open
forcefield

@openforcefield

www.openforcefield.org

# GNNs to calculate AM1-BCC charges

November 10, 2022 | Force field release meeting

# Overview

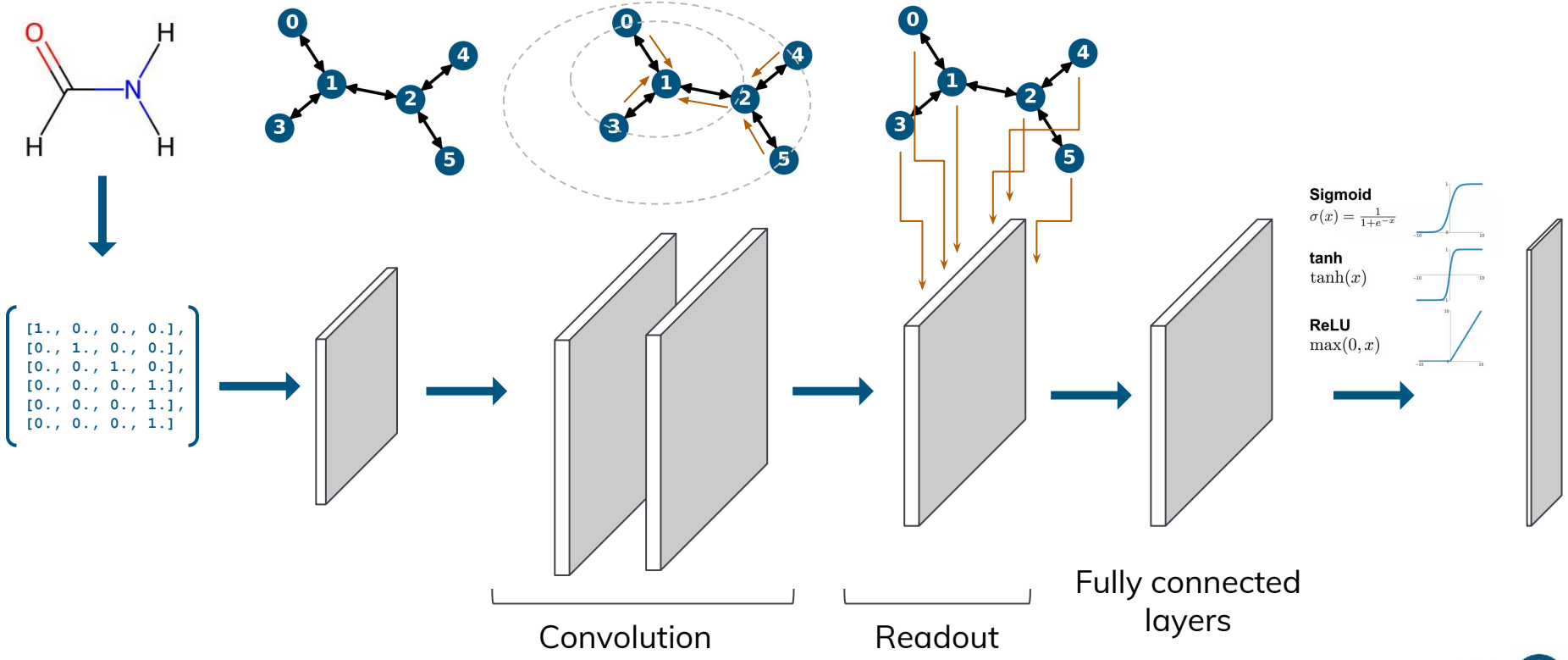
- GNN structure
- Developing a GNN:
  - Hyperparameters / GNN structure
  - Atom features
  - Datasets
  - Targets and metrics
- Current model performances
- Next steps

# Why GNN charges?

- Faster and not conformationally-dependent
- Would be great to get this into Rosemary
- Targeting AM1-BCC for now – could move onto other charge models in future, e.g. RESP

# GNNs

## Structure



# Hyperparameters

## Search space

- Number of convolution layers: 3, 4, 5
- Number of convolution features: 64, 128, **256**
- Number of readout layers: 1, 2
- Number of readout features: 64, 128, 256
- Learning rate: **$10^{-3}$**, $10^{-4}$, $10^{-5}$
- Activation function: ReLU, **Sigmoid**, Tanh

# Atom features



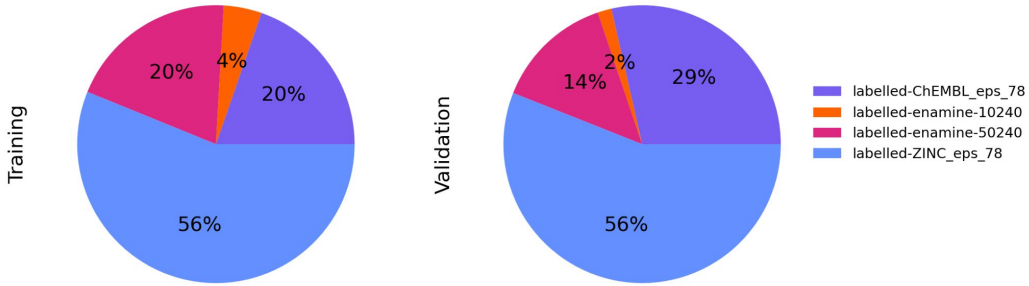
| | v1 | v2 | v3 |
|---|---|---|---|
| AtomicElement | ■ | ■ | ■ |
| AtomConnectivity | ■ | ■ | ■ |
| AtomIsInRing | ■ | | |
| AtomIsInRingOfSize: 3, 4, 5, 6 | | ■ | ■ |
| AtomAverageFormalCharge (resonance) | | ■ | ■ |
| AtomHybridization | | | ■ |

## Datasets

- Curated ZINC and ChEMBL datasets from Bleiziffer *et al.* 2018
  - ChEMBL: 107254 molecules
  - ZINC: 256184 molecules
- Enamine Discovery Diversity sets (10240, 50240 molecules)
  - Filtered with same criteria as Bleiziffer et al. 2018
  - 6537 records
- OpenFF Industry benchmark set
  - 13715 molecules

## Dataset split

- Pooled datasets: enamine, ZINC, ChEMBL
- Selected molecules to cover all atom environments with no more than 10 molecules each (168911 total)
- Split by diversity (Dice similarity between Morgan fingerprints)
- Training: 80%
- Validation: 20%
- Test:
  - SPICE
  - Peptide chains (1-5)
  - OpenFF benchmark set



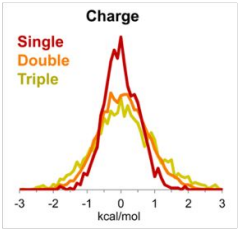
## Ways to achieve AM1-BCC charges

- Training directly to AM1-BCC charges
- Training to AM1 charges and applying (retrained) BCCs

## Metrics

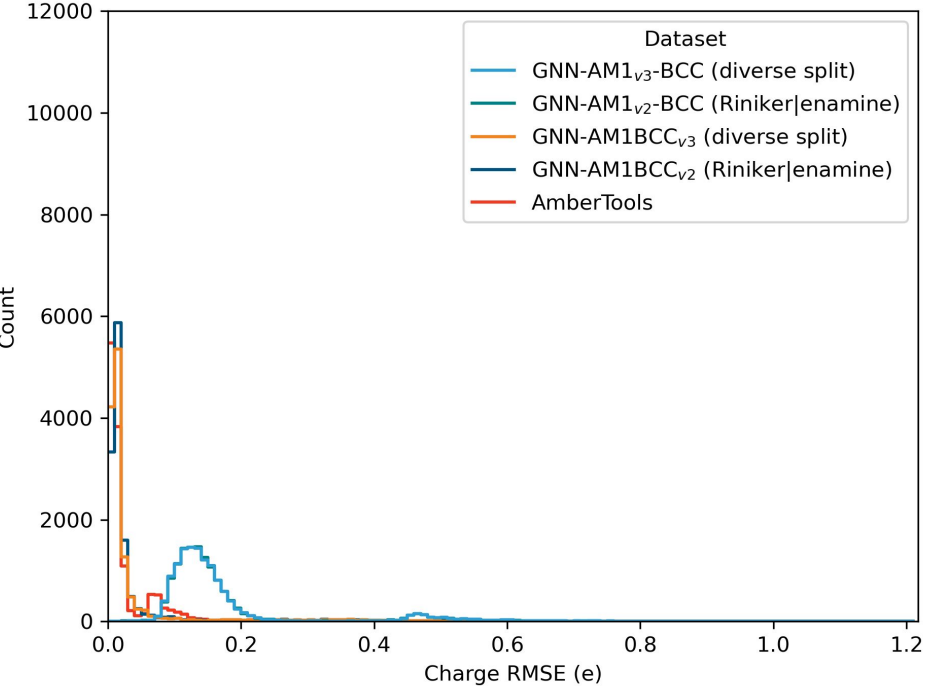
- Charge RMSE
- Properties not used, e.g.:
  - Dipoles
  - ESPs

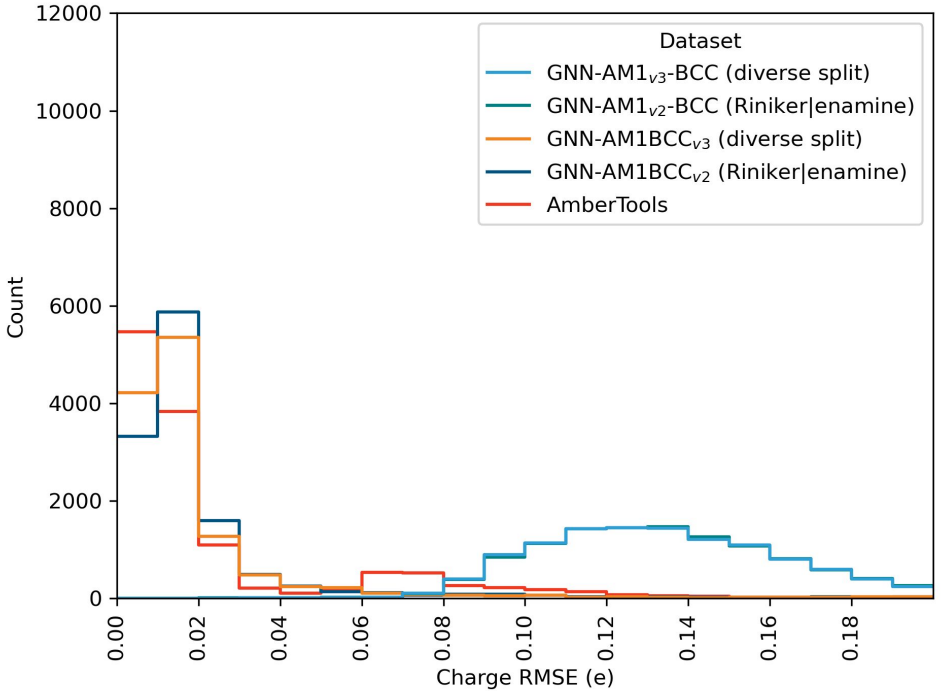
Rocklin, Mobley and Dill 2013



# Current models

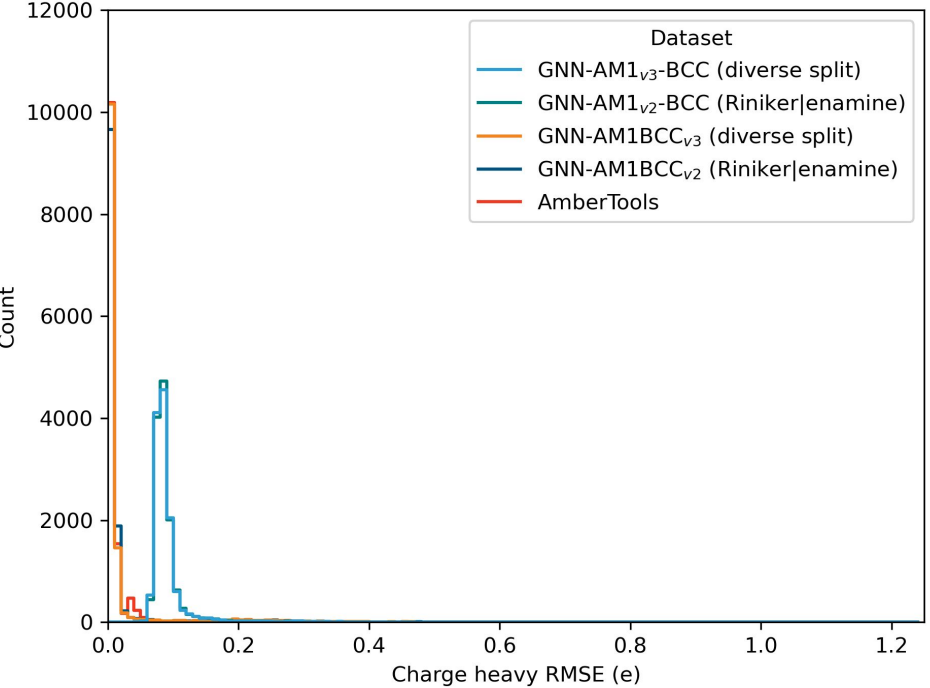
Spice dataset

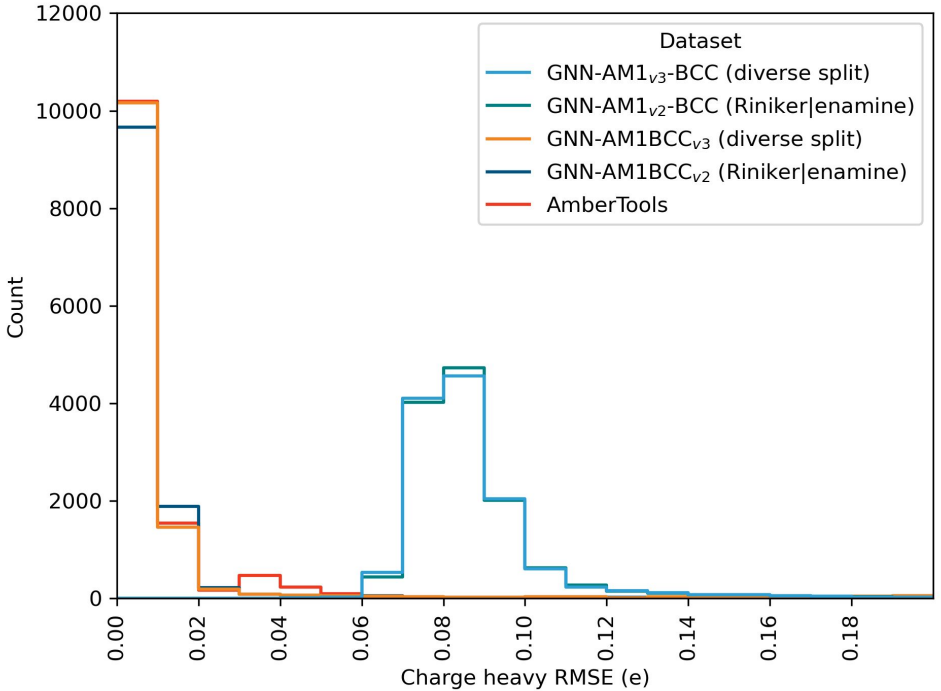


# Current models

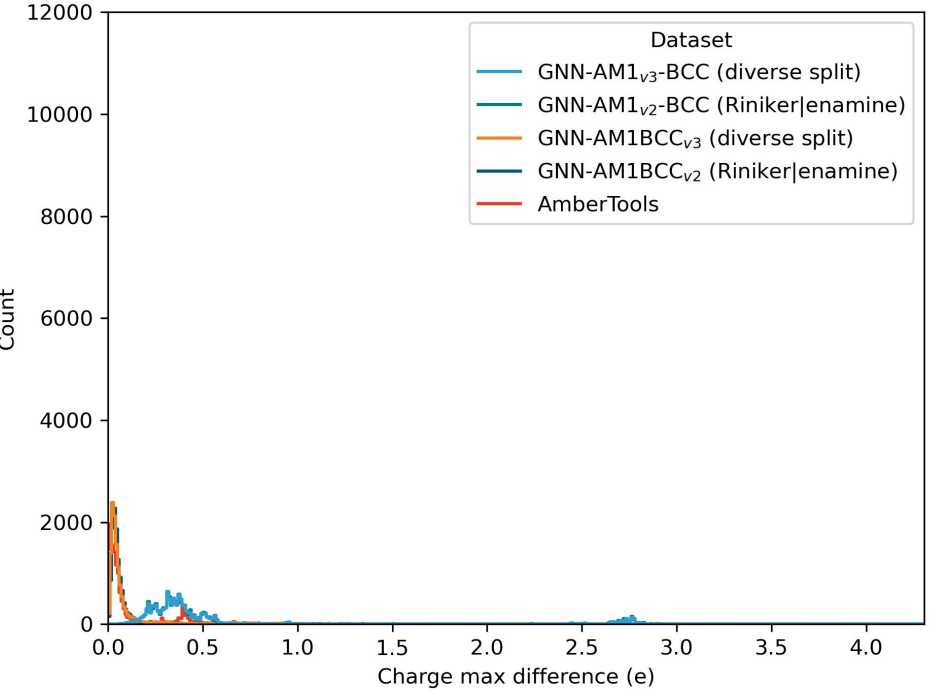
Spice dataset

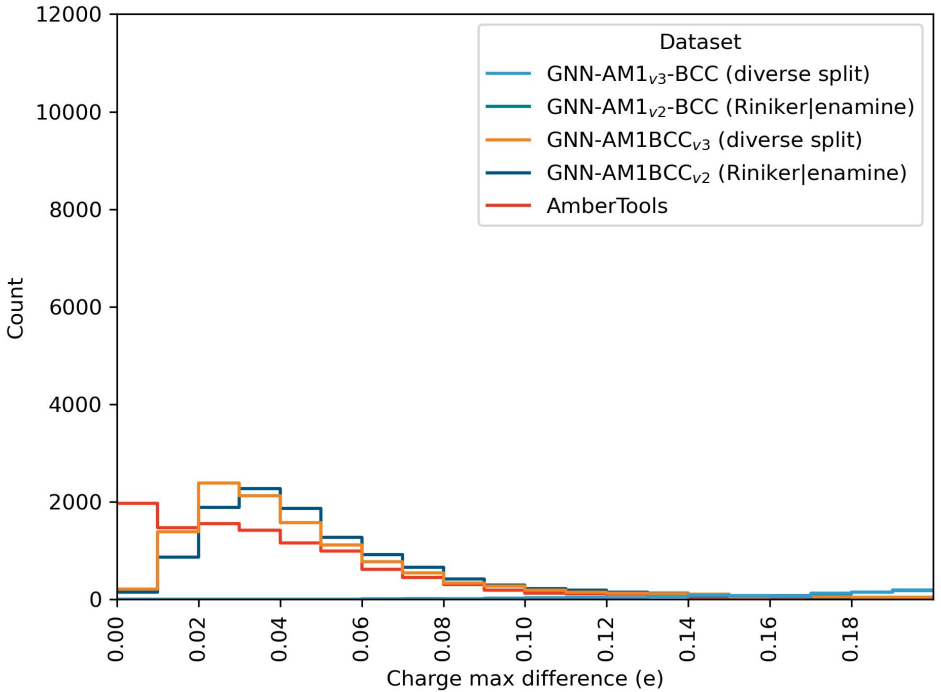


# Current models

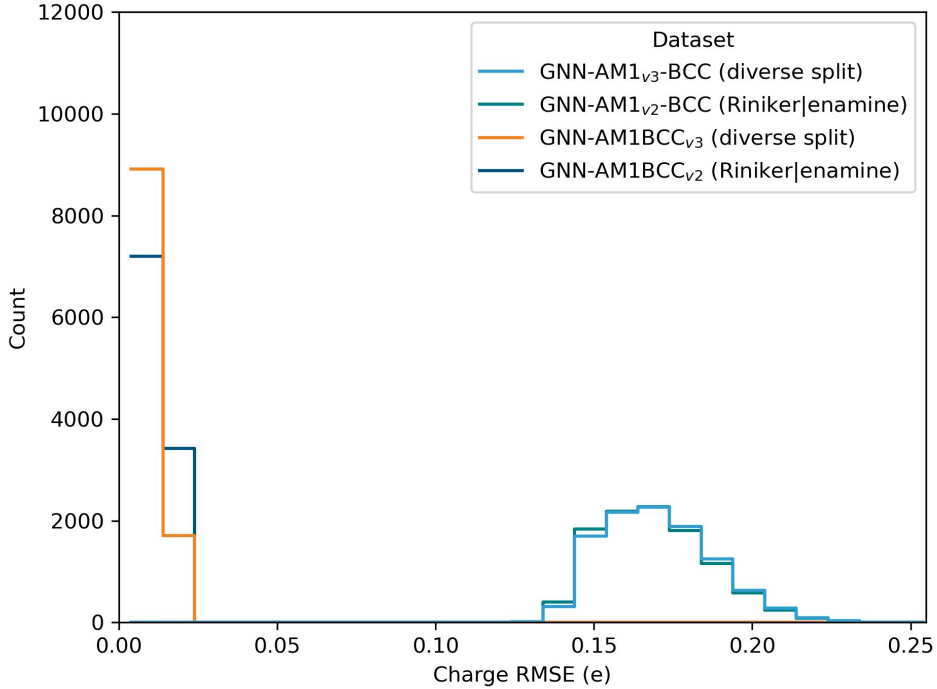
Spice dataset

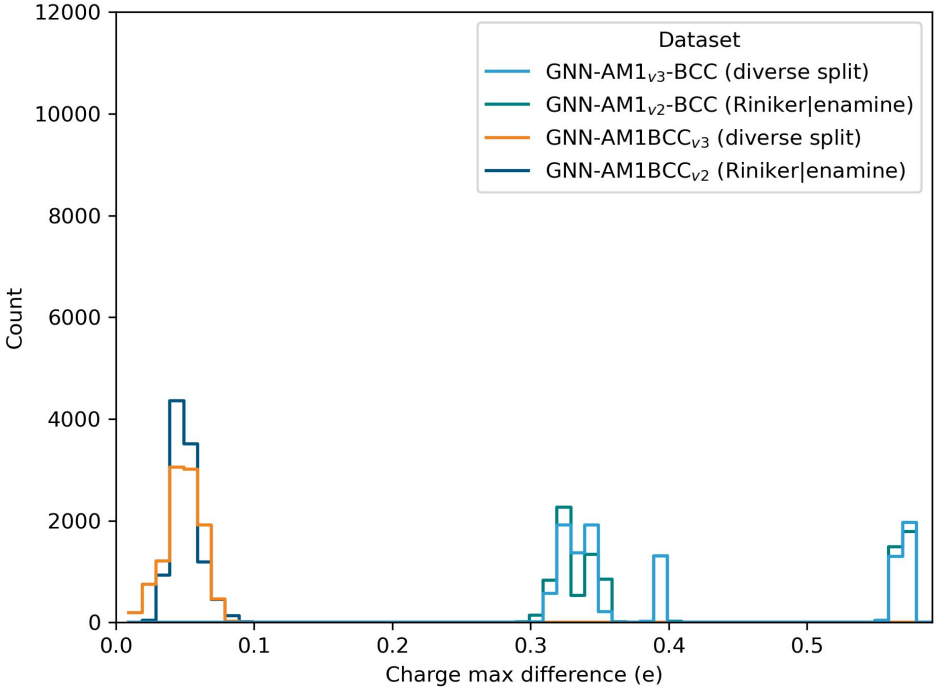


# Current models

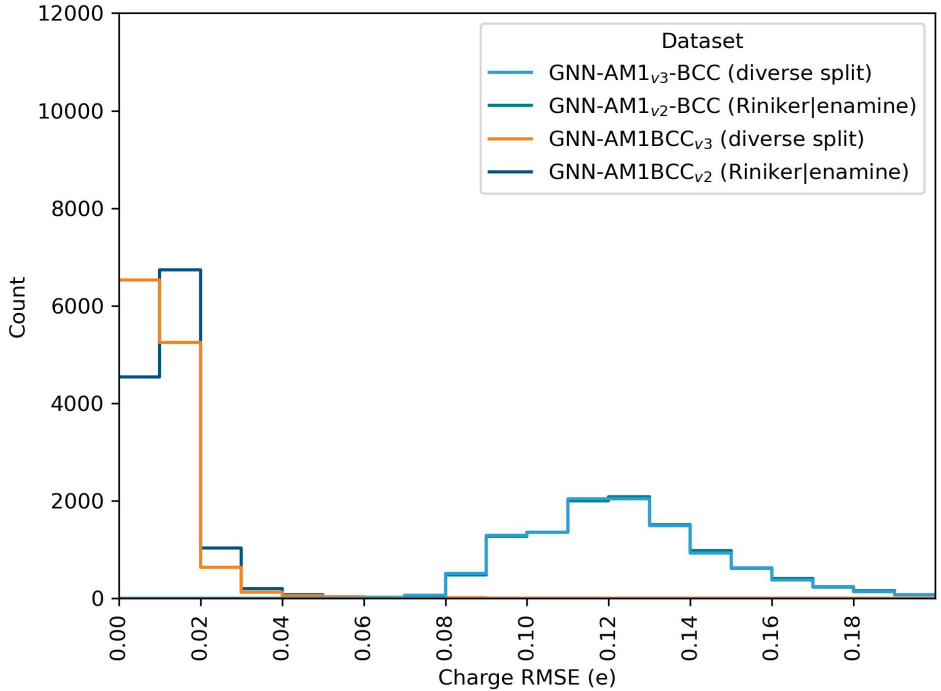
Peptides (length 1-5)

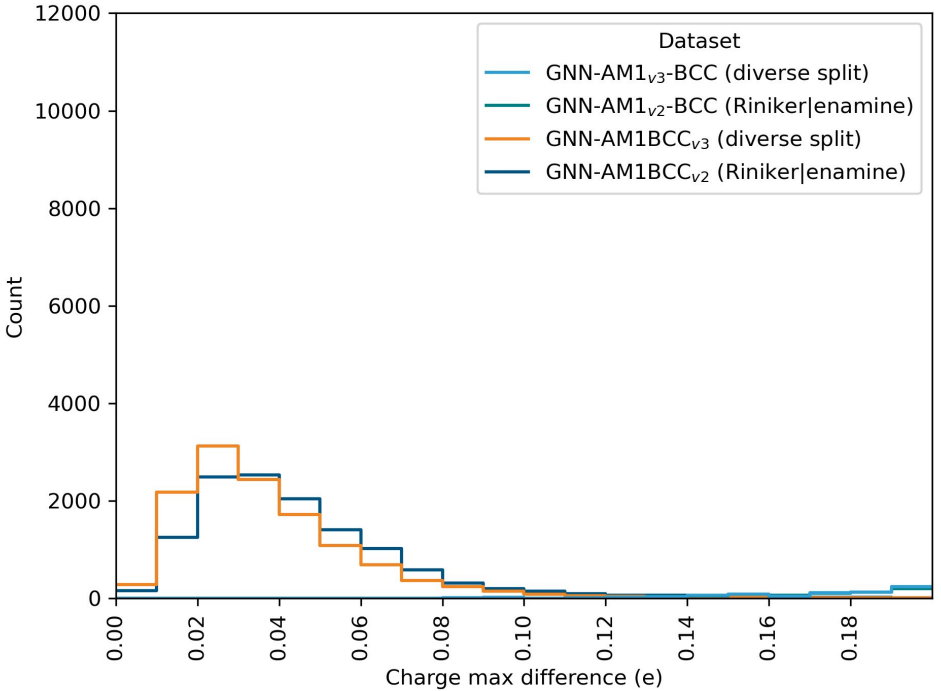


# Current models

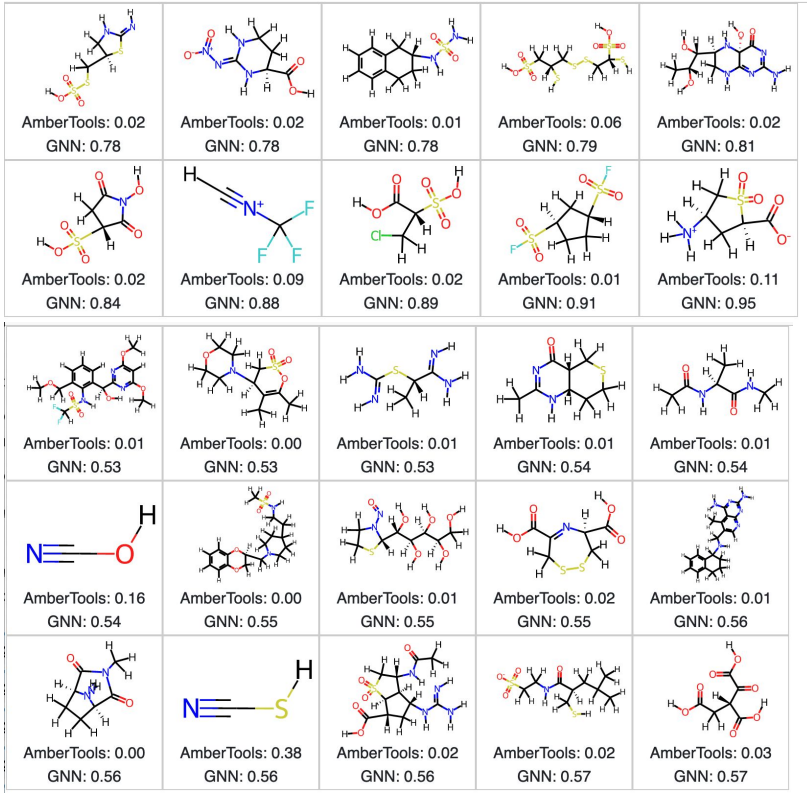
## OpenFF Benchmark set



# Current models

## Examples of poorly performing molecules



# Future directions

## Next steps

- Fix bug in training datasets with some mixed-up records – re-train
- Look closer into outliers
  - Possibly add to training set, especially with more S or H environments
- Finish computing AM1-BCC charges for larger molecules with AmberTools
- Improving efficiency / adding documentation / user-friendliness to OpenFF NAGL

## Other directions

- Improving BCC training
- Investigate architectures with edge features