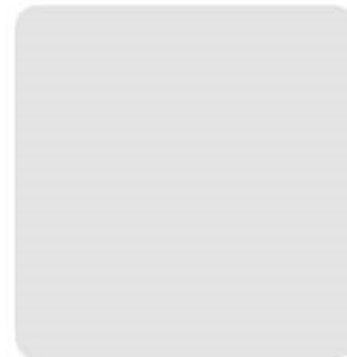


Open source application for small molecule pK_a predictions

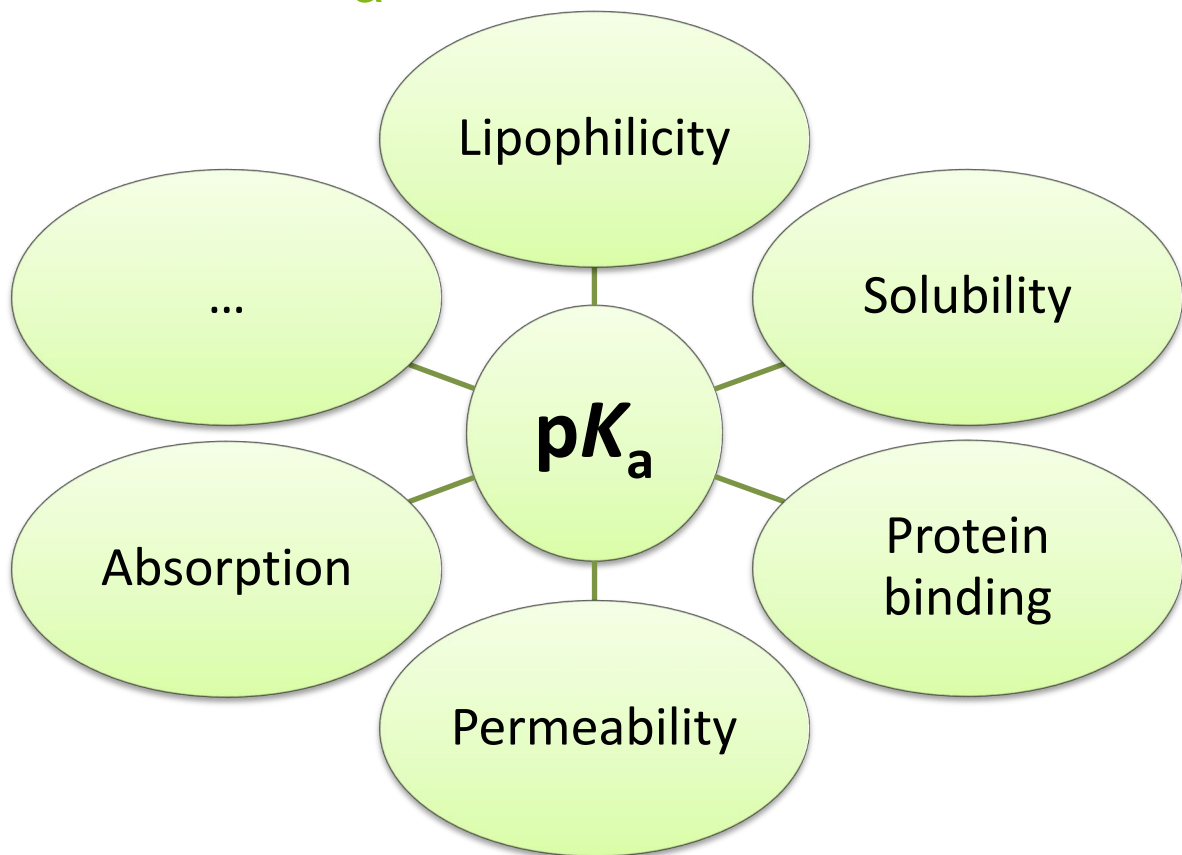
Marcel Baltruschat @CzodrowskiLab

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

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

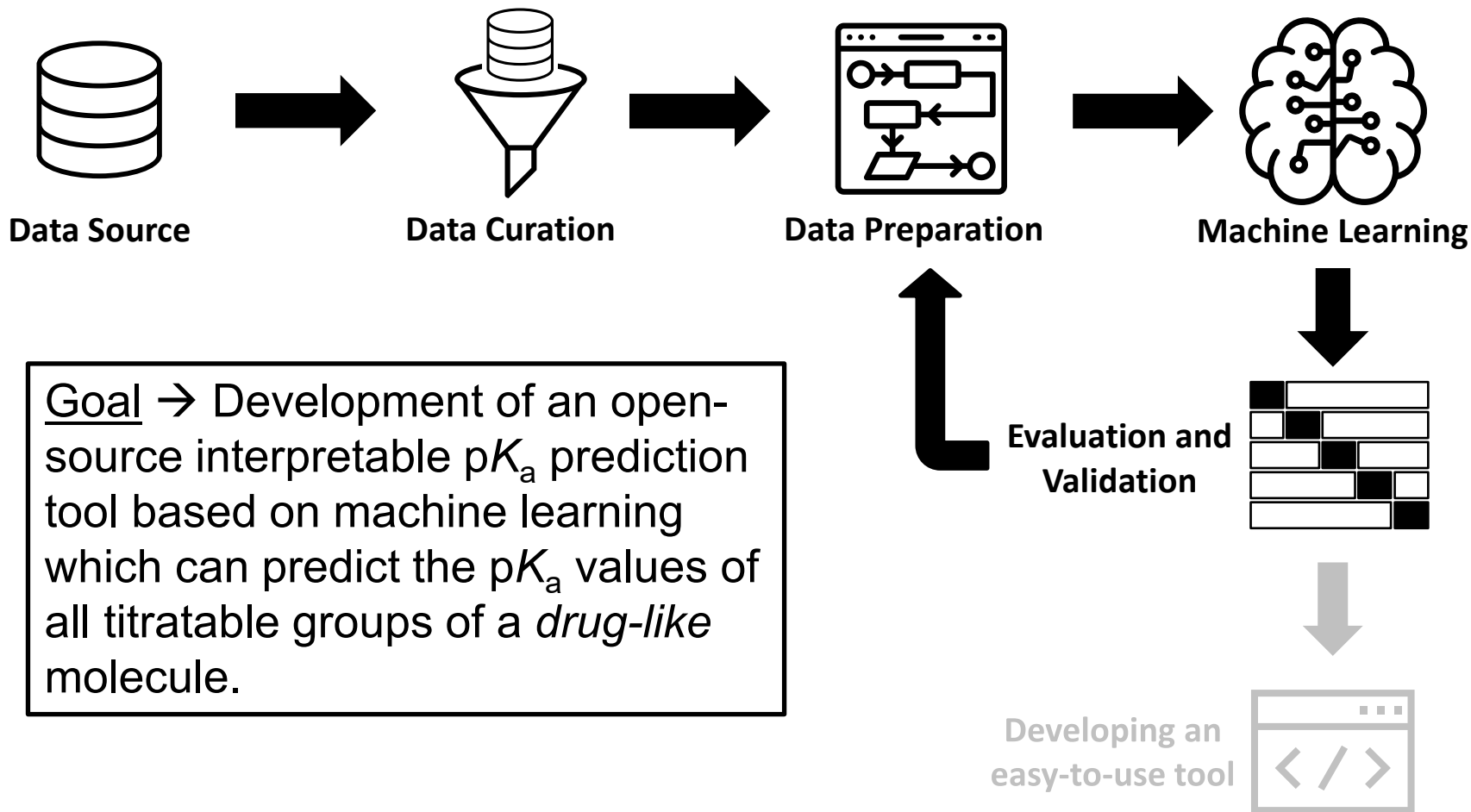


Why pK_a ?



- 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

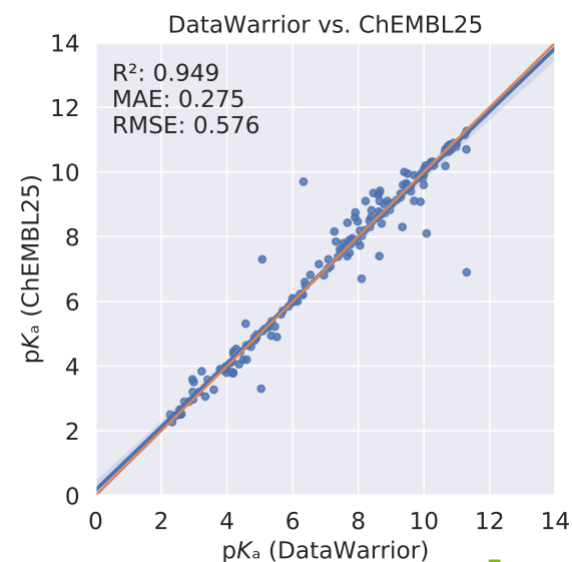
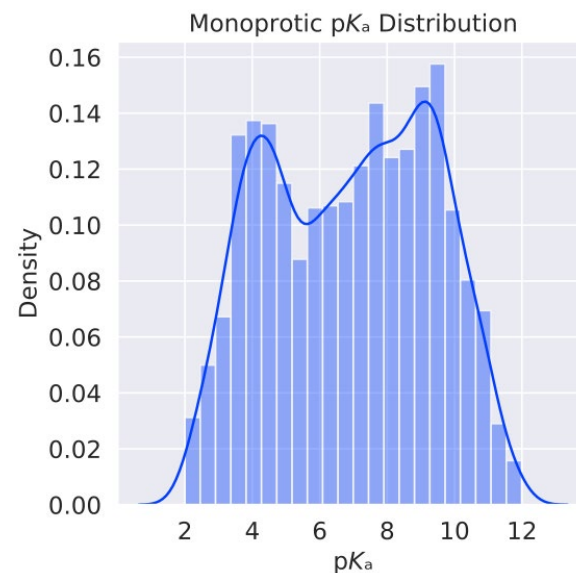


Let's start with monoprotic molecules

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



Machine Learning

Algorithms

- Random Forest (1000 trees)
- Support Vector Regression
- Neural Network (MLP, 3 different architectures)
- Extreme Gradient Boosting

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**

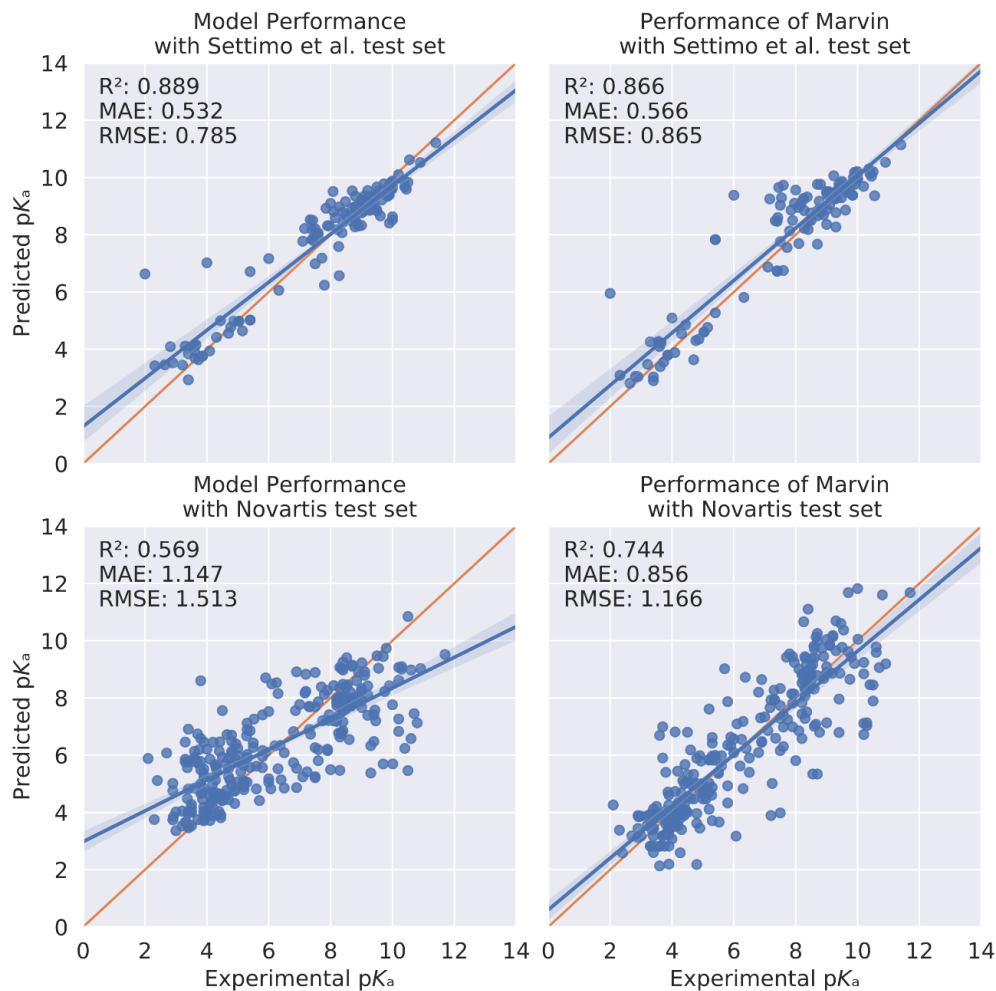


Settimo et. al.
123 mols



Novartis
280 mols

Best Results

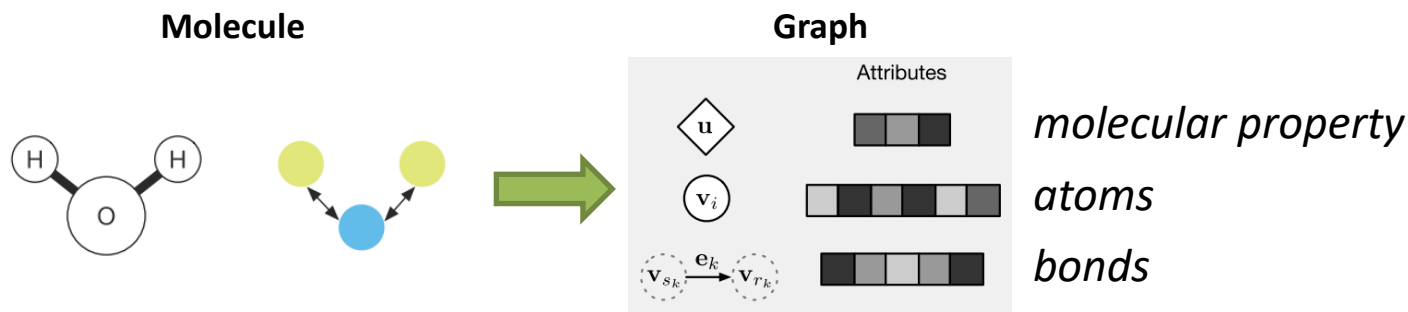


No.	Model	Train data	MAE (CV)	RMSE (CV)	R^2 (CV)
#1	RF (1000 trees)	Descriptors + FP	0.683	1.032	0.820

Graph Convolutional Networks (GCN) and QML

With David Bushiri and Prof. Dr. Enrico Tapavicza

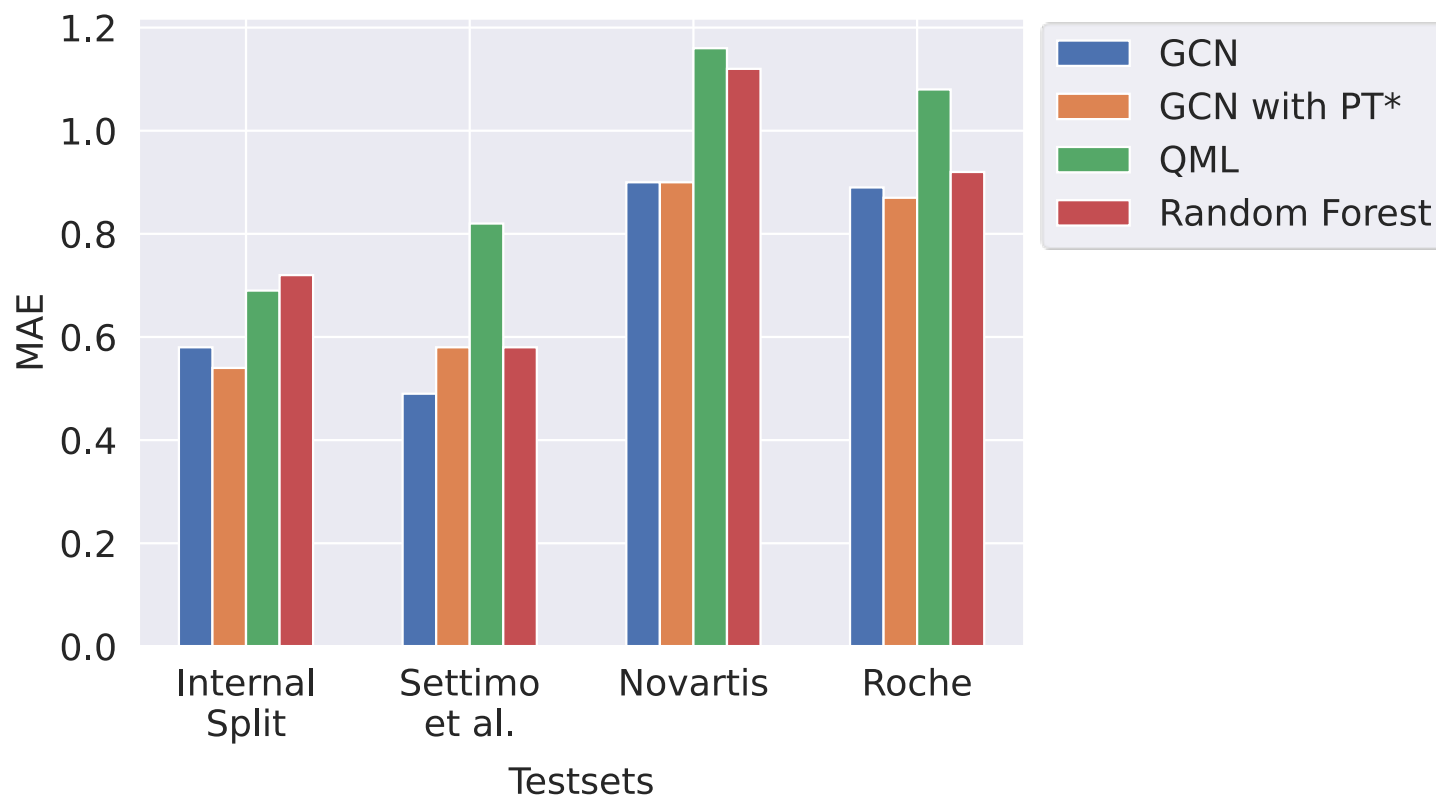
- *PyTorch Geometric* module for GCNs



- *QML*: Kernel-ridge regression based method

➤ **Using 5196 DFT-optimized structures**

Best GCN and QML Results



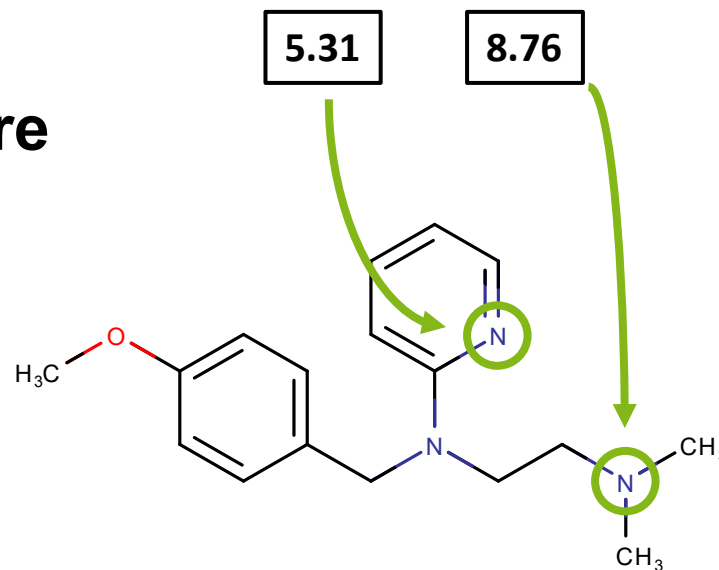
*Pretrained with 900 000 protomers from the ZINC dataset with polar desolvation energy used as target

Let's go multiprotic!

pK_a Predictions of Multiprotic Molecules:

→ The first two problems to solve

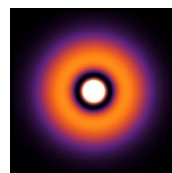
- **Identify and locate titratable groups without licensed software**
 - Must be done for training and every prediction
- Assign the pK_a values from the datasets to the related titratable groups
 - Must be done only for training set



The Idea

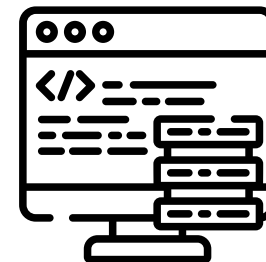
1. Identify the titratable groups with available tools
2. Generate a hardcoded list of SMARTS pattern from the tool results that covers all major groups

- Used Tools:



Dimorphite-DL

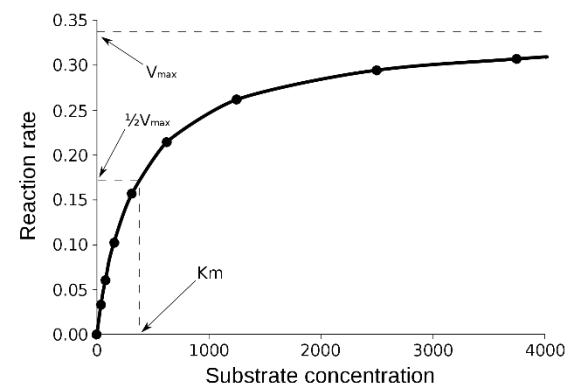
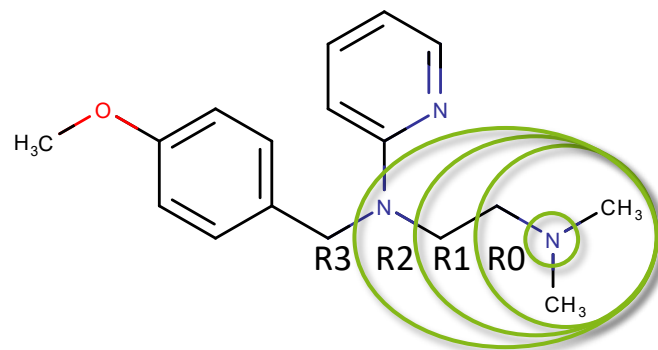
- **17538** unique molecules



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

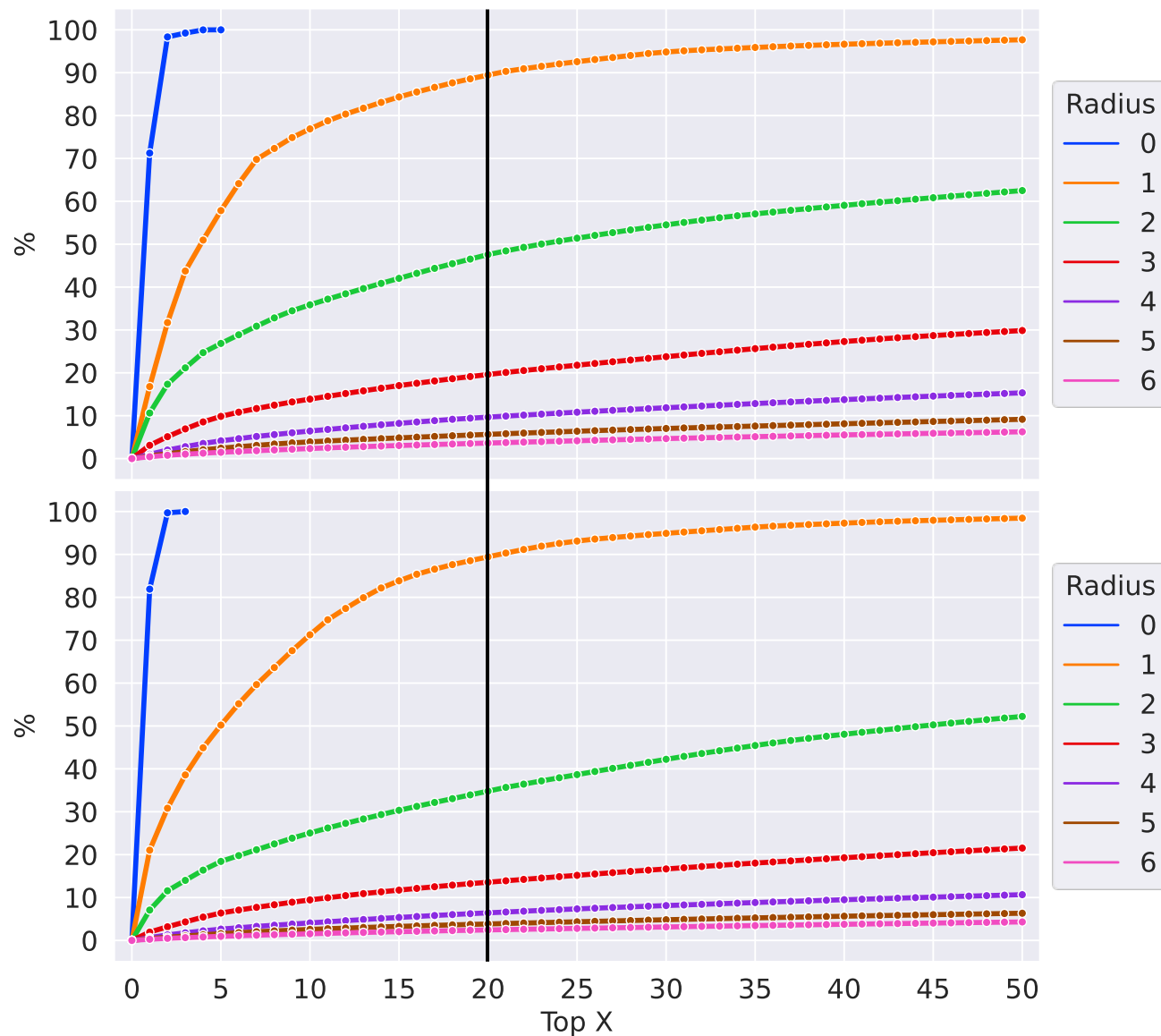
Overview

- 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?

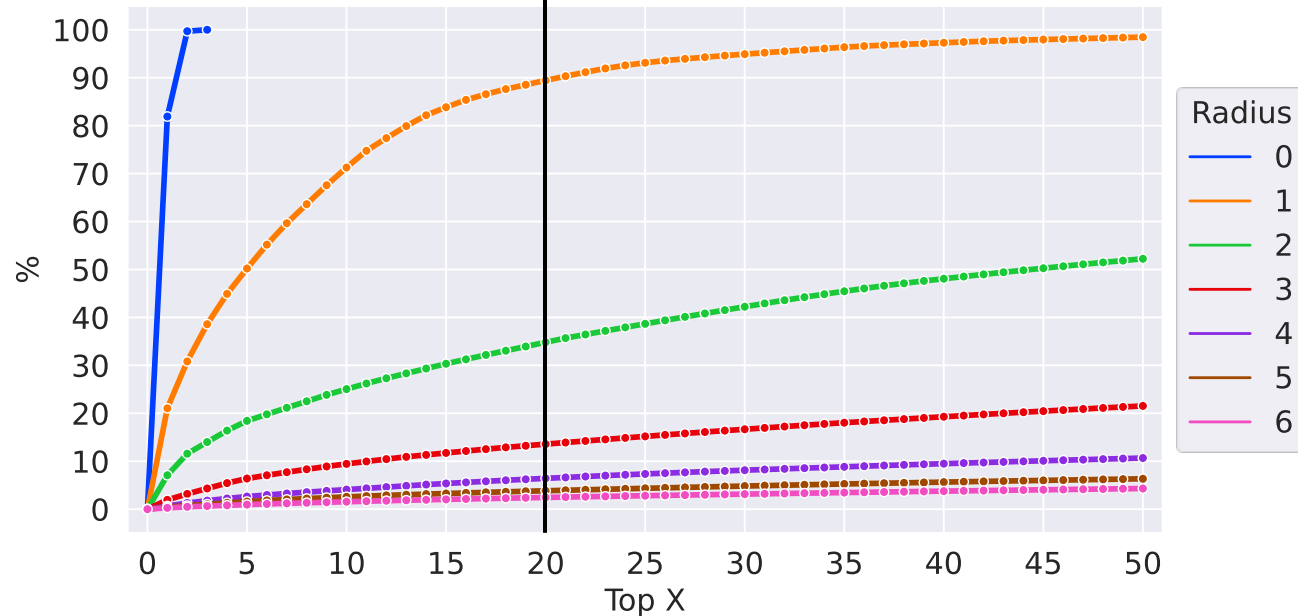


Saturation Curve

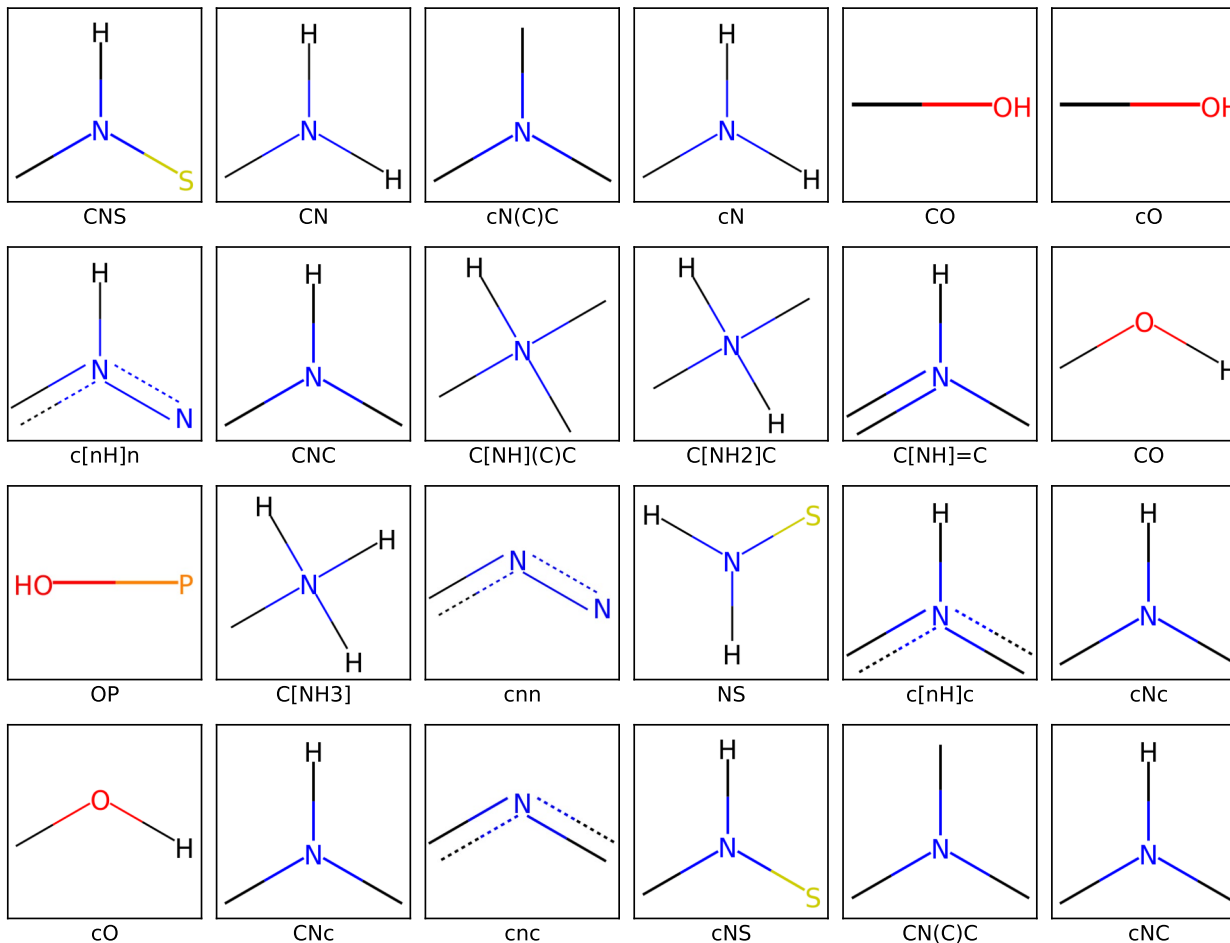
ChemAxon Marvin



Dimorphite-DL

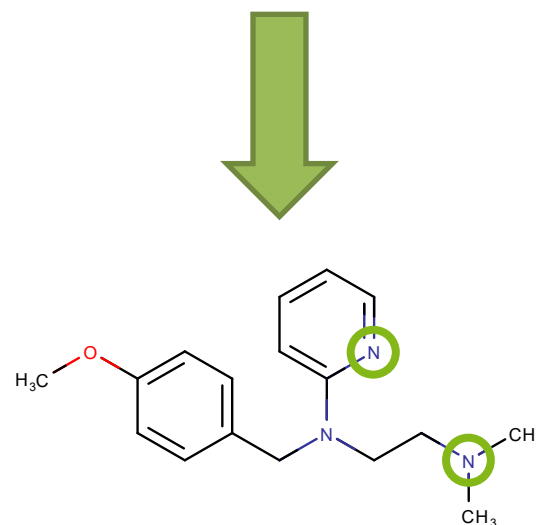
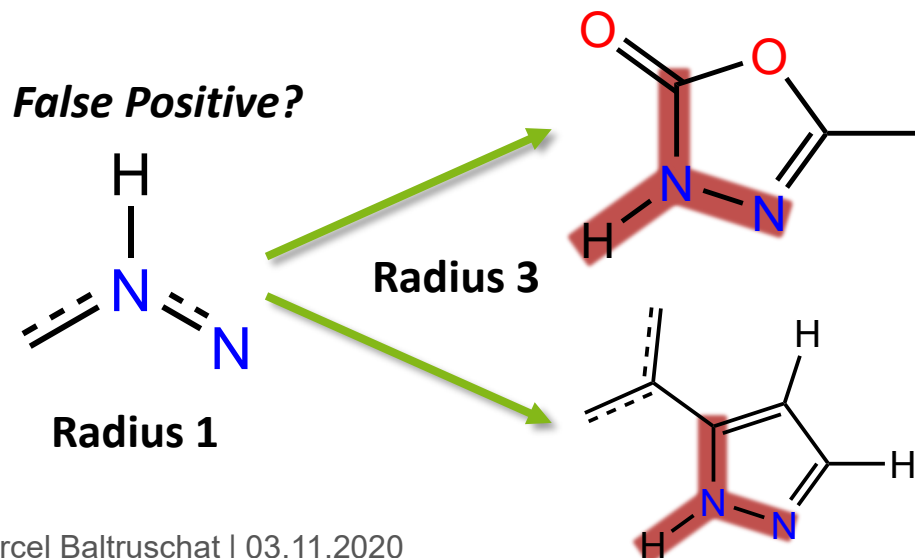
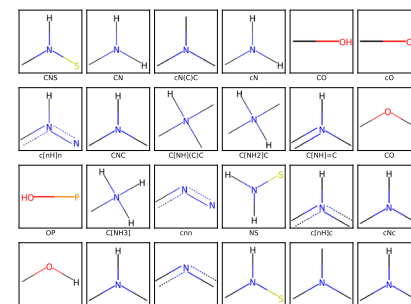


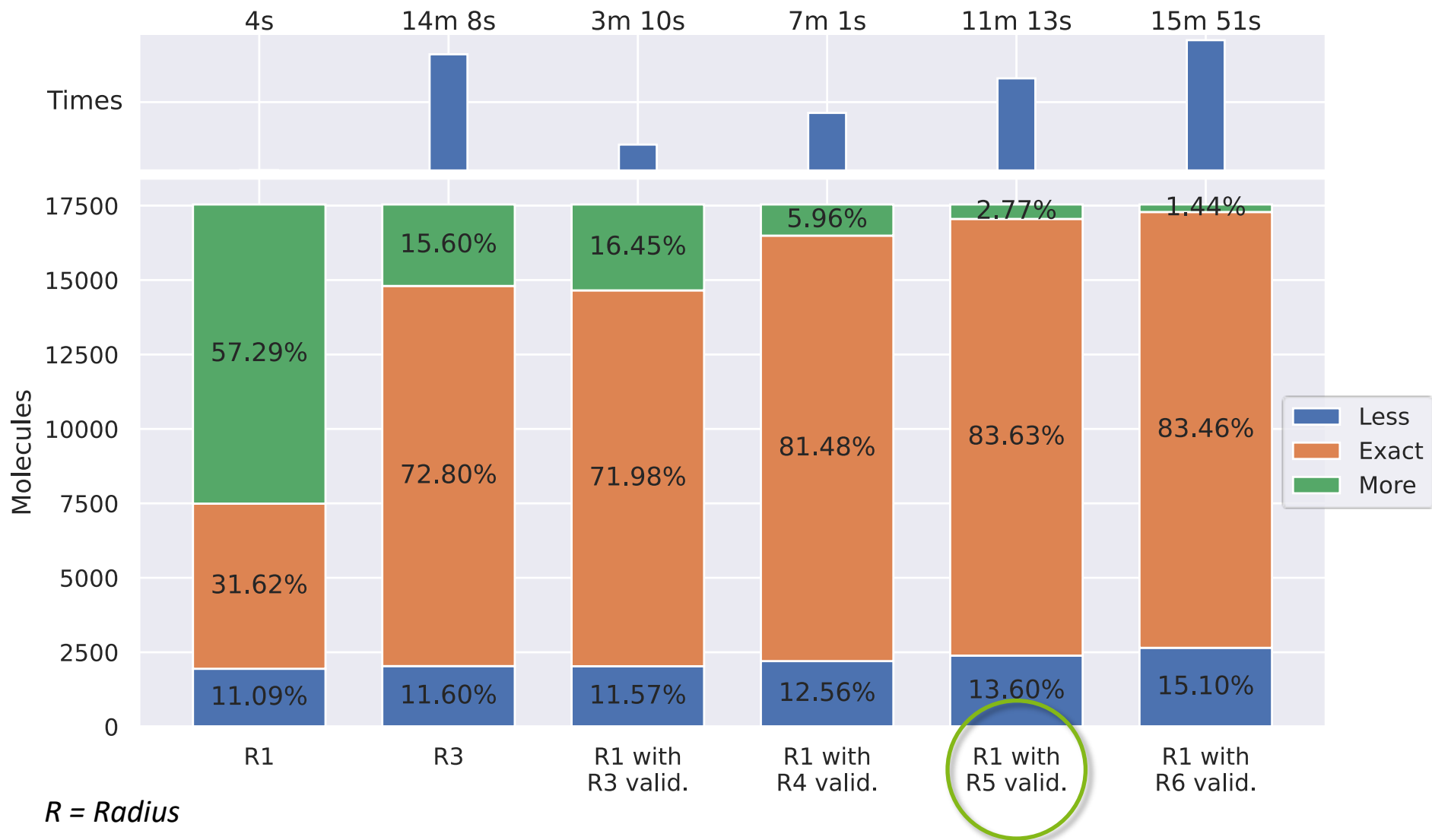
Result → 24 Titratable Fragments



Validation

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

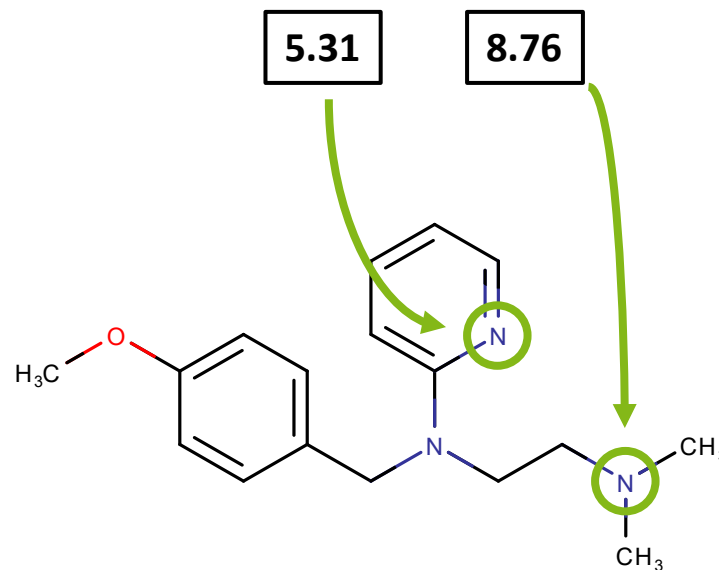




pK_a Predictions of Multiprotic Molecules:

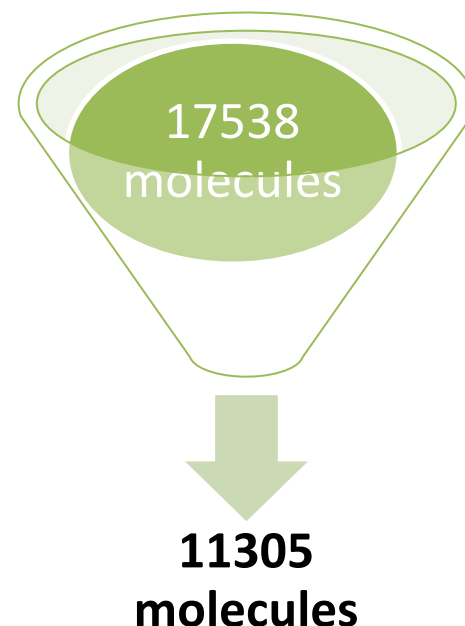
→ The first two problems to solve

- 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

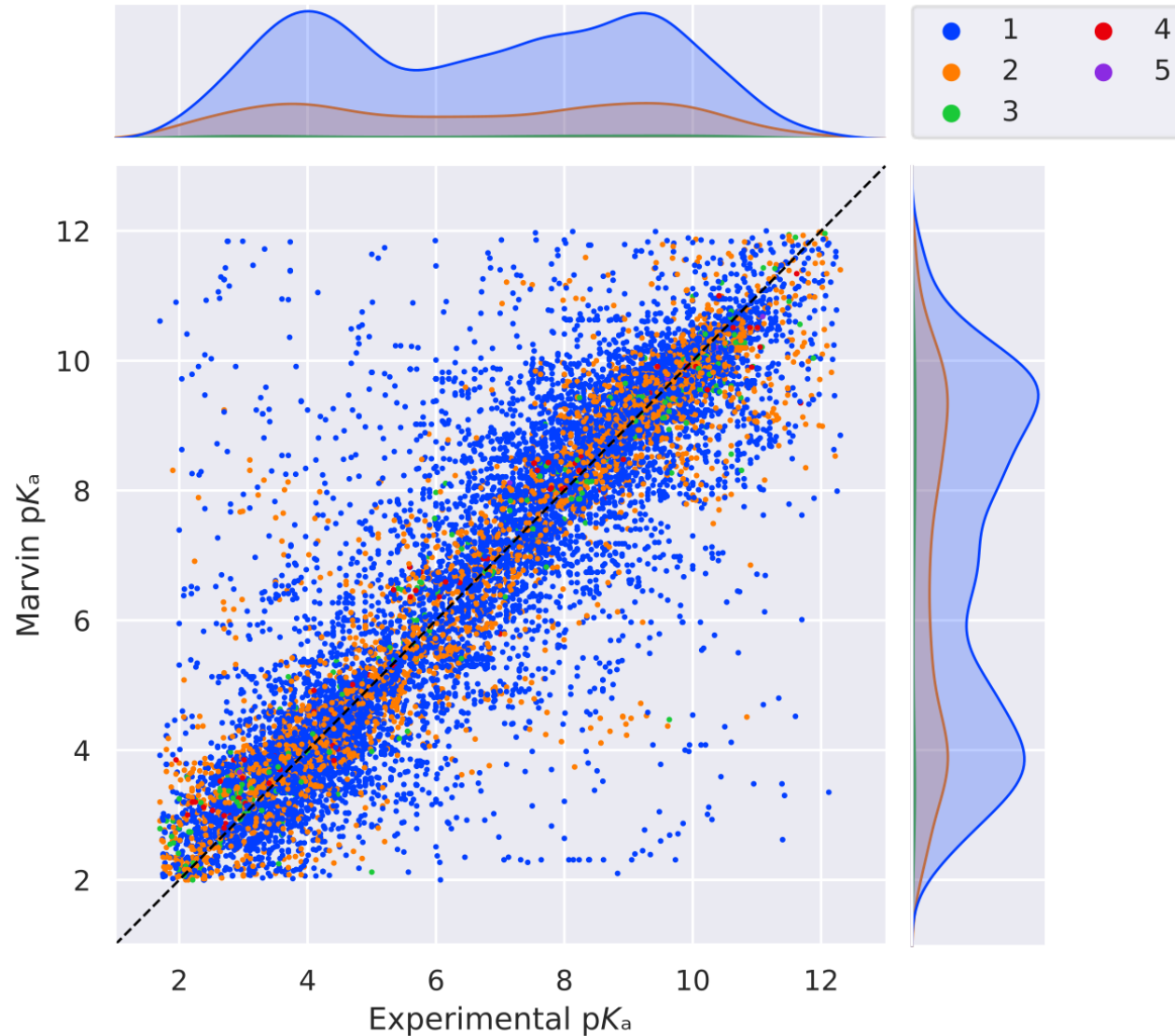


Assign Values to Groups

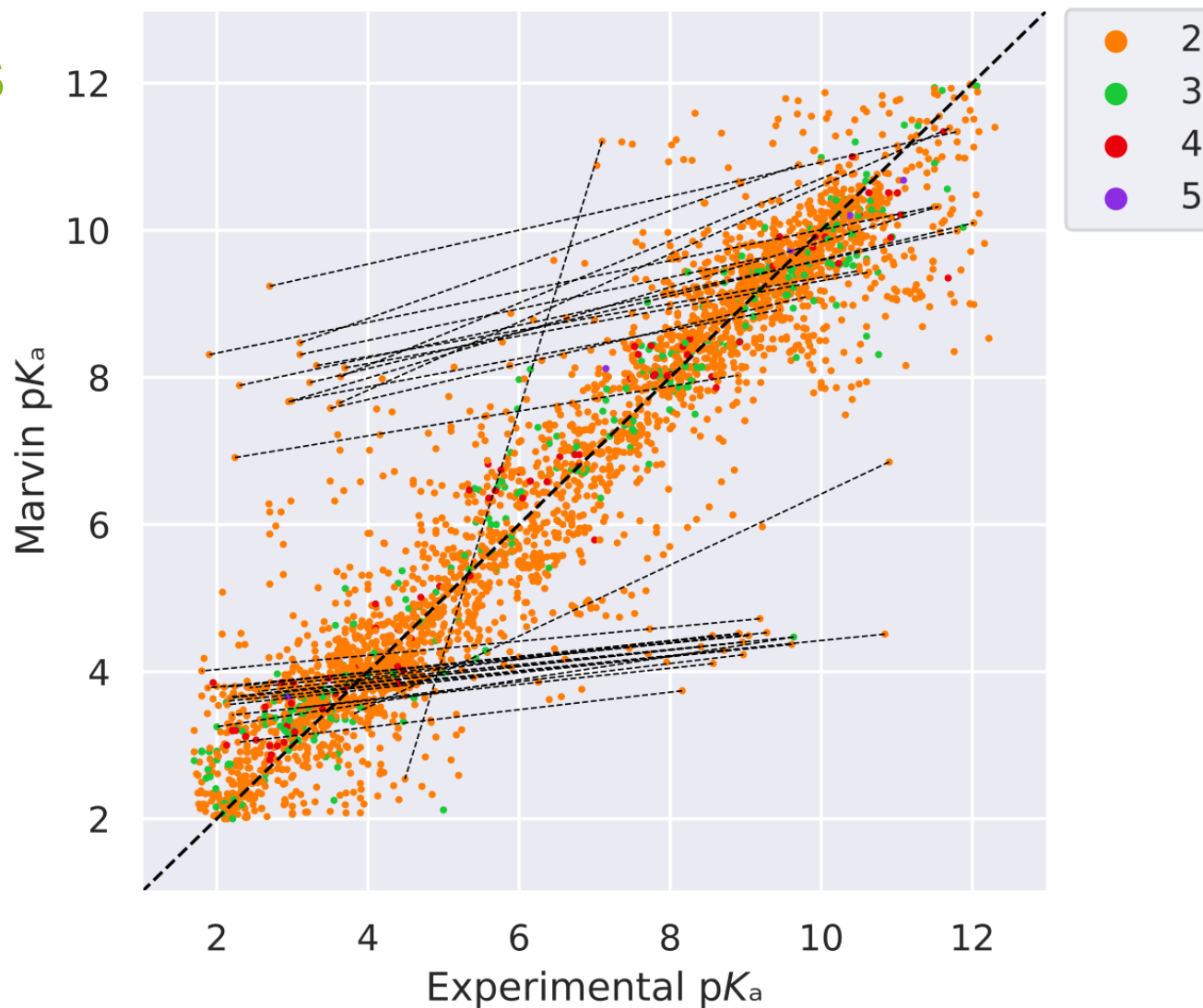
- Combine values that apparently belong to the same titratable group
 - Error range of 0.3 pK_a units
- Find the experimental value that comes closest to the corresponding *Marvin* prediction
- Only consider “exact matches” for now



Results

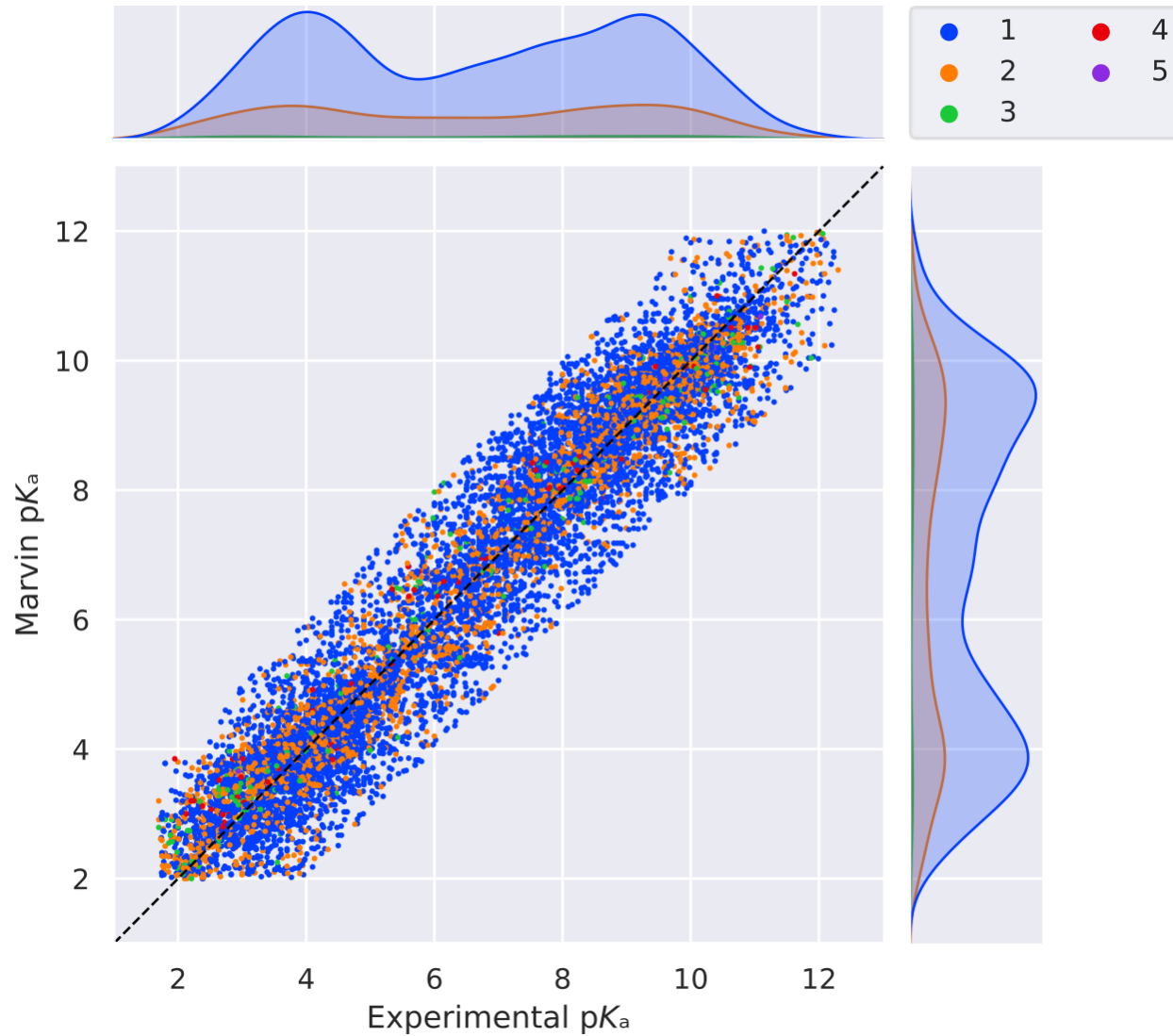
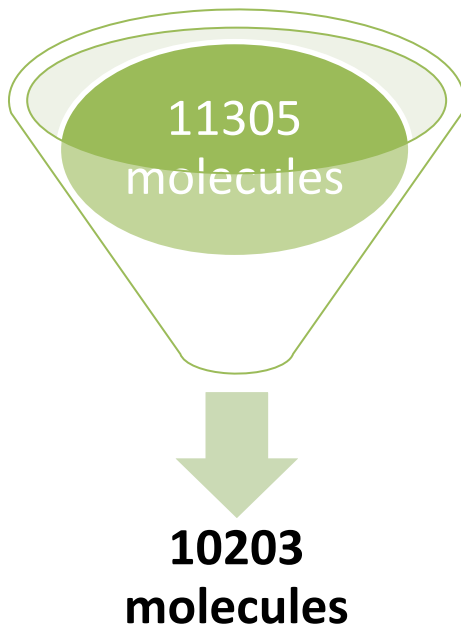


Results





Results

- Cut at max. error = 2





Outlook

 Further investigation of the results and testing with other prediction tools

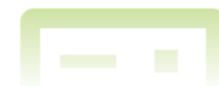
 Improvement of SMARTS pattern extraction

 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

Novartis Pharma AG

Richard A. Lewis
Stephane Rodde

California State University, USA

Enrico Tapavicza

Roche Pharma AG

Christian Kramer

Bayer AG

Michael E. Beck