Integrating Machine Learning and Quantum Chemistry for Micro-pK_a Predictions

OMRI ABARBANEL

DR. GEOFFREY HUTCHISON

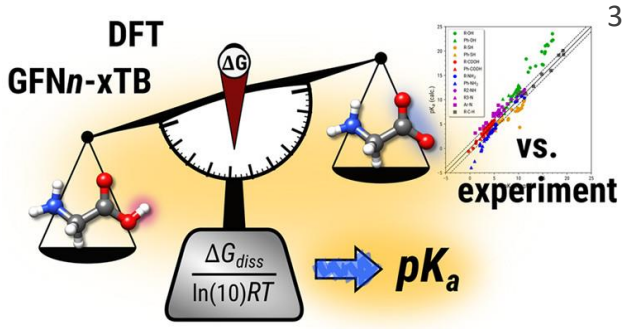
UNIVERSITY OF PITTSBURGH, DEPARTMENT OF CHEMISTRY

ACS FALL MEETING 2023

AUGUST 14, 2023

Current SOTA

Quantum Mechanical Models

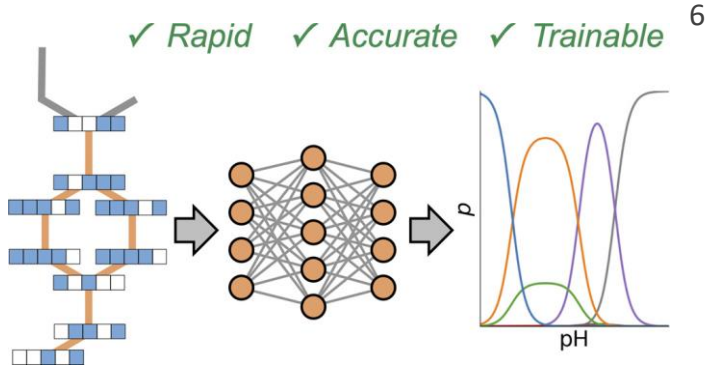


[3]

Model	RMSE Range
Jensen 2017[1]	0.90-1.00
Grimme 2018[2]	0.68-1.01
Grimme 2021[3]	0.84-2.68

[1] 10.1021/acs.jpca.6b10990
[2] 10.1007/s10822-018-0145-7
[3] 10.1021/acs.jpca.1c03463

Machine Learning Models

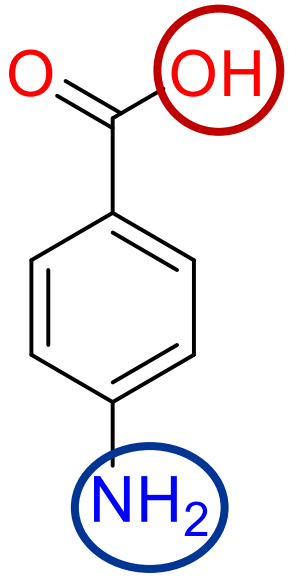


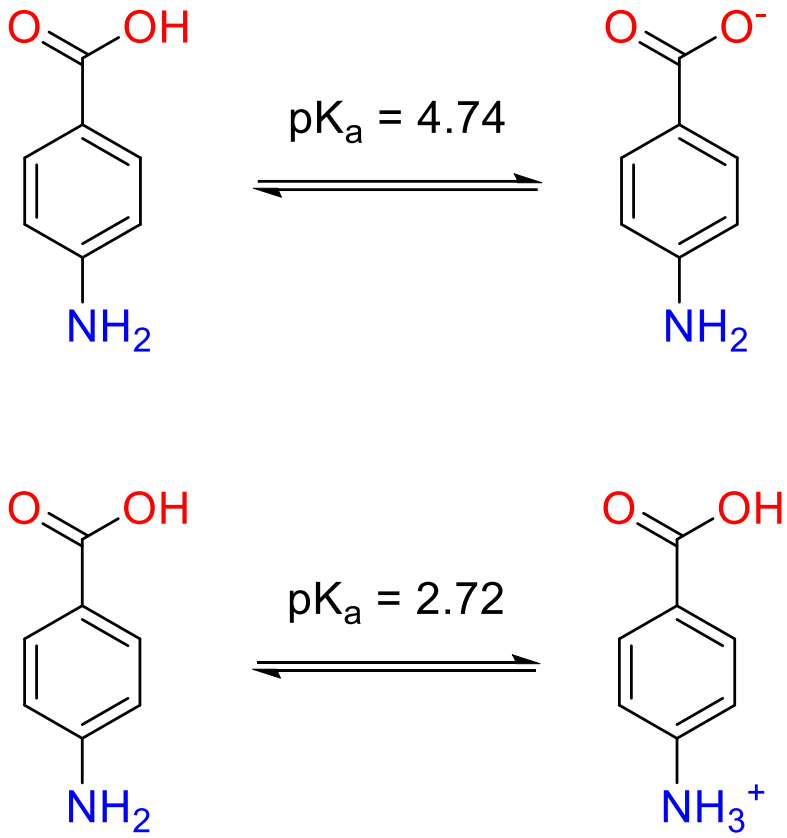
[6]

Model	RMSE Range
Czodrowski 2020[4]	0.79-1.51
Langer 2022[5]	0.97-1.13
Epik 7 2023[6]	0.54-1.01

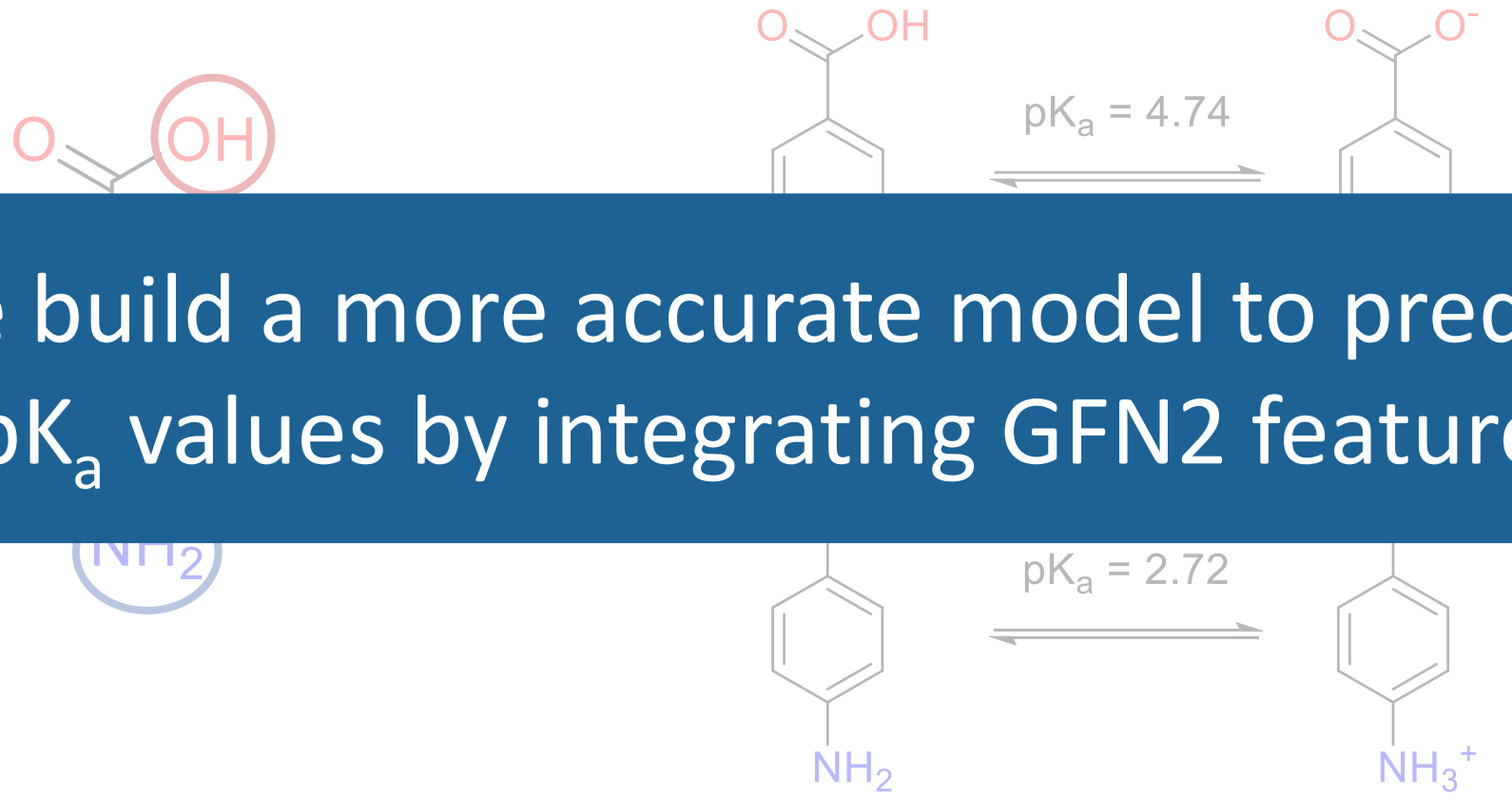
[4] 10.12688/f1000research.22090.2
[5] 10.3389/fchem.2022.866585
[6] 10.1021/acs.jctc.3c00044

Micro-pK_a





Micro-pK_a



Can we build a more accurate model to predict micro-pK_a values by integrating GFN2 features?

Workflow

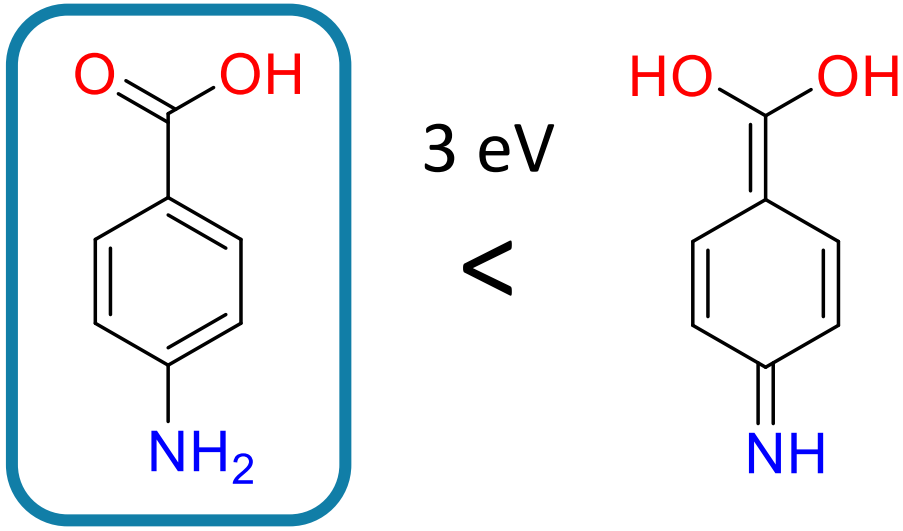

O
OH
NH2
Tautomer Search
Reaction Site Enumeration
Micro-pKa Prediction
4.72
O
OH
NH2
2.72

Workflow


O
OH
NH2
Tautomer Search
Reaction Site Enumeration
Micro-pKa Prediction
4.72
O
OH
NH2
2.72

Tautomer Search

- Tautomer enumeration using RDKit
- GFN2 optimization
  - ALPB implicit solvation model in water
- Lowest energy tautomer is chosen

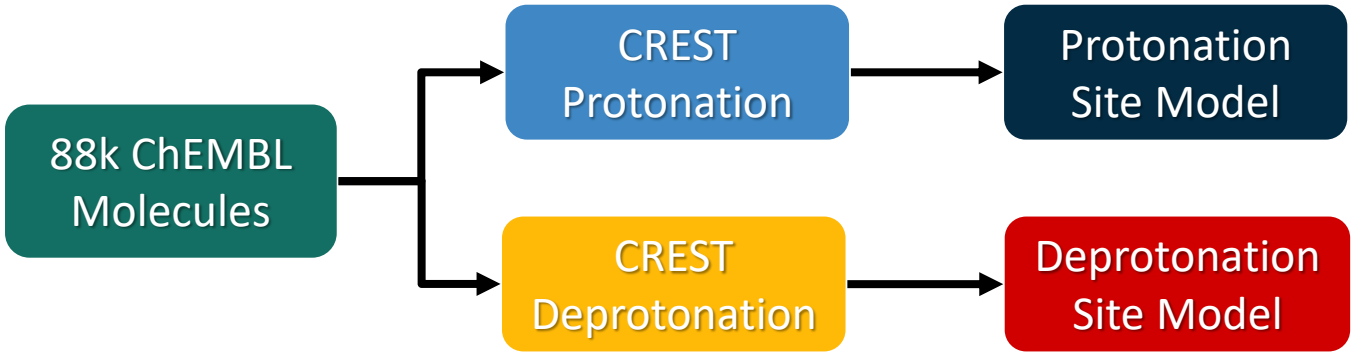


Workflow


O
OH
NH2
Tautomer Search
Reaction Site Enumeration
Micro-pKa Prediction
4.72
O
OH
NH2
2.72

Reaction Site Enumeration

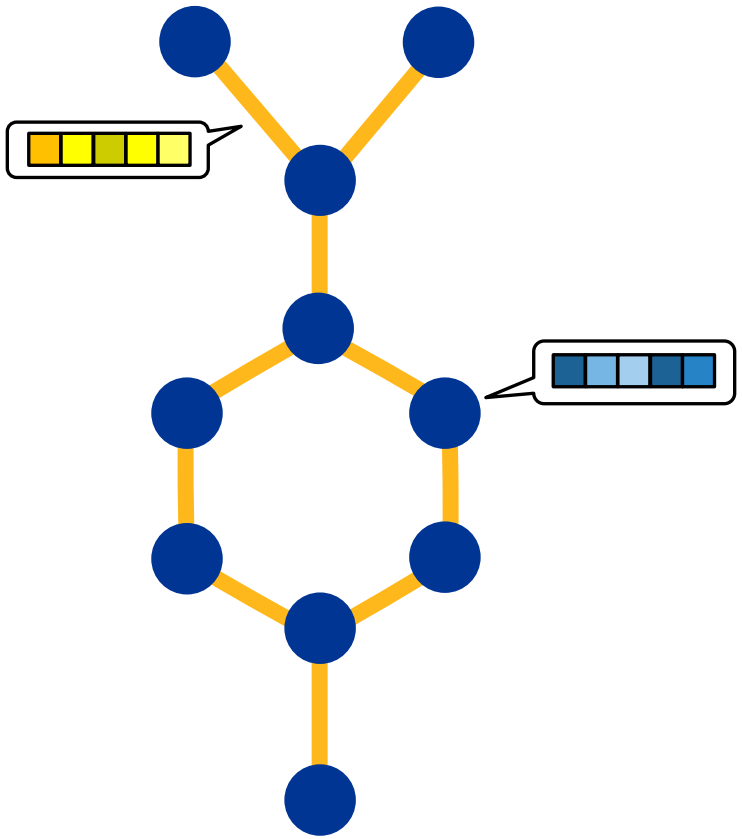
- CREST[1] protonation \ deprotonation site search[2]
- Two surrogate GNN models
  - Protonation site prediction
  - Deprotonation site prediction



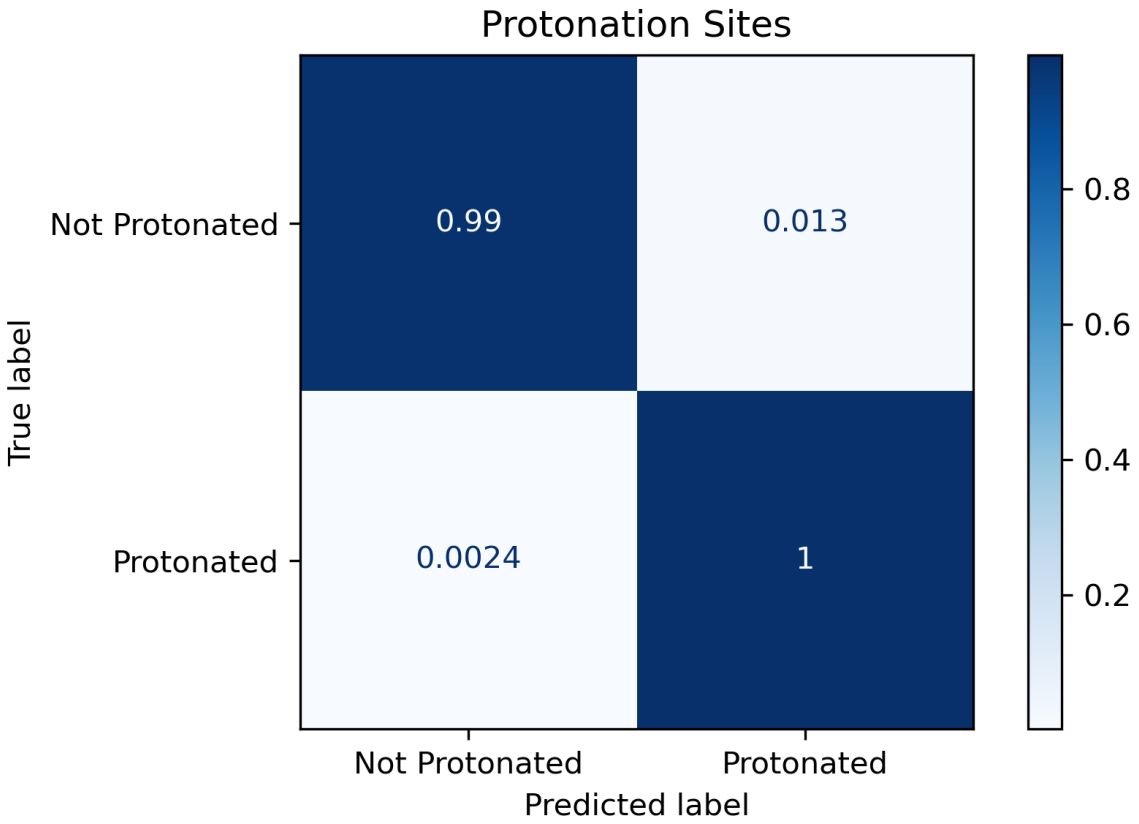
[1] 10.1039/C9CP06869D
[2] 10.1002/jcc.24922

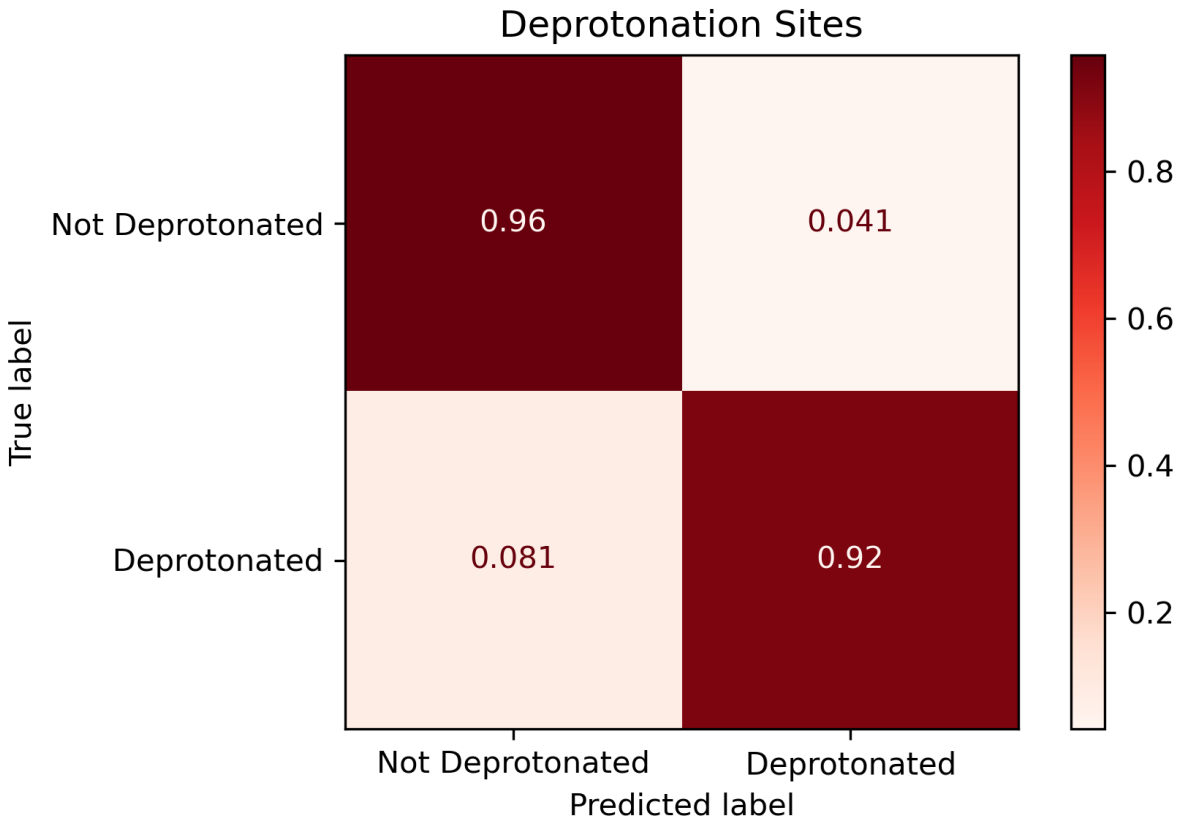
Graph Representation

	RDKit Features	GFN2 Features
Atom	• Atom Type • No. Heavy Neighbors • Formal Charge • Hybridization • Is in Ring • Is Aromatic • Atomic Mass • VDW Radius • Cov. Radius • Chirality • No. Hydrogens • Is HBA\HBD	• Partial Charge • Coord. Number • Polarizability • Fukui Indices
Bond	• Bond Type • Is Conjugated • Is in Ring • Stereochemistry	• Bond Order
Molecule	• Radius of Gyration • Spherocity • Aspherocity • Eccentricity • sp^3 Fraction	• Charge • $\Delta E_{ionization}$



Site Prediction Models



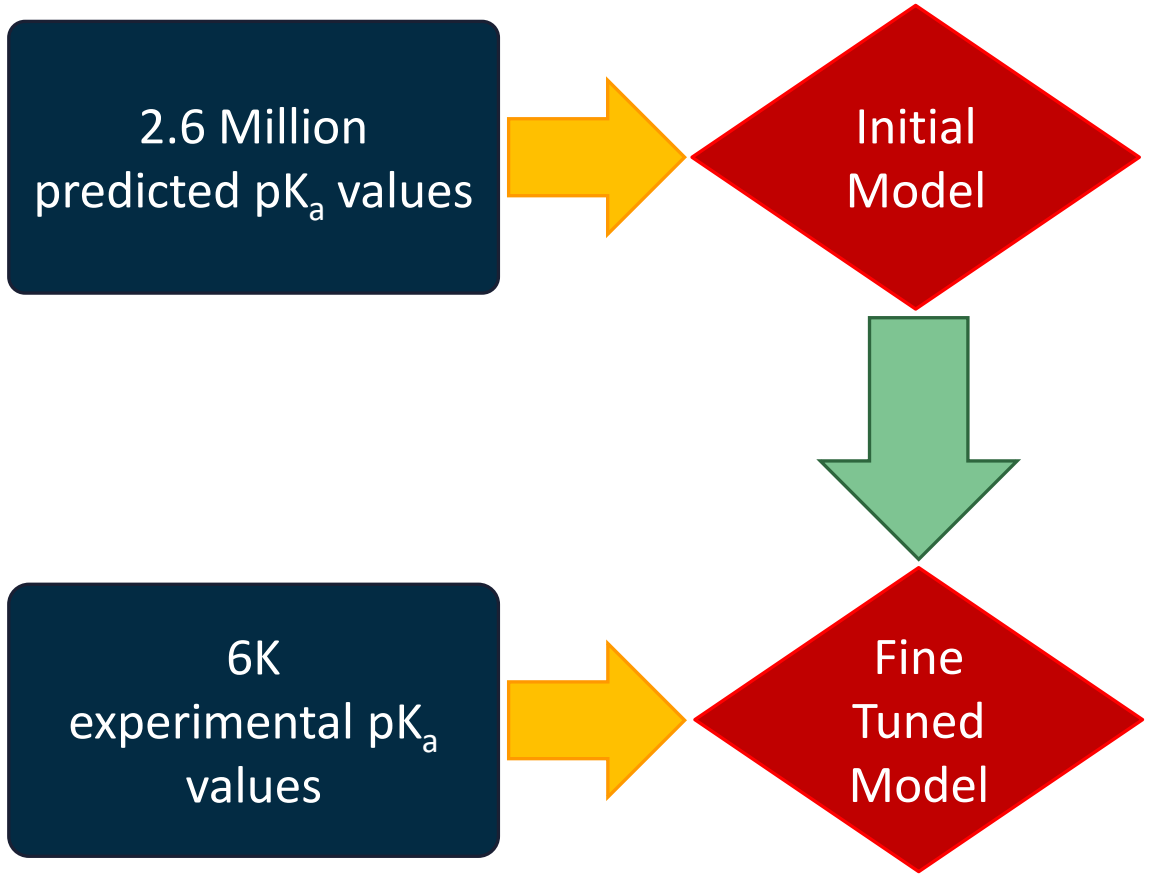


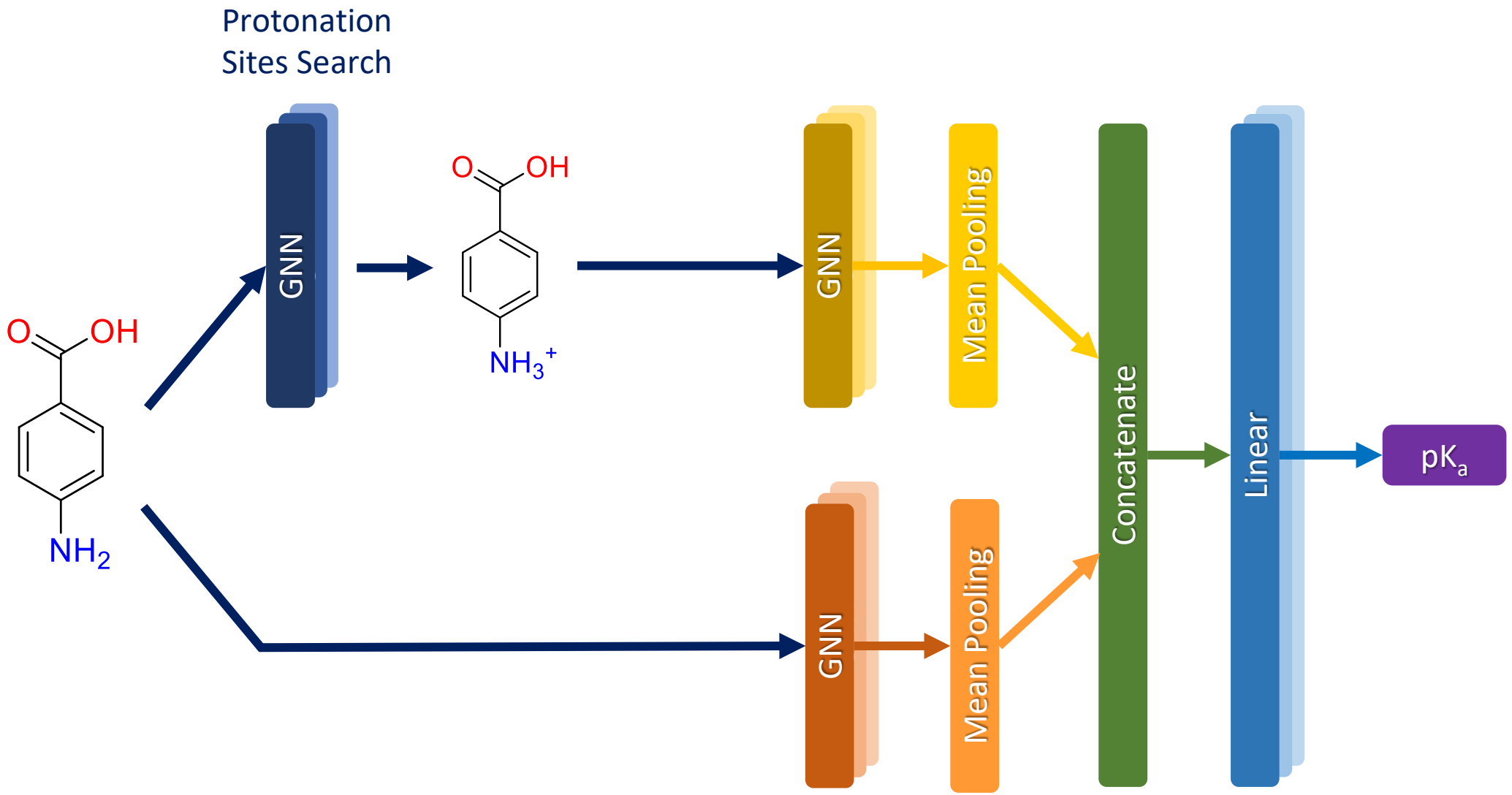
Workflow

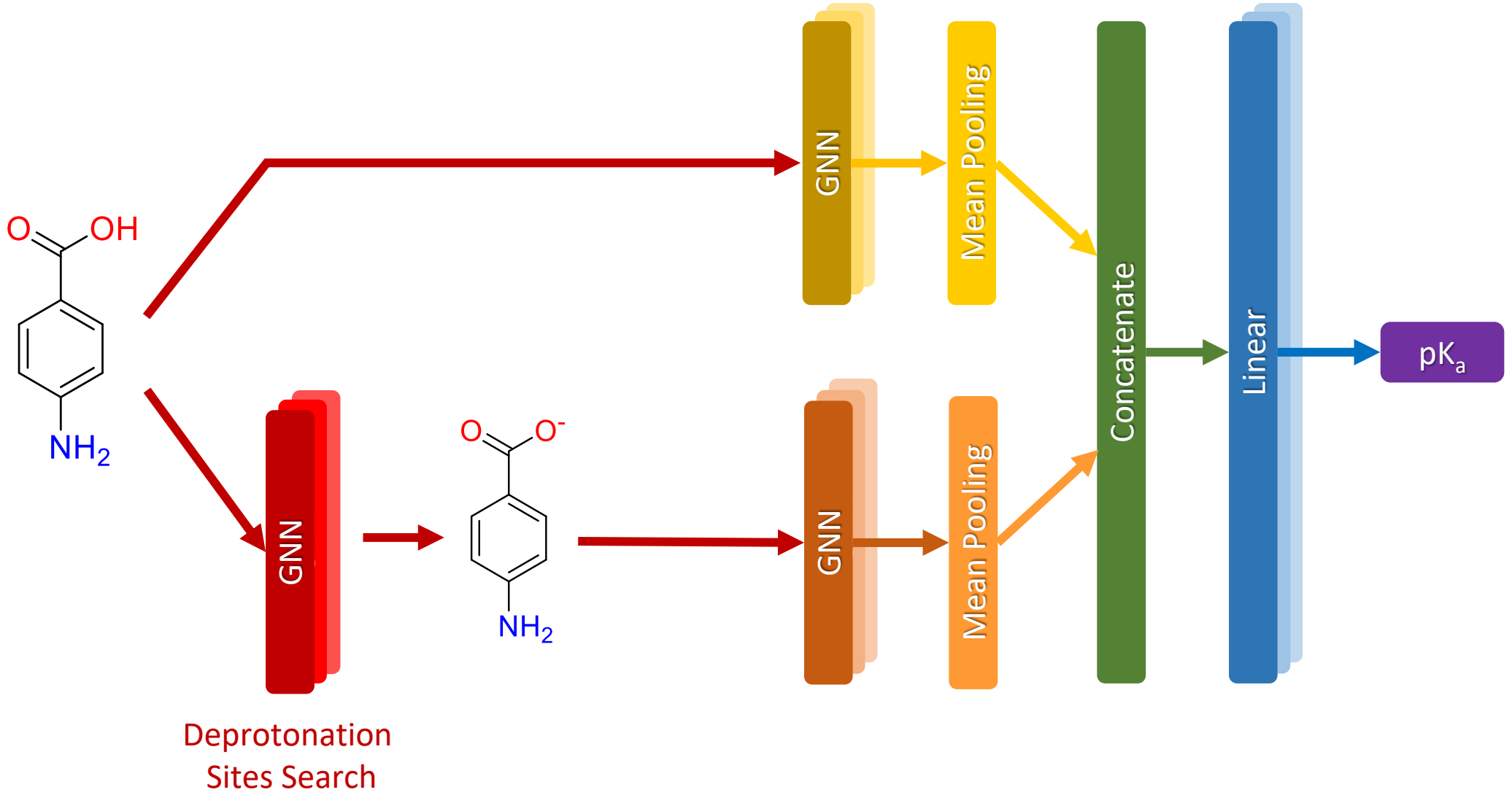

O
OH
NH2
Tautomer Search
Reaction Site Enumeration
Micro-pKa Prediction
4.72
O
OH
NH2
2.72

Micro-pK_a Prediction Model

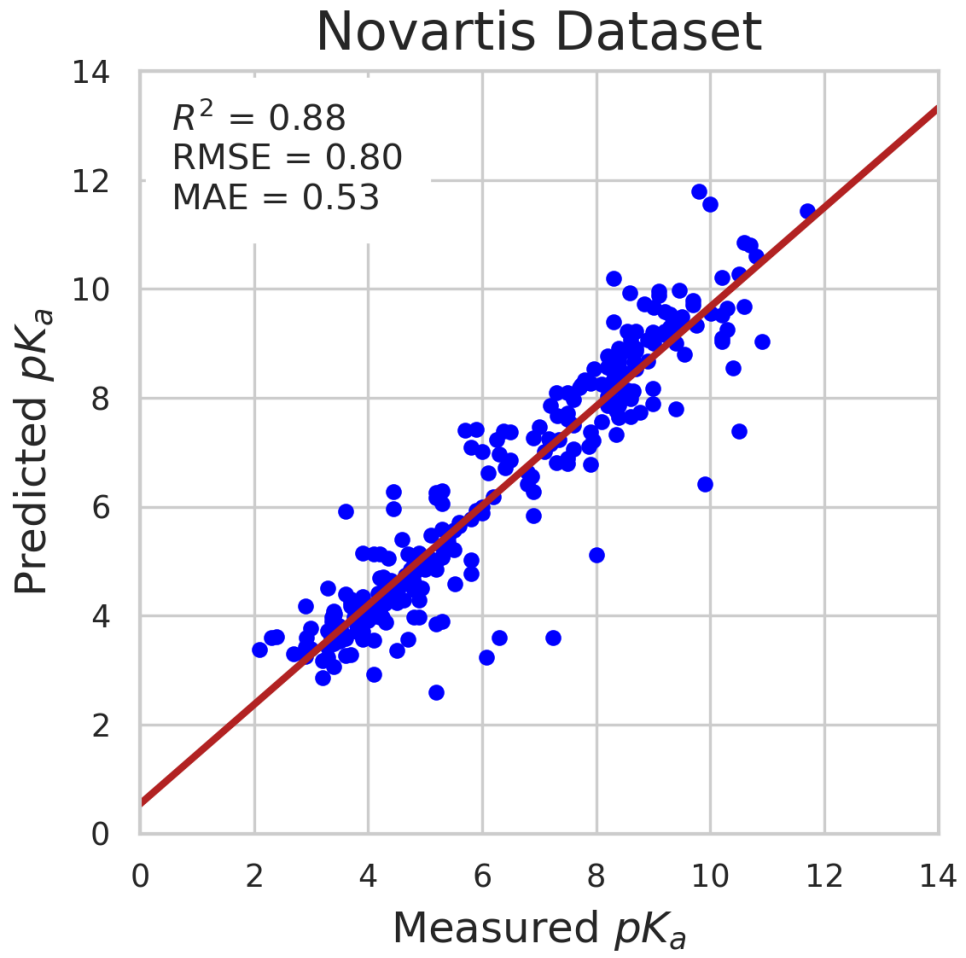
- Problem – Not enough experimental data
  - ~6,000 molecules from the Czodrowski lab
- Solution – Transfer Learning using GNN
  - ~1.5 Million molecules from ChEMBL
  - ~2.6 Million ChemAxon predicted pK_a values

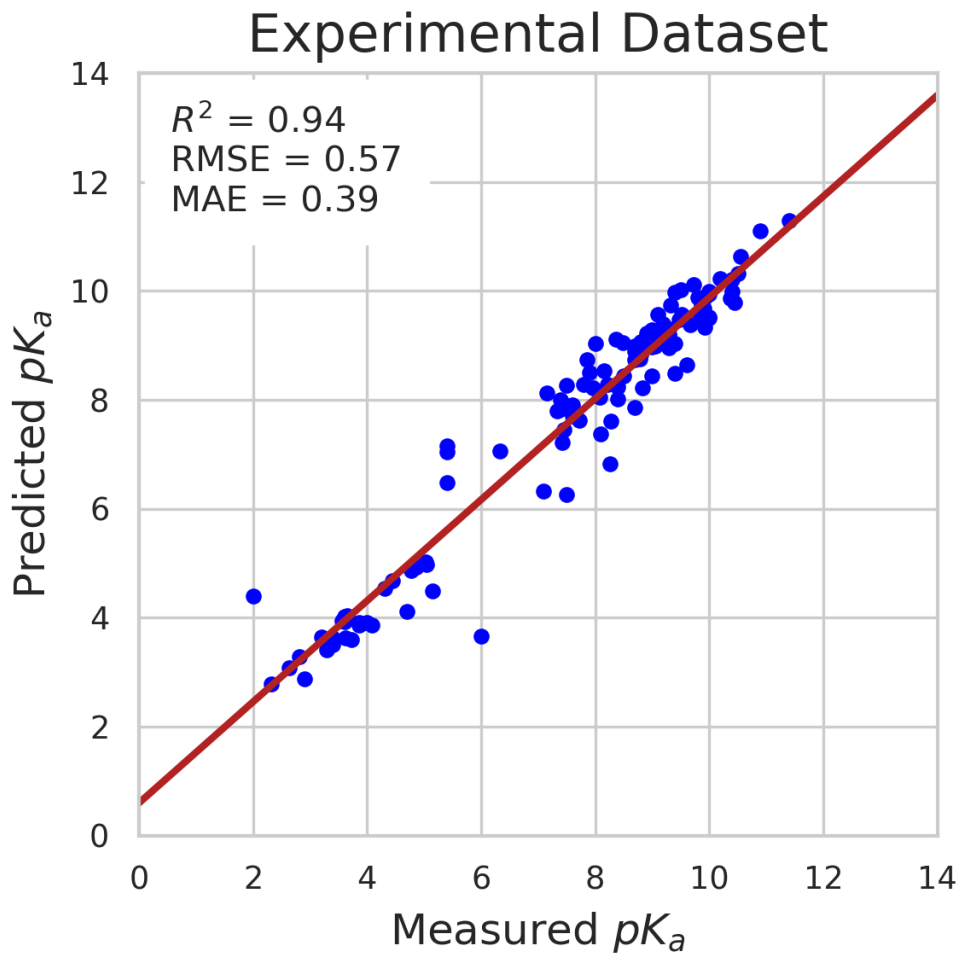



Protonation
Sites Search
O
OH
NH2
GNN
O
OH
NH3+
GNN
Mean Pooling
GNN
Mean Pooling
Concatenate
Linear
pKa

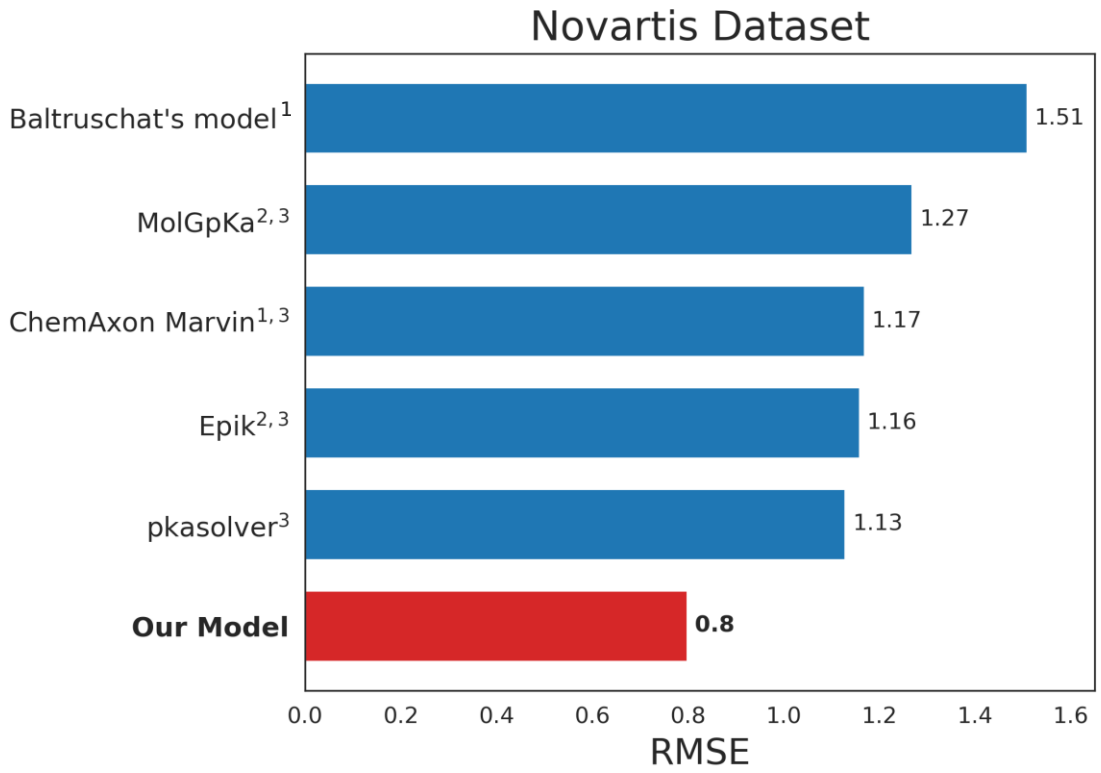

O
OH
NH2
GNN
Mean Pooling
GNN
O
O-
NH2
GNN
Mean Pooling
Concatenate
Linear
pKa
Deprotonation
Sites Search

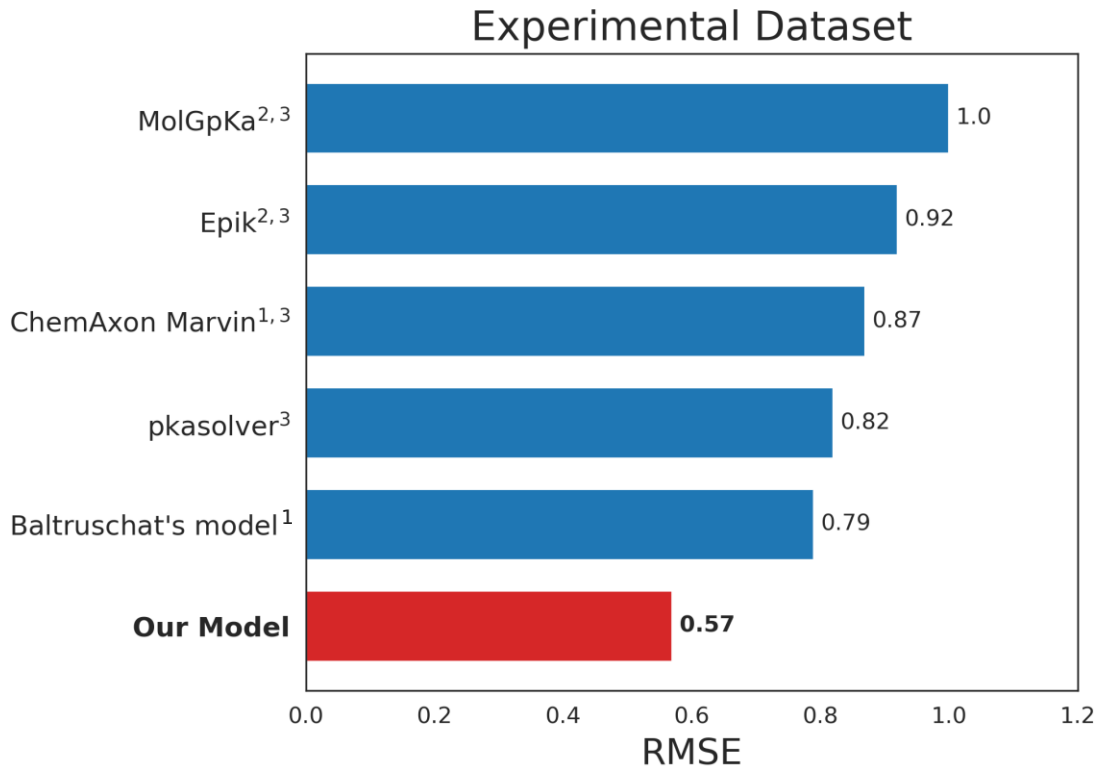
Test Datasets





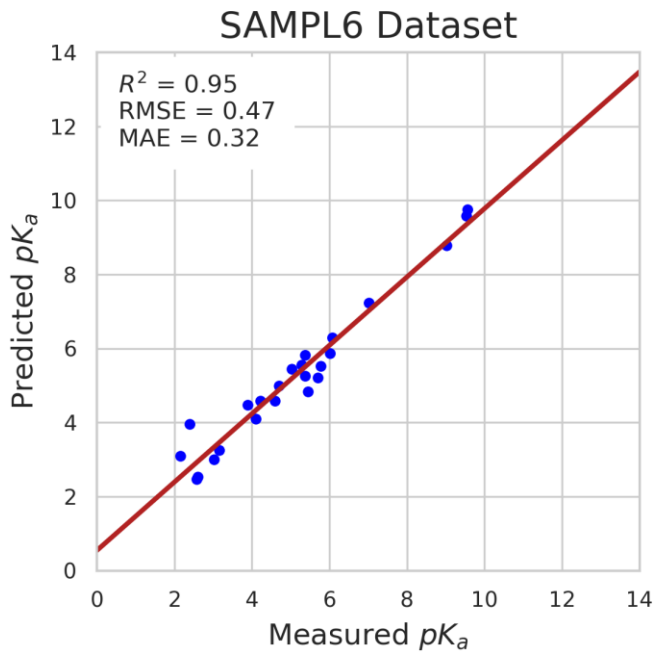
Test Datasets

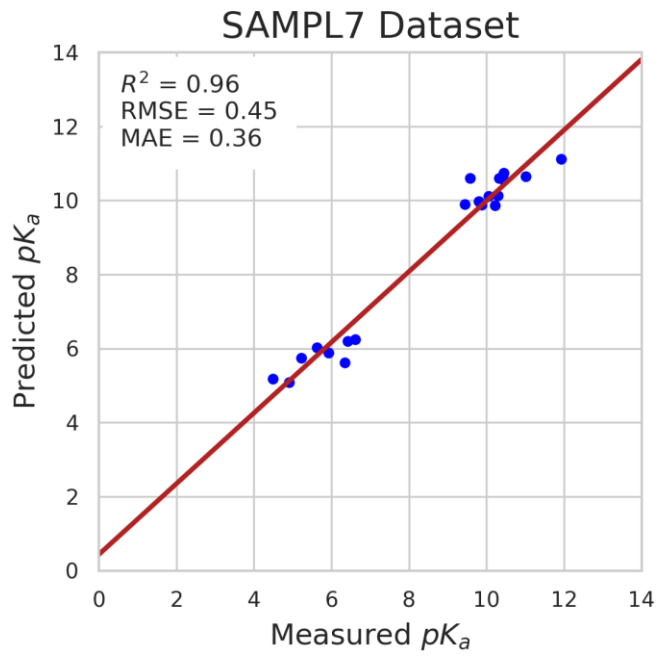


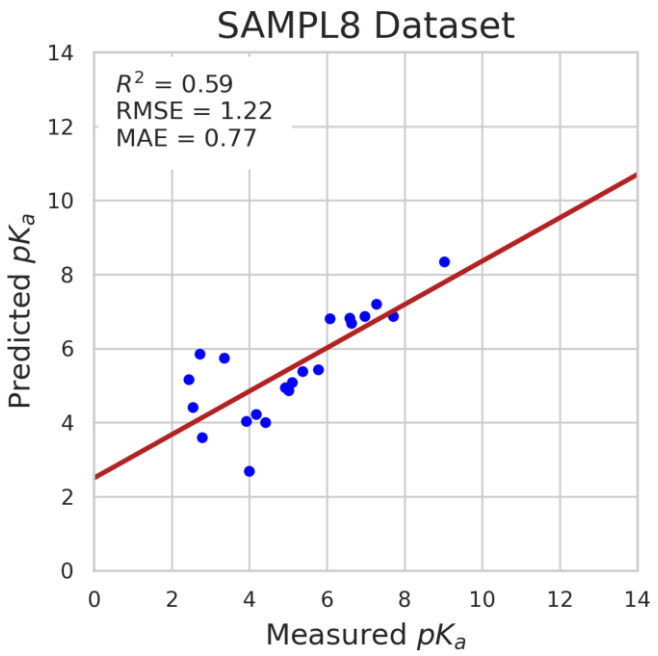


[1] 10.12688/f1000research.22090.2
[2] 10.1021/acs.jcim.1c00075
[3] 10.3389/fchem.2022.866585
[4] 10.1016/j.apsb.2022.11.010

SAMPL Challenges





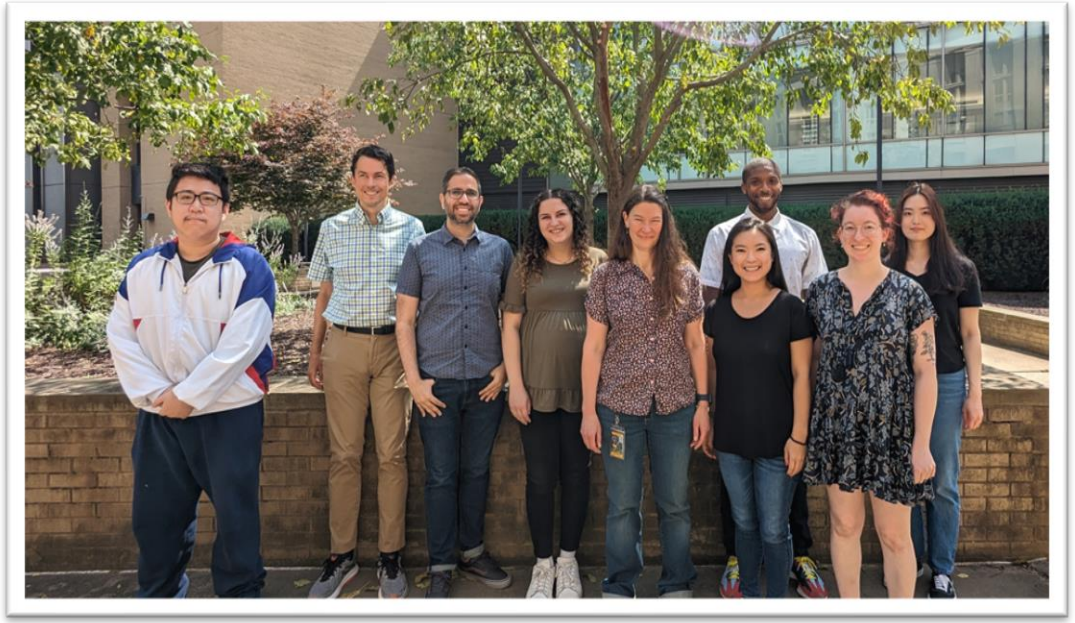


	SAMPL6[1]	**SAMPL7**[2]	**SAMPL8**[3]
	RMSE	RMSE	RMSE
Best Submission	0.68	0.71	1.45
Our Model	**0.47**	**0.45**	**1.22**

[1] 10.5281/zenodo.2651393, [2] 10.5281/zenodo.5637494, [3] 10.5281/zenodo.7535037

Conclusions

- First free, open model with micro-pK_a predictions with RMSE below 0.8
- Adding semi-empirical QM features can help with micro-pK_a predictions
  - Comparisons with ML models (e.g., ANI / AIMNET) to come
- More experimental data is needed (please!)
  - Transfer learning is a viable way to remedy that
- ChemRxiv manuscript, code and data to come soon



Thank You

QUESTIONS?

Office of Naval Research
USN ONR
Science & Technology

PittCRC
Center for Research Computing

University of Pittsburgh