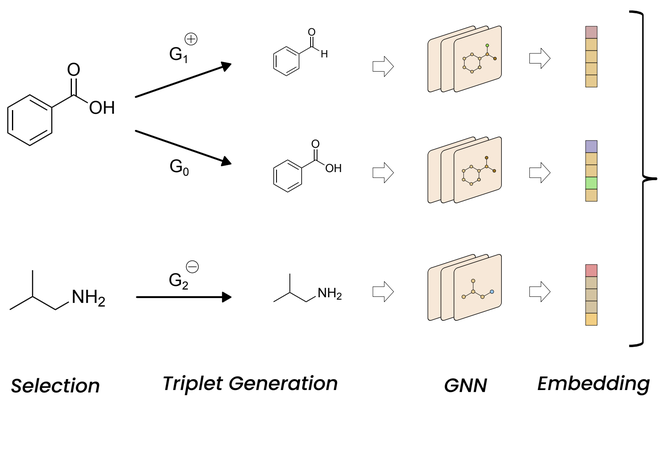
Augmentations

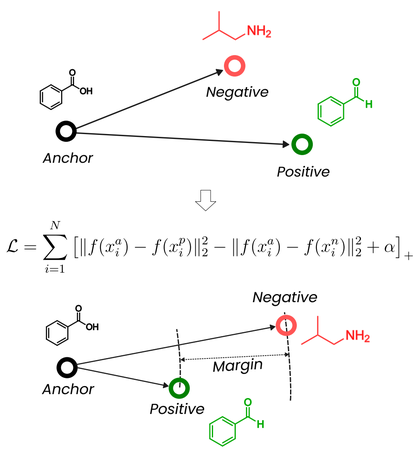
120M Molecules

label	smiles
0	N#Cc1cc(O)c2oncc2c1
0	N#Cc1cc(O)c2onc(Cl)c2c1
0	N#Cc1cc(N)c2oncc2c1
0	N#Cc1cc(O)c2oncc2c1F
1	Cc1nccc(C=CC(C)Cl)n1
1	Cc1nccc(C=CC(Cl)CF)n1
1	Cc1nccc(C=PC(C)Cl)n1
2	O=PCOCC(COCP=O)OCP=O
2	O=PCOCC(OCP=O)C(F)OCP=O
2	O=PCOCC(OCP=O)C(Cl)OCP=O
...	

- Input: **10M Molecules**
- 130 **Augmentation Rules**
- Out: **120M Molecules**

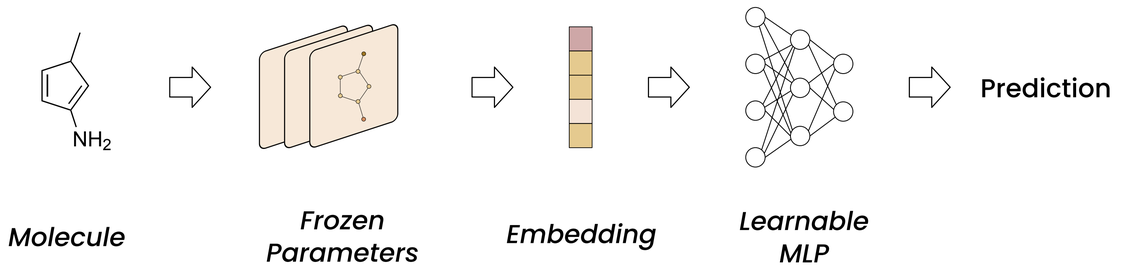
4. Self-Supervised Pre-Training



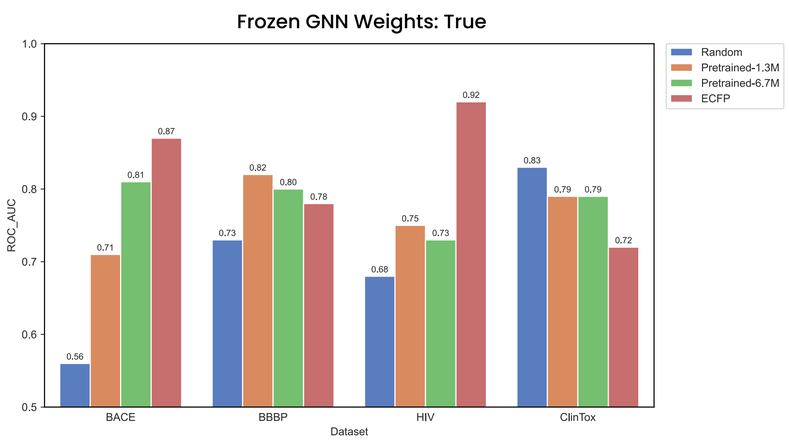


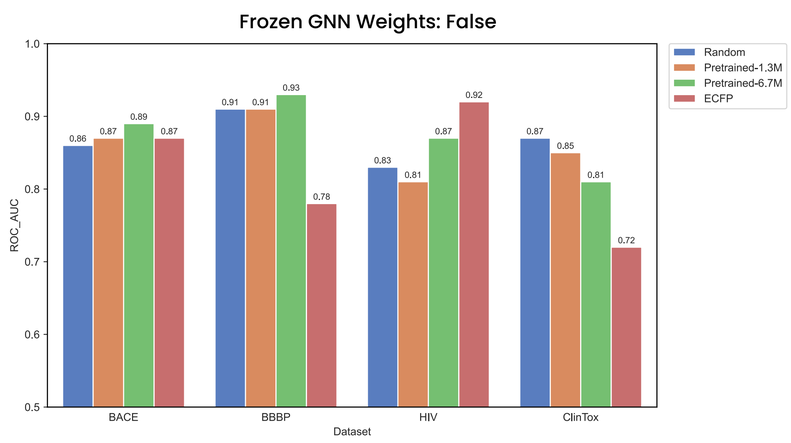
Triplet Loss

5. Fine-Tuning



The Results





6. Conclusion

- Pre-Training **improves embeddings** vs. randomly initialized GNNs
- ECFPs still have **comparatively good embeddings**
- Further investgation of embeddings is required