technische universität
dortmund

fakultät für chemie
und chemische biologie

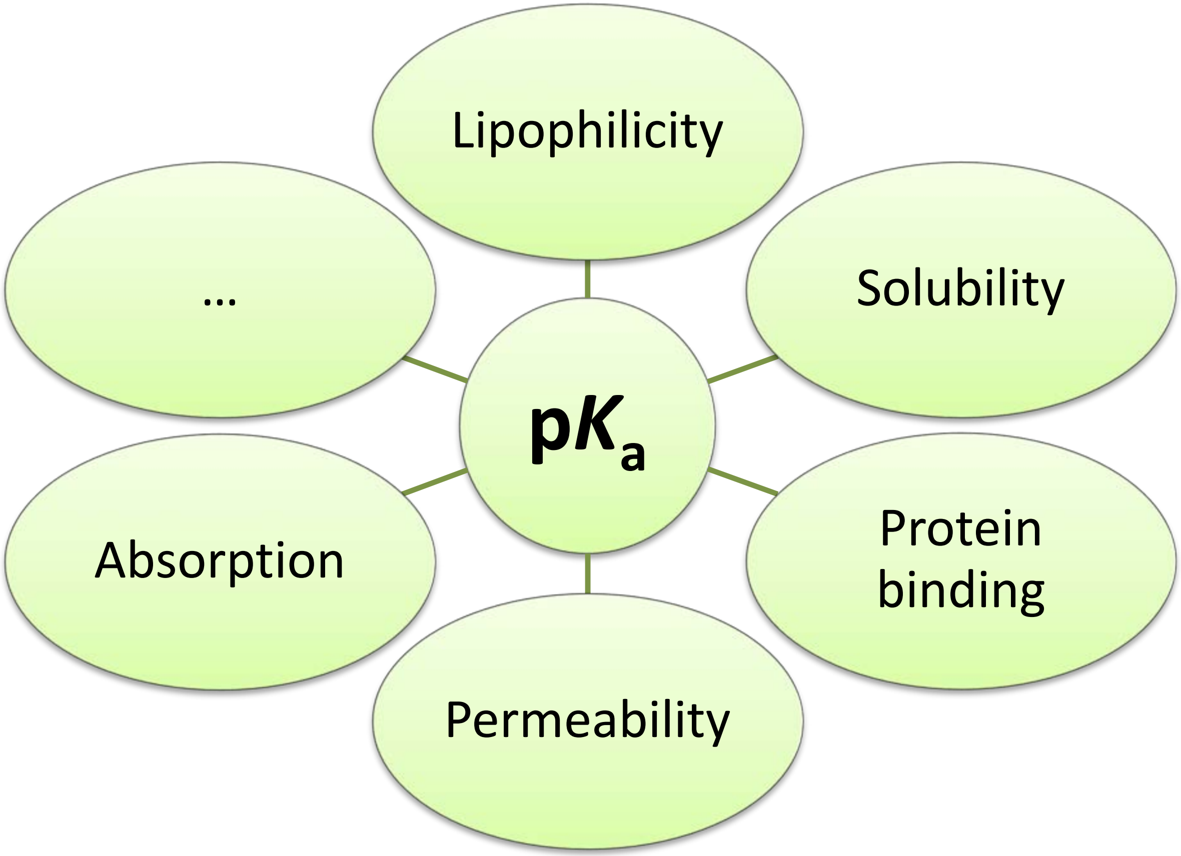
pK_a predictions on top of the RDKit

Marcel Baltruschat @CzodrowskiLab

https://github.com/czodrowskilab/Machine-learning-meets-pKa

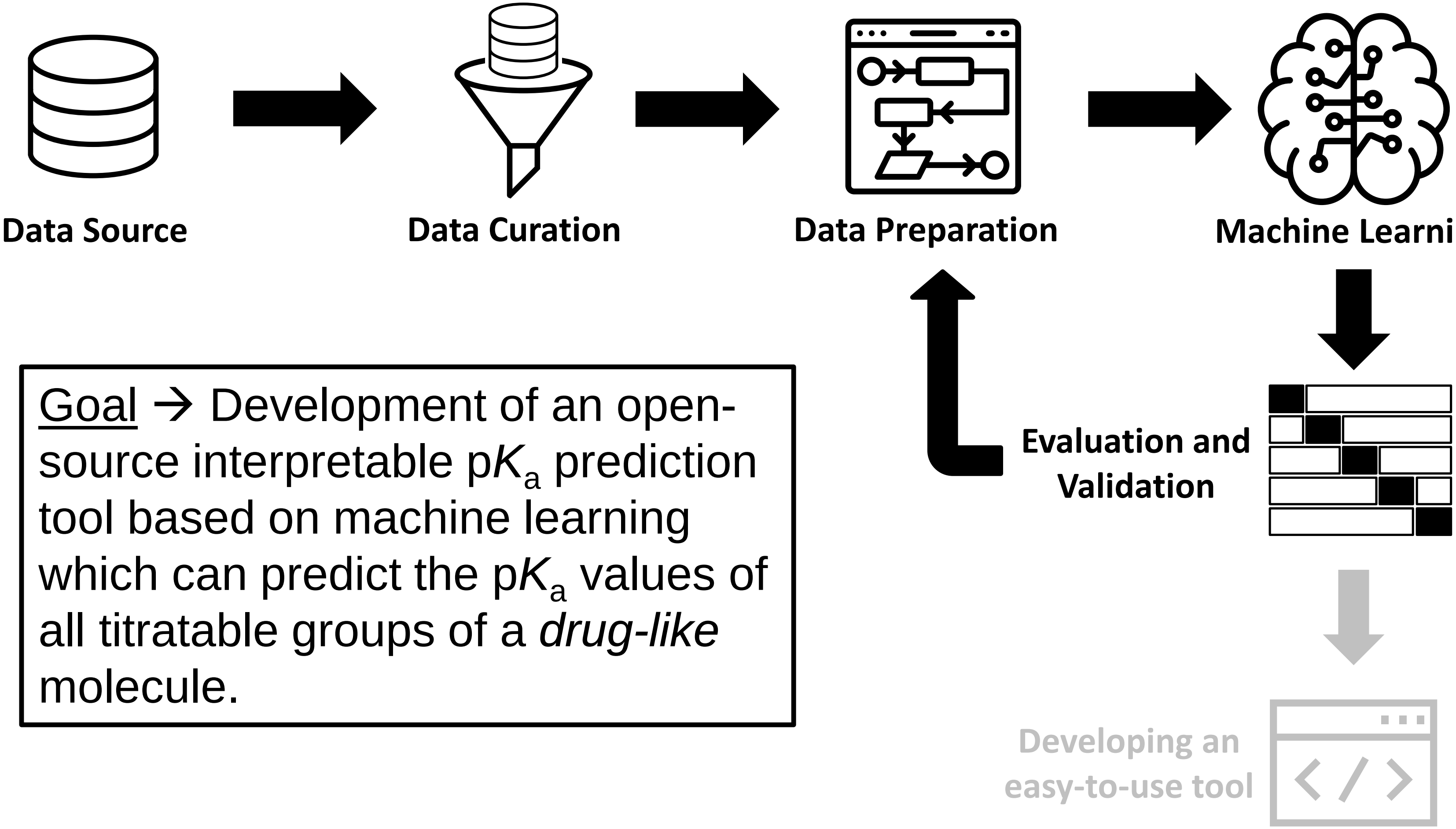
https://github.com/czodrowskilab/Multiprotic-pKa-Processing

Why pK_a?



- Latest published free pK_a predictors can‘t reach the quality of commercial tools
- Free pK_a predictors lack of features e.g. locating titratable groups

Research Topic



Data Source

Data Curation

Data Preparation

Machine Learning

Evaluation and Validation

Developing an easy-to-use tool

<u>Goal</u> → Development of an open-source interpretable pK_a prediction tool based on machine learning which can predict the pK_a values of all titratable groups of a *drug-like* molecule.

Icons made by Smashicons from www.flaticon.com

Let’s start with monoprotic molecules

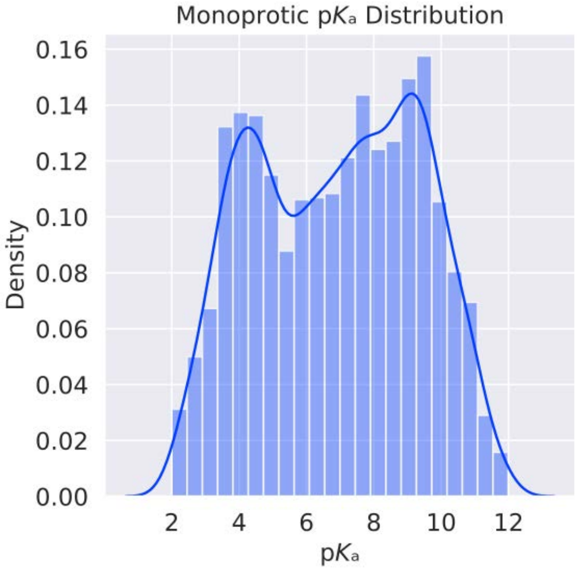
Baltruschat M and Czodrowski P. Machine learning meets pKa [version 2; peer review: 2 approved]. *F1000Research* 2020, **9**(Chem Inf Sci):113 (https://doi.org/10.12688/f1000research.22090.2)

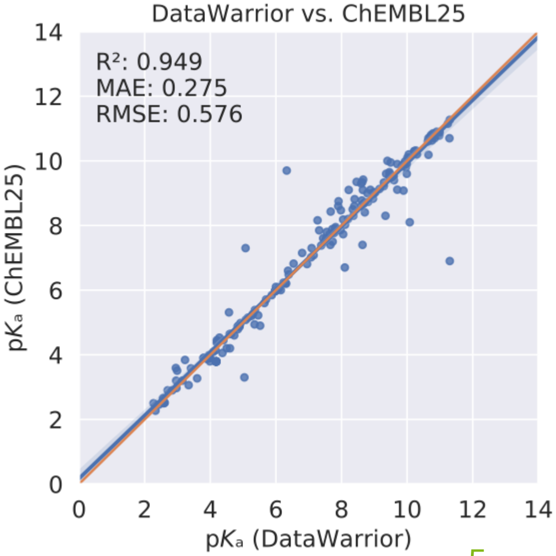
Monoprotic Dataset

- Using the curated *ChEMBL25* and *DataWarrior* datasets
- **5994** curated unique monoprotic structures

- No source is specified for the values from *DataWarrior*
- *ChEMBL25* data points are completely taken from literature

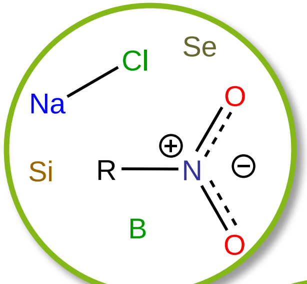
- Good correlation of the intersection

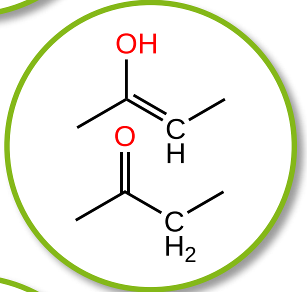


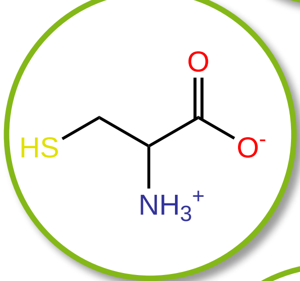


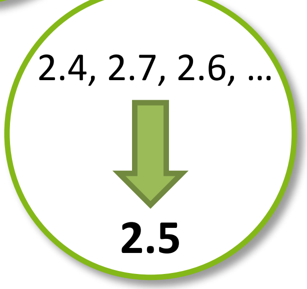
Data Curation

- Removal of salts, nitro groups, B, Se, Si
- Lipinski‘s rule of five (one violation allowed)
- pK_a between 2 and 12
- Tautomer standardization
- Protonation at pH 7.4
- Combination of data points from duplicated structures while removing outliers









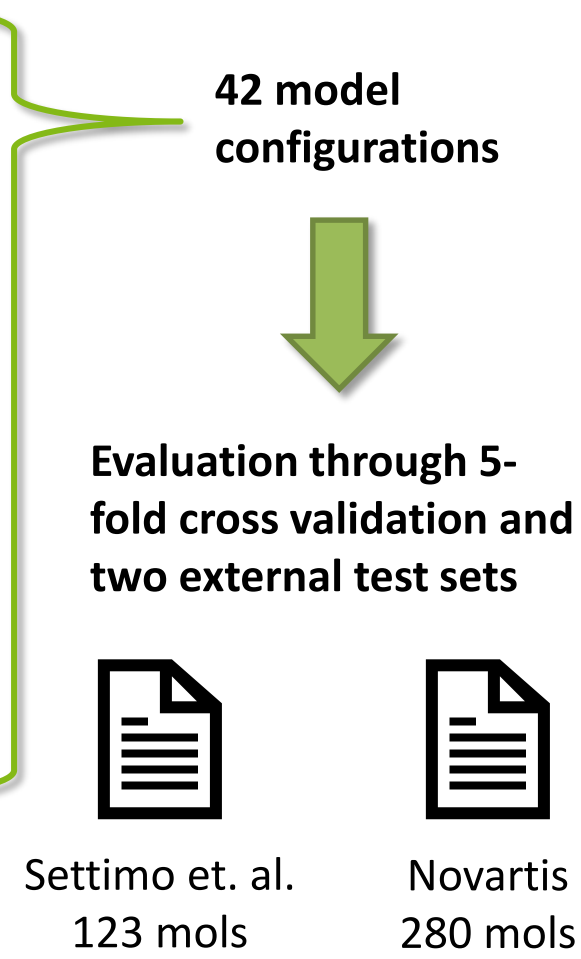
Machine Learning

Algorithms

- Random Forest (1000 trees)
- SVR (gamma=“auto”/”scale”)
- Neural Network (MLP, 3 different architectures)
- XGB

Training data

- 196/200 RDKit descriptors
- FeatureMorgan FP, radius 3, 4096 bits (FCFP6-like)
- Both combined
- Scaling for each of the three above



42 model configurations

Evaluation through 5-fold cross validation and two external test sets

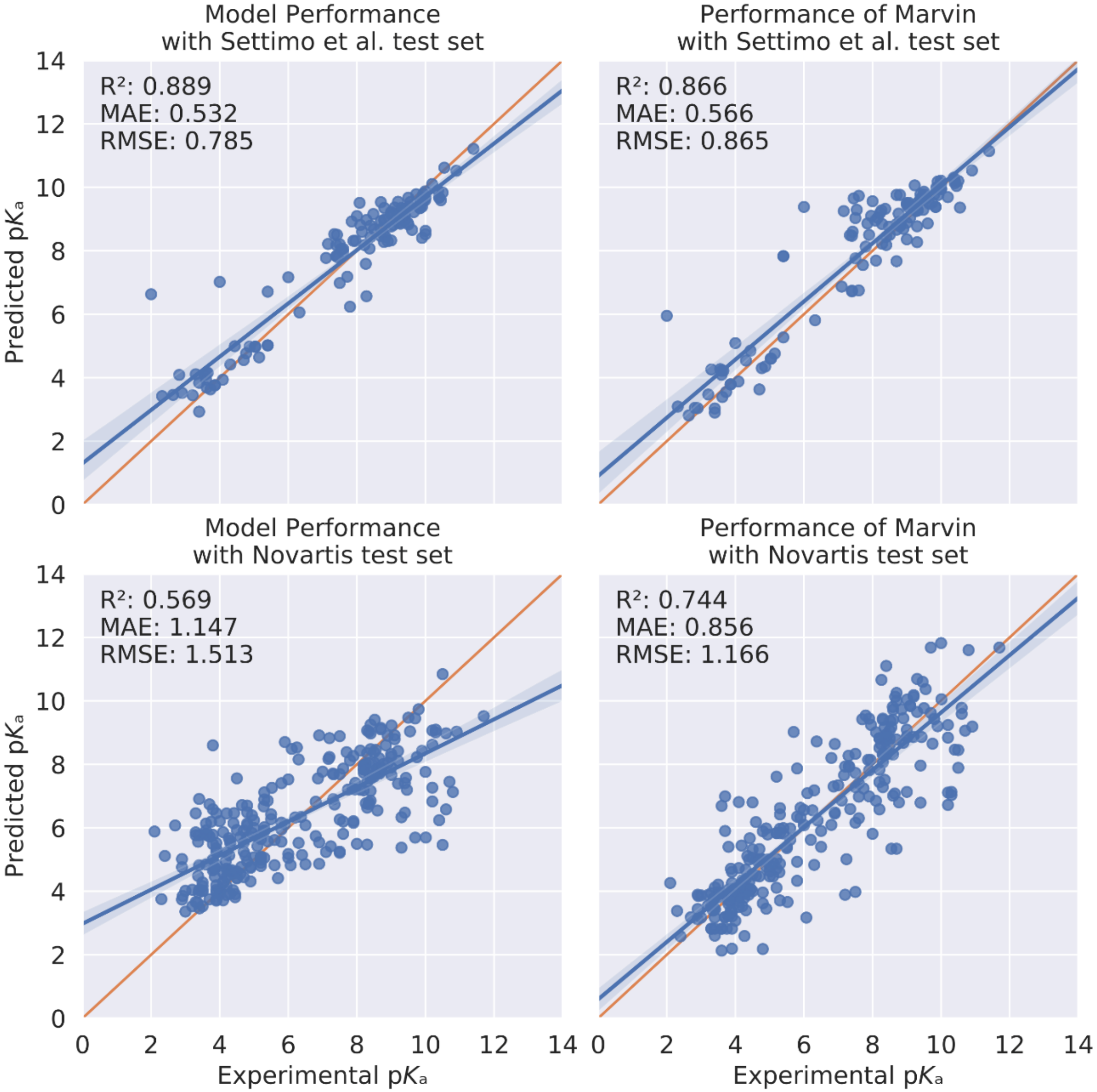
Settimo et. al.
123 mols

Novartis
280 mols

Best Results

No.	Model	Train data	MAE (CV)	RMSE (CV)	R^2 (CV)
#1	RF (1000 trees)	DESC+MF3 (scaled)	0.682	1.032	0.820
#2		DESC+MF3	0.683	1.032	0.820
#3		MF3	0.708	1.094	0.797
#4		MF3 (scaled)	0.708	1.094	0.797

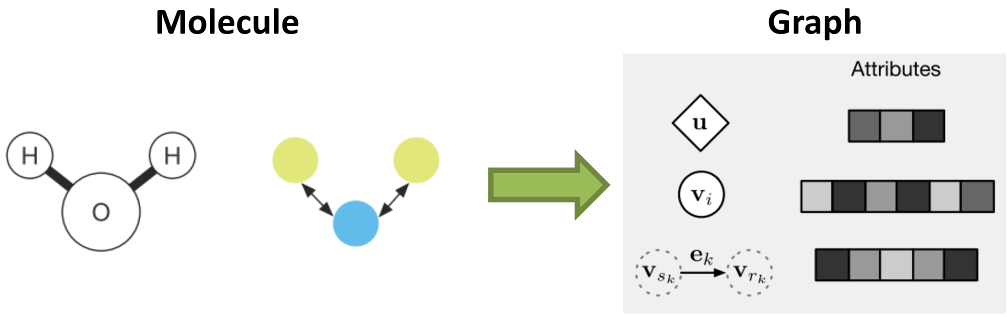
Best Results



Graph Convolutional Networks (GCN) and QML

With David Bushiri and Prof. Dr. Enrico Tapavicza

- *PyTorch Geometric* module for GCNs



Molecule

Graph

molecular property

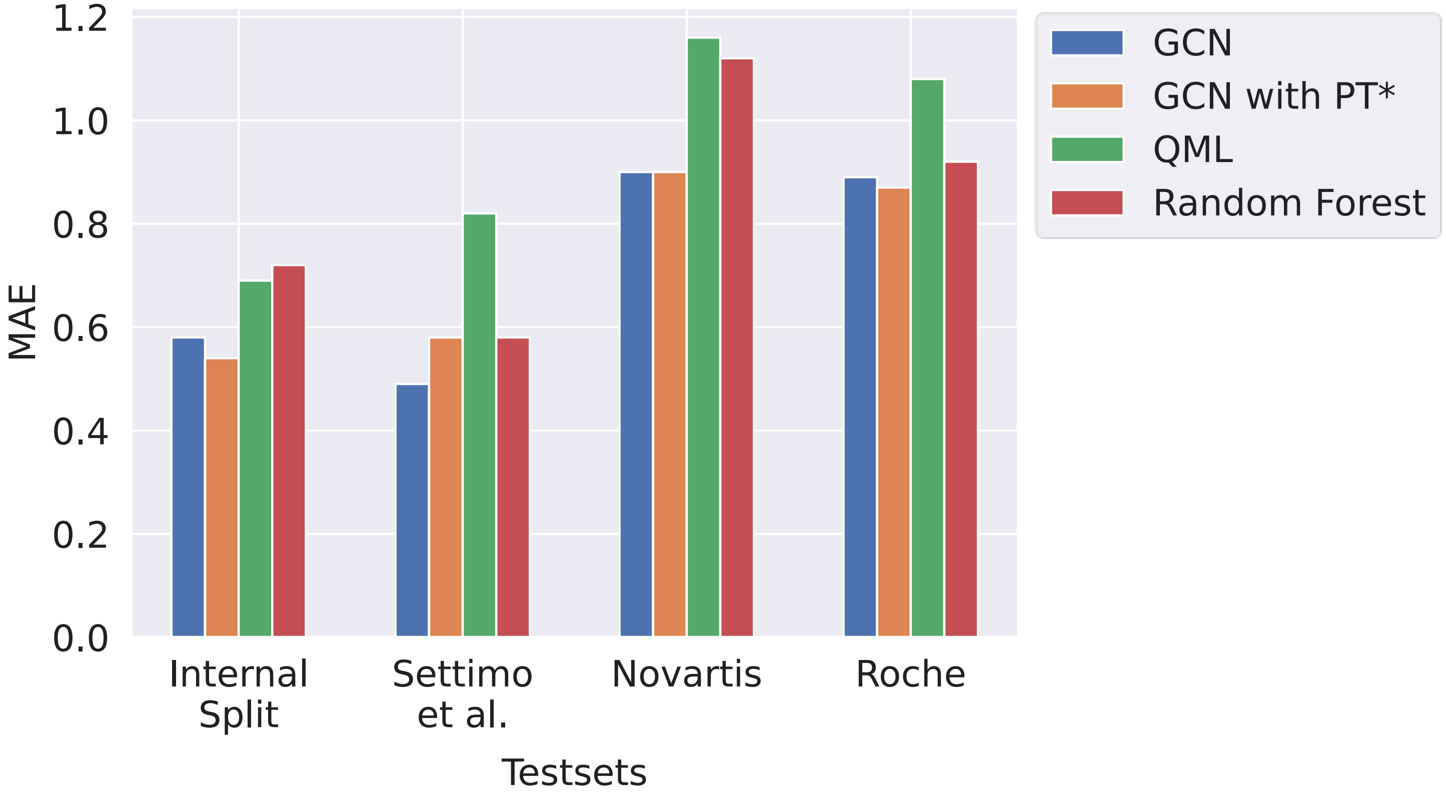
atoms

bonds

- *QML*: Kernel-ridge regression based method

➢ **Using 5196 DFT-optimized structures**

Best GCN and QML Results



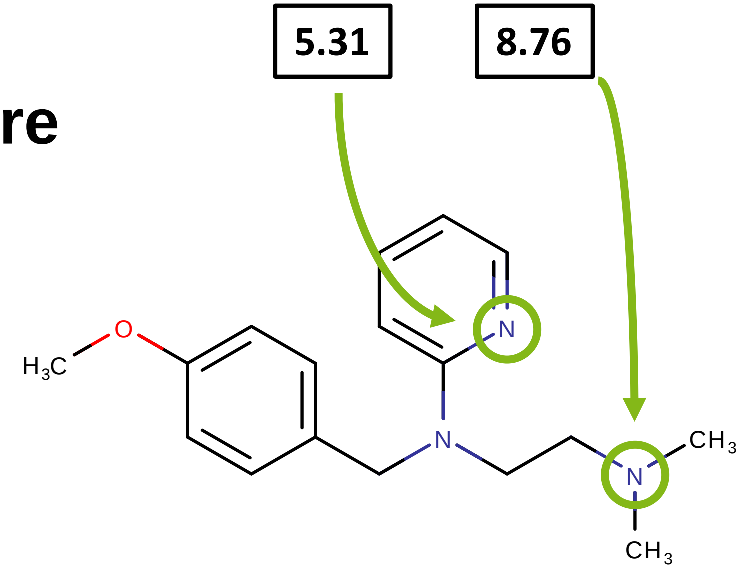
*Pretrained with 900 000 protomers from the ZINC dataset with polar desolvation energy used as target

Let’s go multiprotic!

pK_a Predictions of Multiprotic Molecules:
→ The first two problems to solve

- **Identify and locate titratable groups without licensed software**
  - **Must be done for training and every prediction**

- Assign the pK_a values from the datasets to the related titratable groups
  - Must be done only for training set



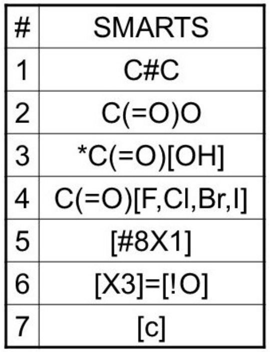
The Idea

1. Identify the titratable groups with available tools
2. Generate a hardcoded list of SMARTS pattern from the tool results that covers all major groups

- Used Tools:

Marvin Suite

Dimorphite-DL



#	SMARTS
1	C#C
2	C(=O)O
3	*C(=O)[OH]
4	C(=O)[F,Cl,Br,I]
5	[#8X1]
6	[X3]=[!O]
7	[c]

Datasets

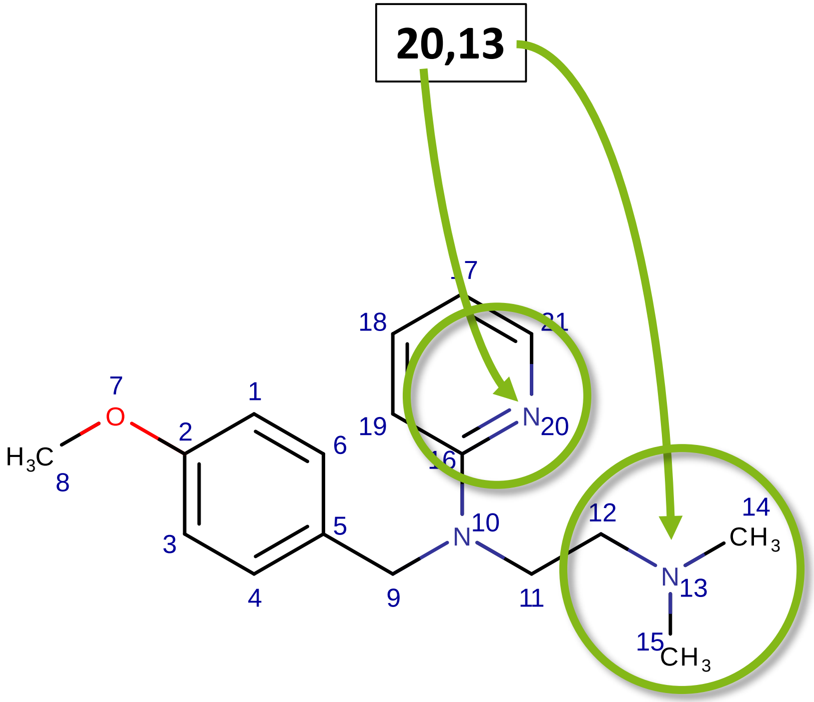
Source	pK_a Values	Unique Molecules
ChEMBL26	8503	6617
DataWarrior	7911	7463
Hunt et al.	2488	2277
Settimo et al.	612	511
Literature Compilation	1765	1353
SAMPL6	31	24
Novartis	1025	646
Roche	1762	1738
OpenEye	55322	23875
Total (curated)	**49349**	**17538**

ChemAxon Marvin

Marvin provides atom ids for titratable groups during pK_a calculation

Extract environment around all atom ids

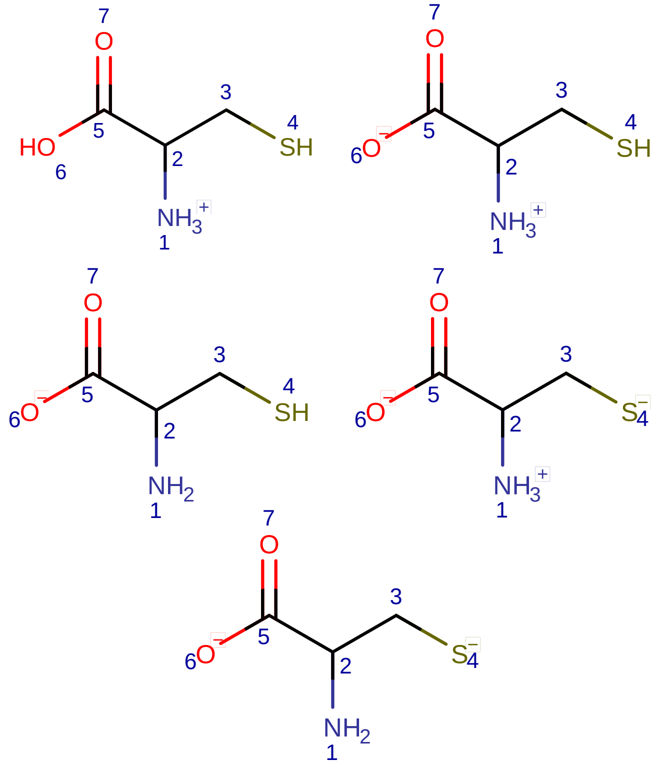
Group by environment and count



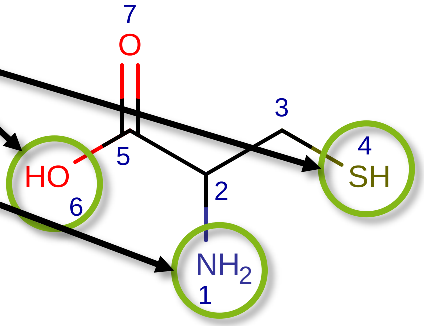
https://git.durrantlab.pitt.edu/jdurrant/dimorphite_dl

Dimorphite-DL

- Calculates all possible microstates of a molecule in a specified pH range

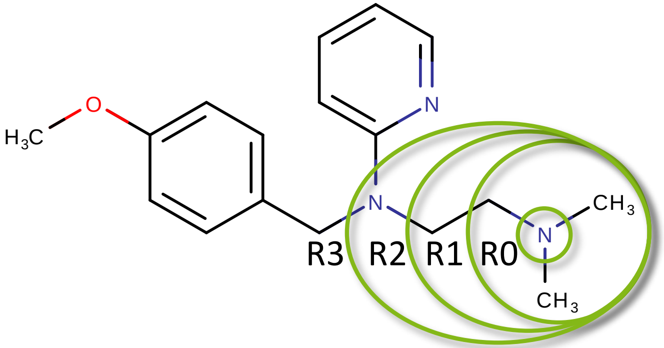


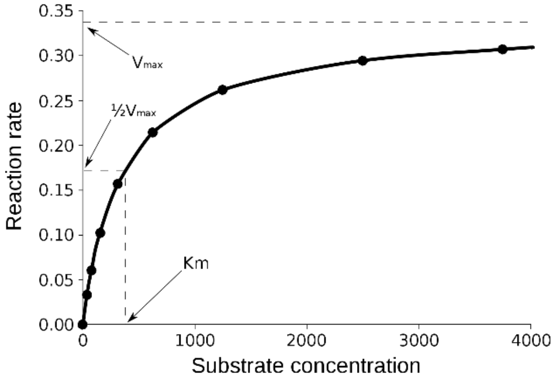
1	2	3	4	5	6	7
1	0	0	0	0	0	0
1	0	0	0	0	-1	0
0	0	0	0	0	-1	0
1	0	0	-1	0	-1	0
0	0	0	-1	0	-1	0



Overview

- For both Marvin and Dimorphite-DL
  - Investigate environment distributions with radius 0 to 6
  - Looking at the saturation curves, how many environments / groups do we *really* need?

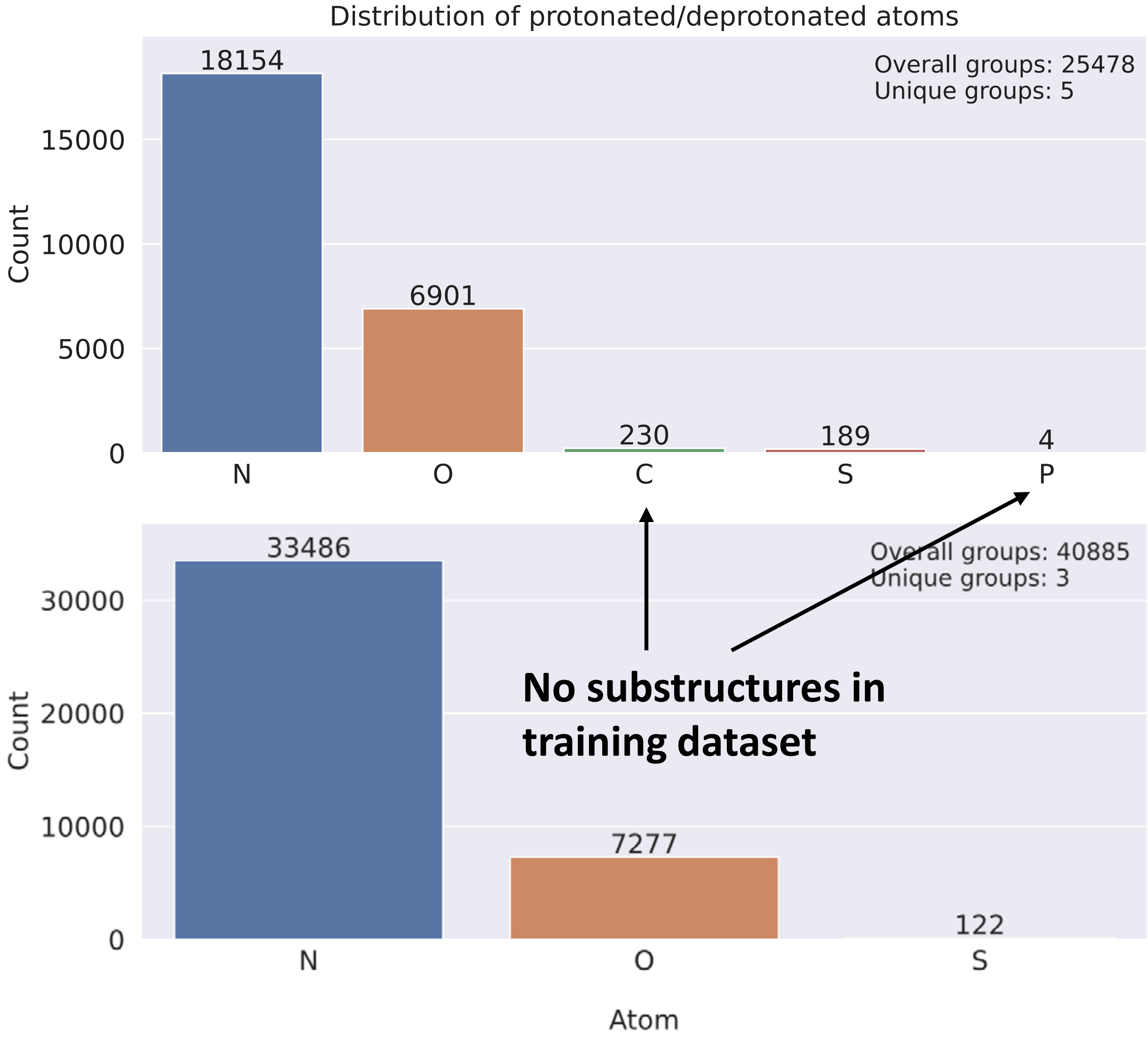




Radius 0

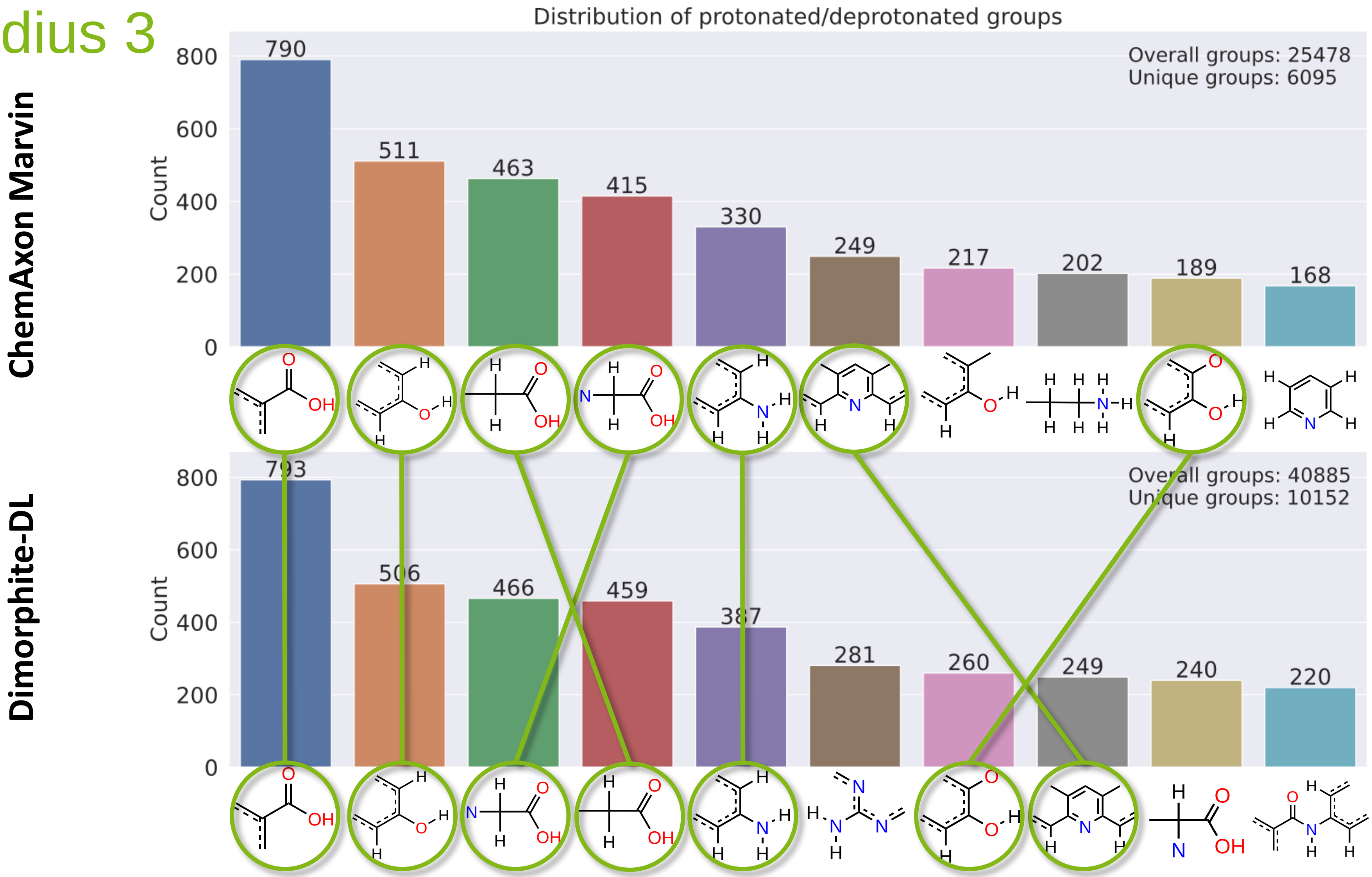
ChemAxon Marvin

Dimorphite-DL



No substructures in training dataset

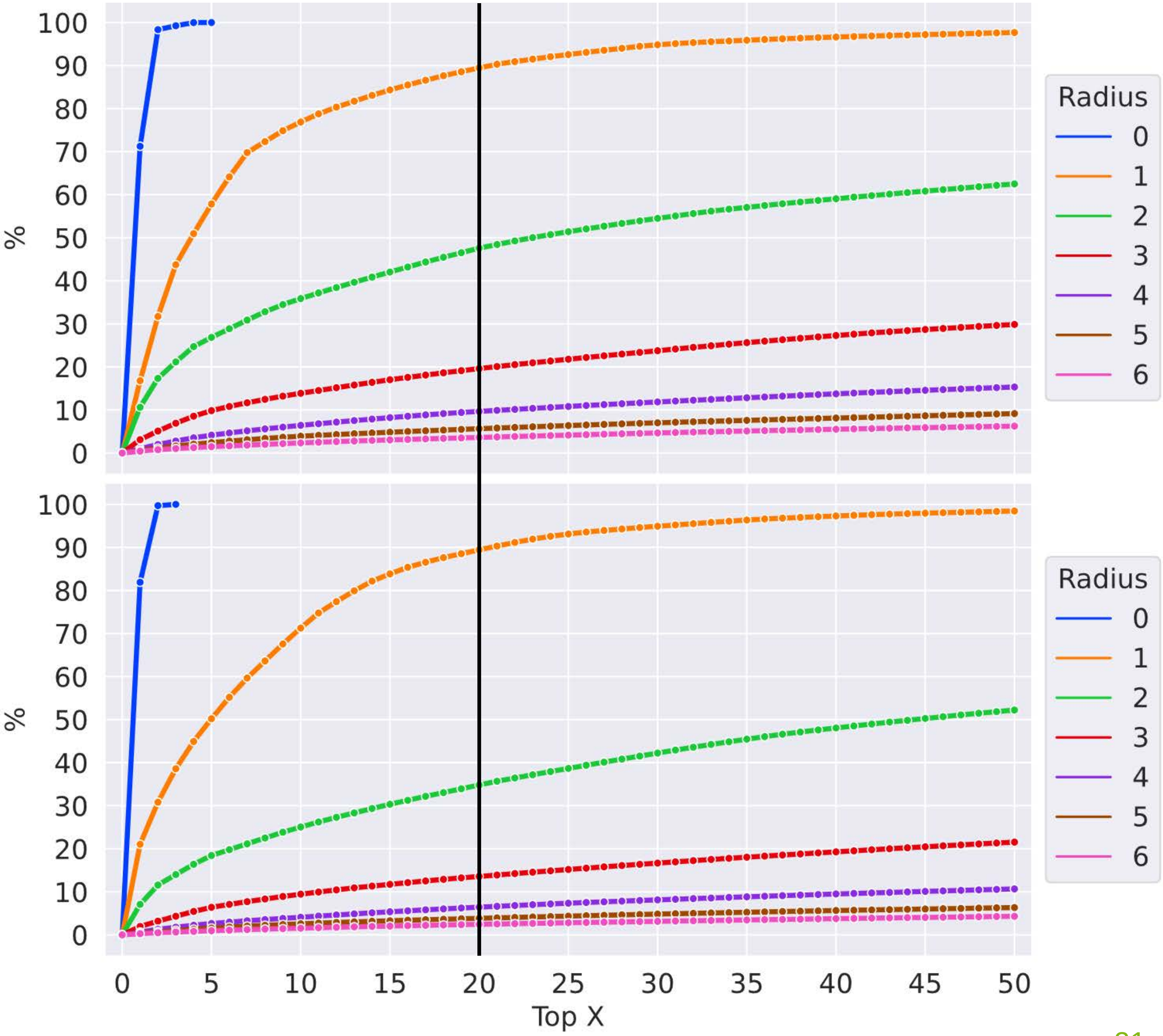
Radius 3


Distribution of protonated/deprotonated groups
ChemAxon Marvin
Count
800
600
400
200
0
790
511
463
415
330
249
217
202
189
168
Overall groups: 25478
Unique groups: 6095
Dimorphite-DL
Count
800
600
400
200
0
793
506
466
459
387
281
260
249
240
220
Overall groups: 40885
Unique groups: 10152

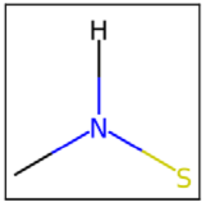
Saturation Curve

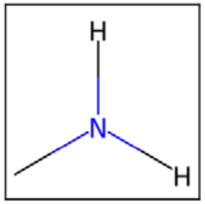
ChemAxon Marvin

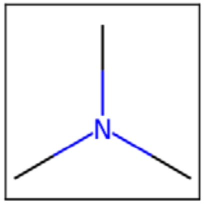
Dimorphite-DL

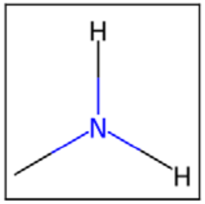


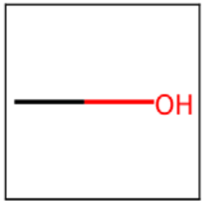
Result → 24 Titratable Fragments

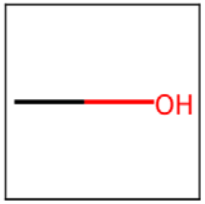

CNS

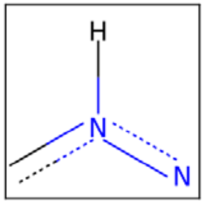

CN

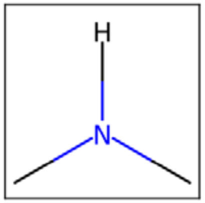

cN(C)C

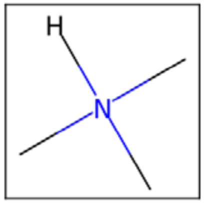

cN

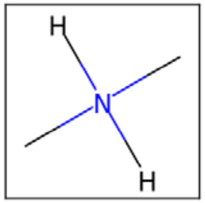

CO

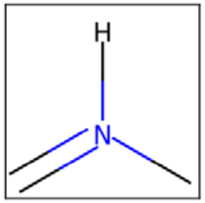

cO

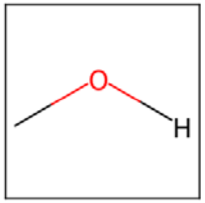

c[nH]n

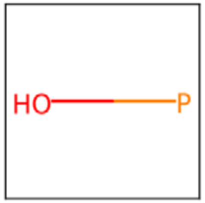

CNC

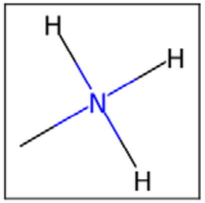

C[NH](C)C

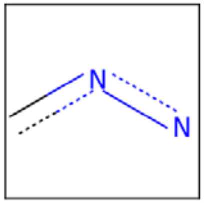

C[NH2]C

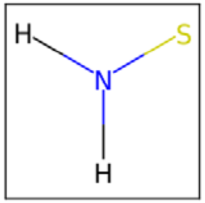

C[NH]=C

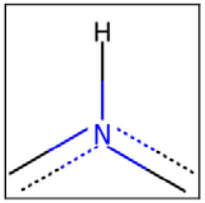

CO

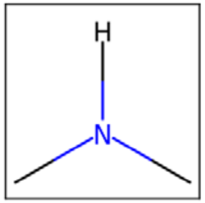

OP

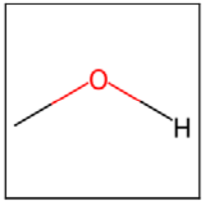

C[NH3]

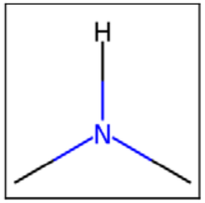

cnn

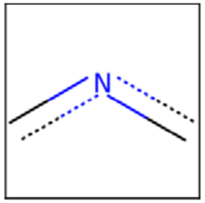

NS

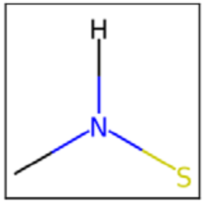

c[nH]c

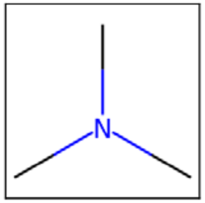

cNc

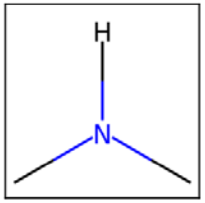

cO


CNc


cnc

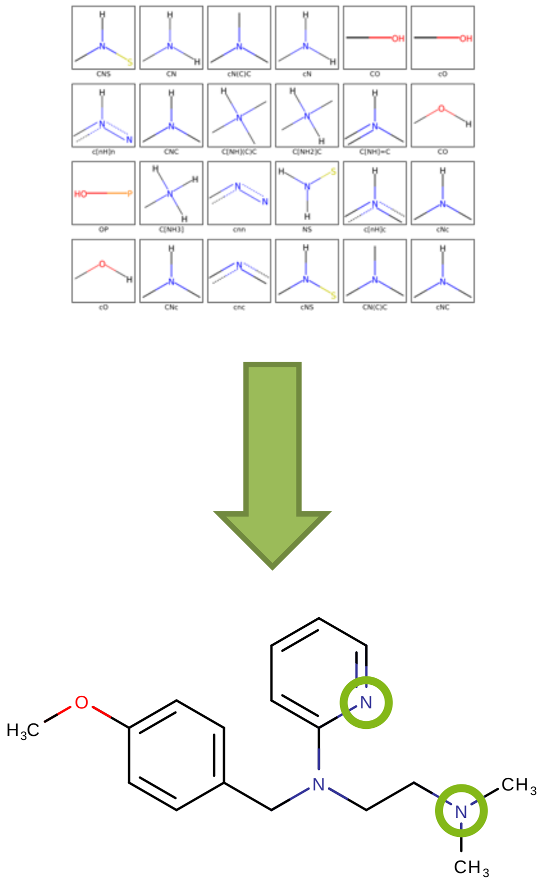

cNS


CN(C)C

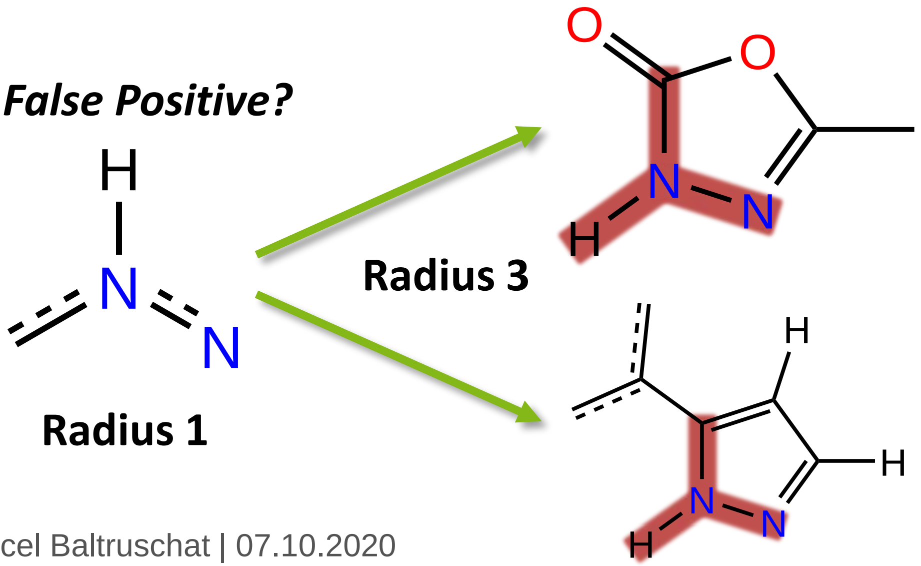

cNC

Validation

- Find the locations with the extracted titratable fragments
- Validate for all radii
- Test a hierarchical structure

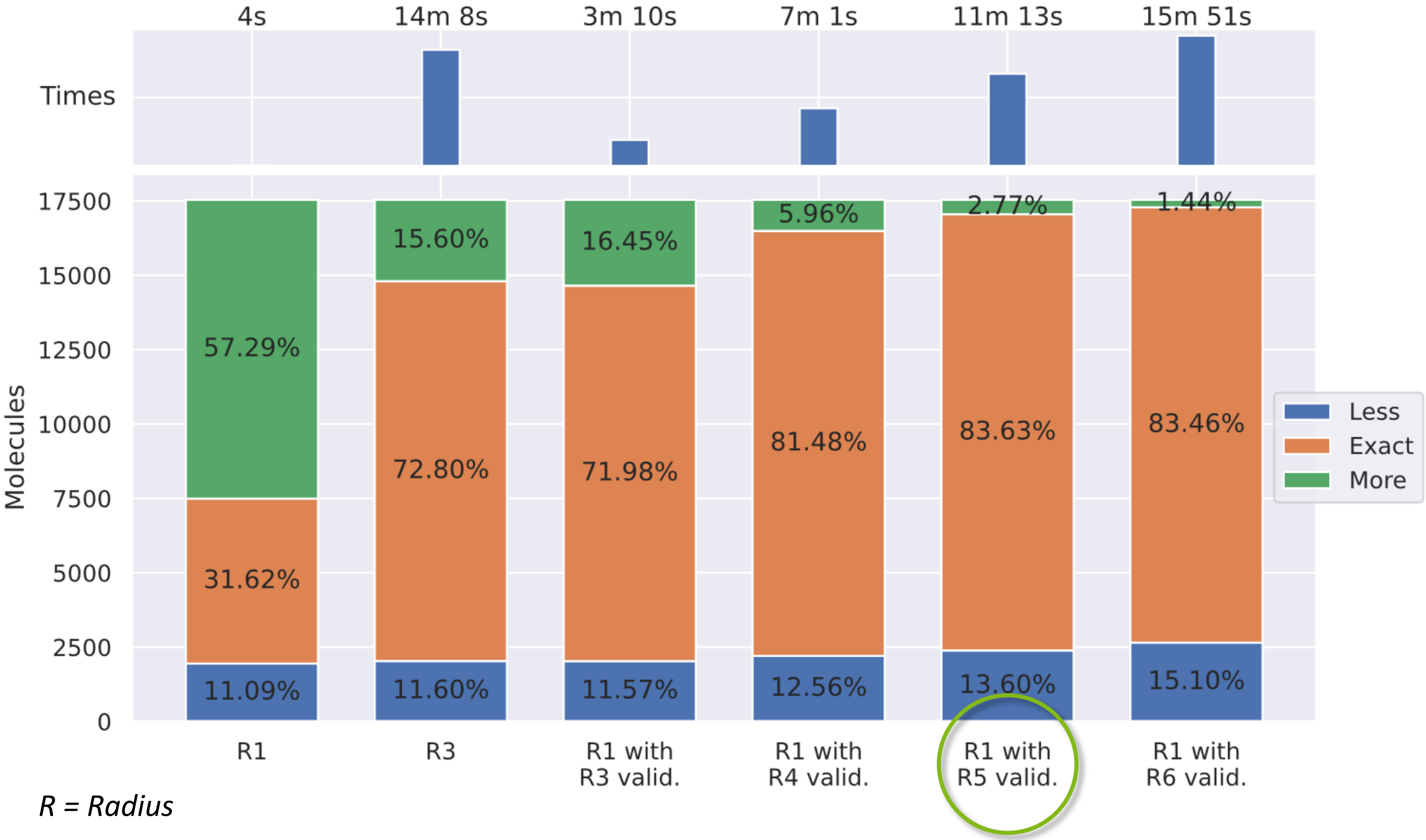


False Positive?



Radius 3

Radius 1

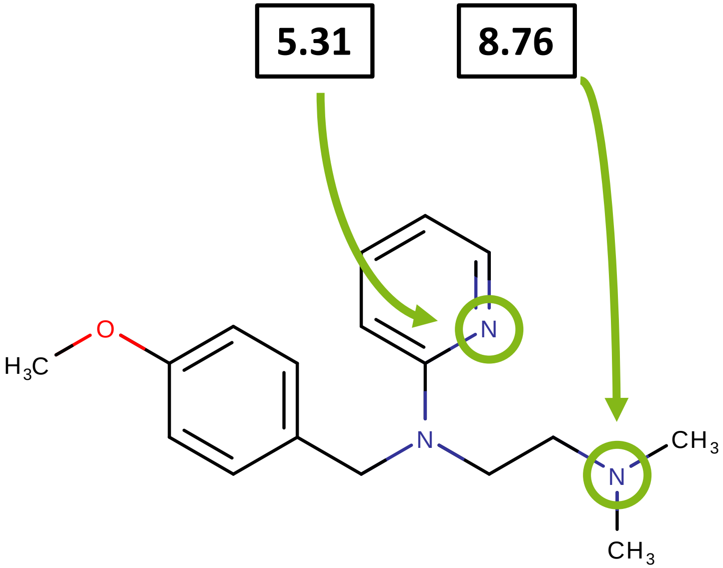

4s
14m 8s
3m 10s
7m 1s
11m 13s
15m 51s
Times
17500
15000
12500
10000
7500
5000
2500
0
Molecules
57.29%
31.62%
11.09%
15.60%
72.80%
11.60%
16.45%
71.98%
11.57%
5.96%
81.48%
12.56%
2.77%
83.63%
13.60%
1.44%
83.46%
15.10%
Less
Exact
More
R1
R3
R1 with R3 valid.
R1 with R4 valid.
R1 with R5 valid.
R1 with R6 valid.

R = Radius

pK_a Predictions of Multiprotic Molecules: → The first two problems to solve

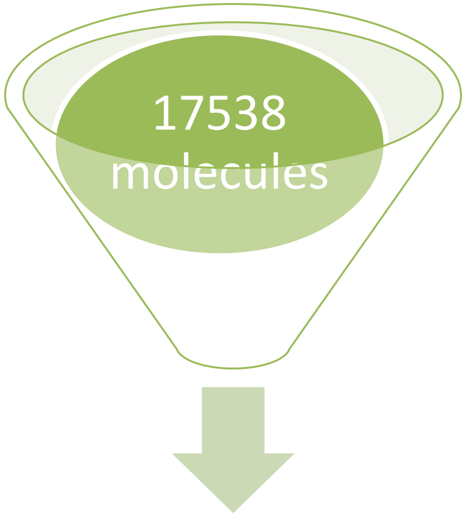
- Identify and locate titratable groups without licensed software
  - Must be done for training and every prediction

- **Assign the pK_a values from the datasets to the related titratable groups**
  - **Must be done only for training set**



Assign Values to Groups

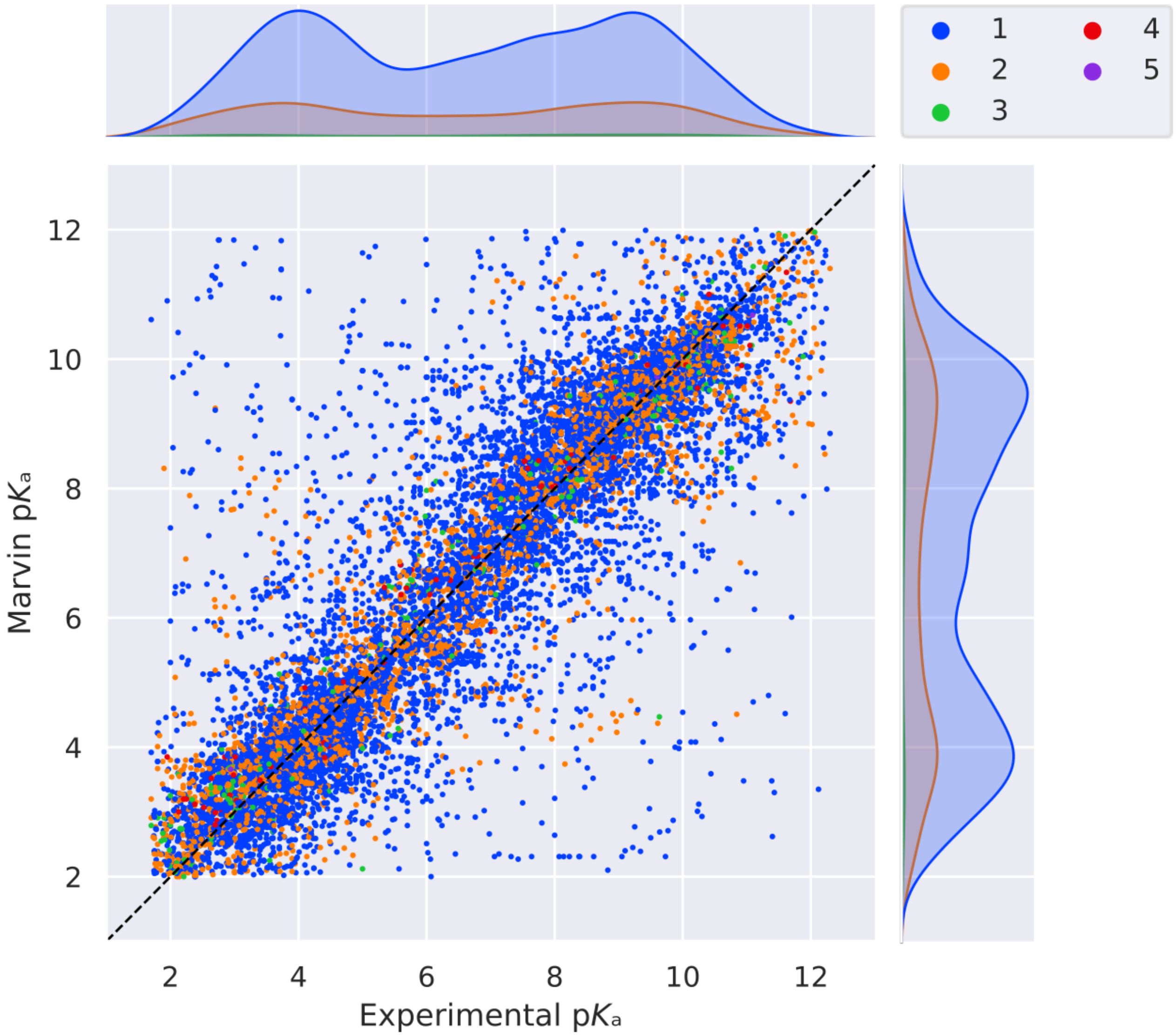
- Combine values that apparently belong to the same titratable group
  - Error range of 0.3 pK_a units
- Find the experimental value that comes closest to the corresponding *Marvin* prediction
- Only consider “exact matches” for now



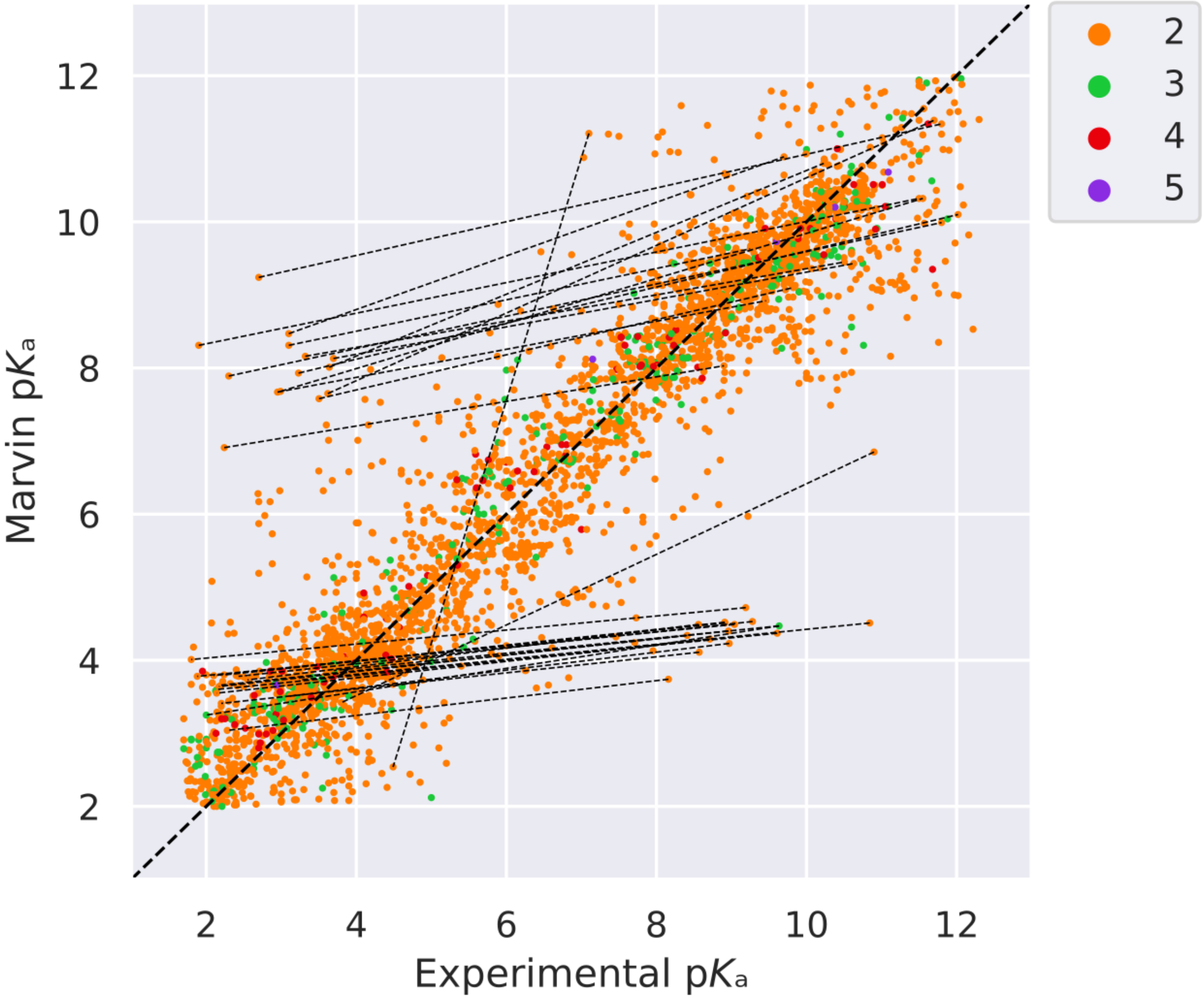
17538 molecules

11305 molecules

Results

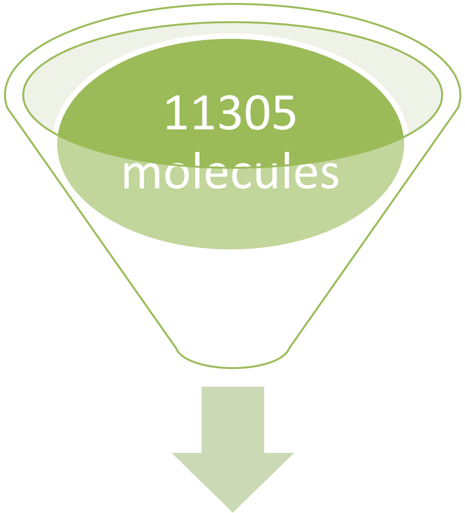


Results



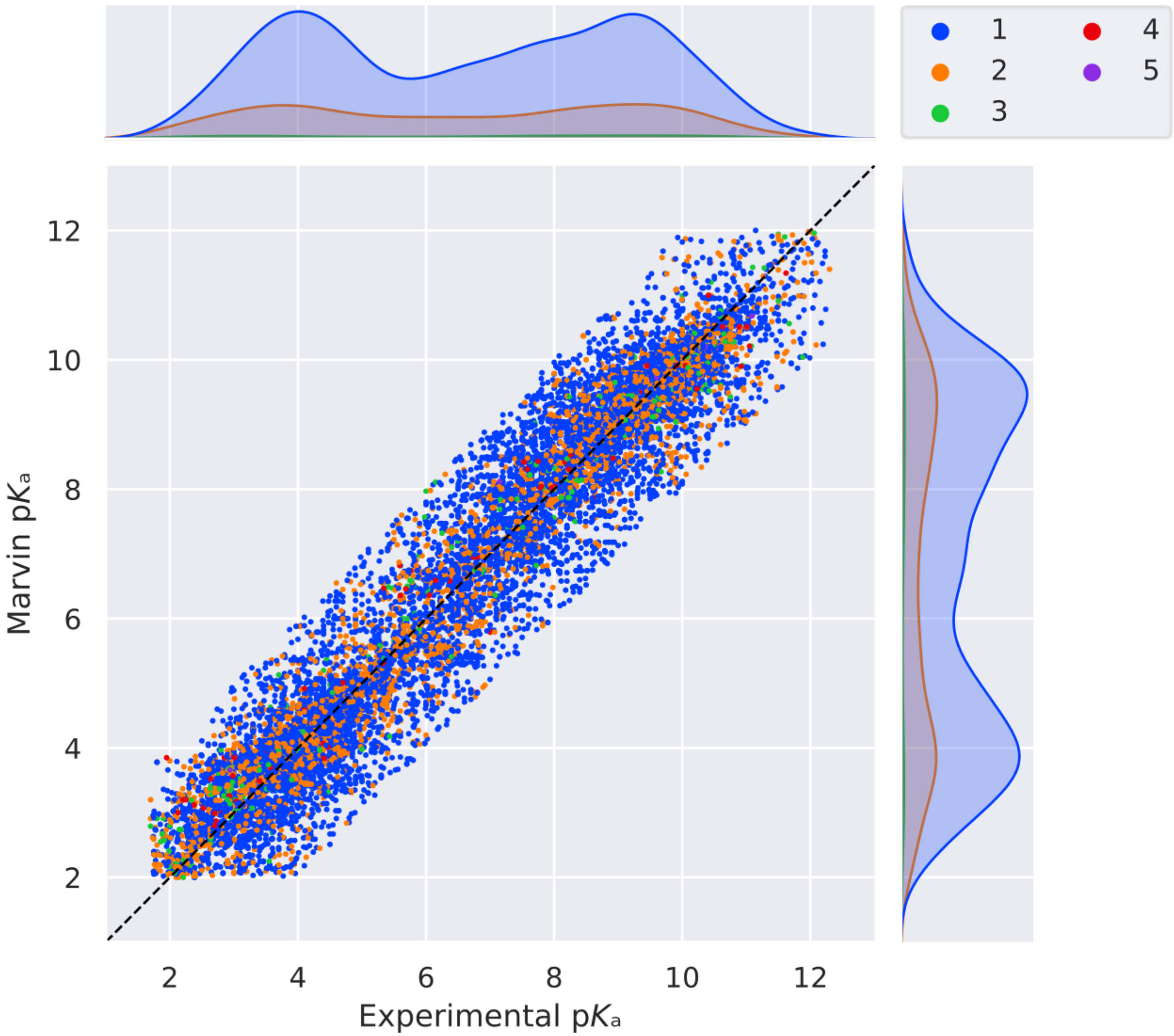
Results

- Cut at max. error = 2



11305 molecules

10203 molecules



Outlook

- Further investigation of the results and testing with other prediction tools
- Improvement of SMARTS pattern extraction
- Reducing amount of rejects through value assignment
- Replace OpenEye tautomers with RDKit integrated MolVS
- Start with machine learning for multiprotic molecules
- Develop and publish an easy-to-use toolkit

Acknowledgements

CzodrowskiLab
Paul Czodrowski
David Bushiri

California State University, USA
Enrico Tapavicza

Novartis Pharma AG
Richard A. Lewis
Stephane Rodde

Roche Pharma AG
Christian Kramer

Bayer AG
Michael E. Beck