The absolute importance of applicability domain in QSAR:

A new *in silico* multiprotic pK_a prediction tool with significantly improved prediction accuracy and new functionality for PhysChem, MedChem, CompChem, and Cheminformatics applications

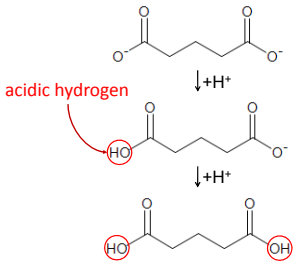
Robert Fraczkiewicz#, Mario Lobell*, Robert D. Clark#, Alexander Hillisch*, Andreas H. Göller*, Ursula Krenz*, Rolf Schoenneis*

Simulations Plus, * Bayer Pharma

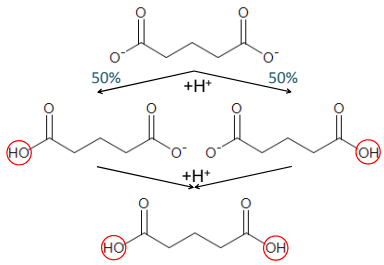
This poster is not an independent entity – it serves as a companion to our podium lecture with the same title presented on Monday morning.

MULTIPROTIC IONIZATION

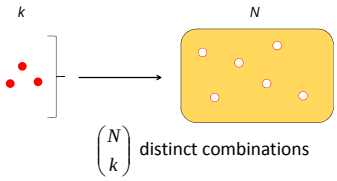
This is how the multiprotic ionization of organic molecules is usually presented:



This is how it really happens:

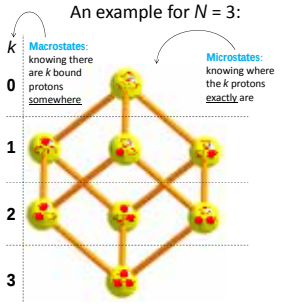


It's a simple combinatorial problem:
Distribute k protons among N sites; $0 \le k \le N$

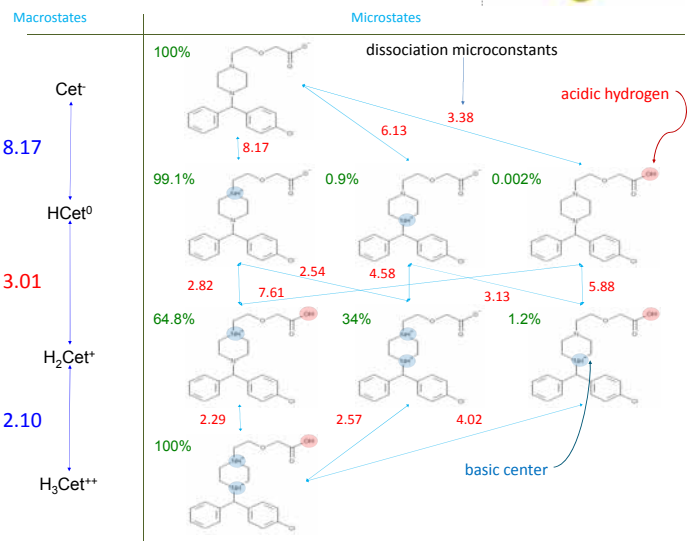


$\binom{N}{k}$ distinct combinations

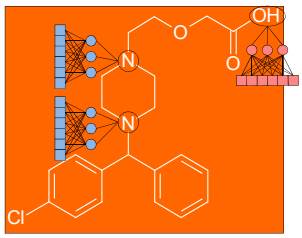
An example for $N = 3$:



Another example for $N = 3$ (Cetirizine) [1]:

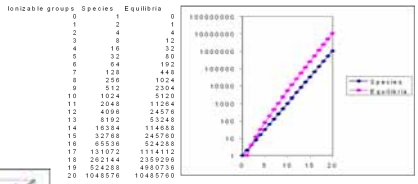


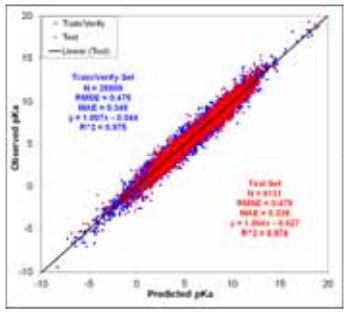
THE S+pKa MODEL



- 10 Artificial Neural Network Ensembles (ANNE)
- ANNEs use localized atomic descriptors as inputs
- ANNEs predict ionization microconstants
- Macroconstants calculated with microequilibria theory
- One ANNE for each of the following 10 classes of ionizable atoms: (1) Hydroxyacids, (2) Acidic amides, (3) Acids of aromatic NH, (4) Thioacids, (5) Carboacids, (6) Amines, (7) Bases of aromatic N, (8) N-oxides, (9) Thiones, (10) Carbobases (protonable C in certain π–excessive rings)

The most significant challenge: the number of microstates grows exponentially with the number of ionizable groups in a molecule. This puts a sizable strain on the CPU and demands a very efficient computer code.





The model has been trained with 27123 molecules (33640 apparent pK_a values) from public + Bayer sources. 20469 molecules (25509 pK_a values) were used in the actual ANNE training, while 6654 molecules (8131 pK_a values) were used as an external test set.

EXTERNAL VALIDATION AT BAYER

Bayer pK_a set	No of cmpds	Average closest Tanimoto similarity to BTr	Tanimoto sim to BTr ≥0.80	No of pK_a	MAE		RMSE		R^2	
					v 6.1	v 7.0	v 6.1	v 7.0	v 6.1	v 7.0
Bayer training set (BTr)	15983	1	100%	19467	0.85	0.29*	1.14	0.40*	0.84	0.98*
Tanimoto similars to BTr	4730	0.88	98%	5644	0.82	0.41	1.03	0.58	0.85	0.95
Strongest acid or base	8931	0.82	60%	9168	0.79	0.52	1.04	0.71	0.76	0.89
Newest measurements	12951	0.79	45%	16404	0.72	0.50	0.94	0.67	0.87	0.93

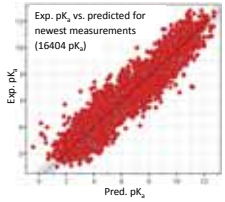
Tanimoto similarity based on MACCS 166 Fingerprints
MAE=Mean Absolute Error , RMSE=Root Mean Squared Error , R^2=Squared correlation coefficient
* BTr has been used for training of the new pK_a calculation tool and is therefore not a true test set in this instance

- Bayer training set (BTr): Bayer compounds and pK_a appended to the training set of the new pK_a calculation tool (ADMET Predictor™ v 7.0)
- Tanimoto similarity to BTr: Average Tanimoto similarity to closest BTr compound is 0.88. 98% have Tanimoto similarity ≥0.8
- Strongest acid or base: All exp. pK_a in this set represent the strongest acid (provided $pK_a \le 9$) or base (provided $pK_a \ge 5$) in a molecule (all structures have been visually inspected and annotated)
- Newest measurements: pK_a measurements performed in time period after sourcing of exp. pK_a for BTr

Significant leap forward in pK_a prediction accuracy with a mean absolute error of only 0.50 for the newest, most challenging set of test compounds

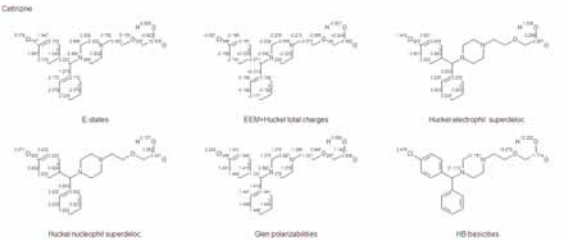
- ACD/Percepta v. 12 and ADMET™ Predictor v 6.1 show comparable pK_a prediction accuracy
- ADMET Predictor™ v 7.0 (after retraining with BTr) shows significantly improved pK_a prediction accuracy

Predicted by	Trained with	MAE	RMSE	R^2
ACD/Percepta v 12	15932 lit pK_a	0.77	1.05	0.84
ADMET Predictor v 6.1	14147 lit pK_a	0.73	0.95	0.86
ADMET Predictor v 7.0	14149 lit pK_a + 19467 Bayer pK_a	0.51	0.67	0.93



- A subset of the Bayer pK_a set "Newest measurements" (see above) comprising 1000 compounds with 1000 pK_a values had been processed with the pK_a calculation tool from ACD/Labs (Advanced Chemical Development) version 12
- Both versions of ADMET Predictor could process all 1000 compound structures
- 19 compound structures could not be processed by the ACD software since they contained certain functional groups (e.g.) sulfoximine not recognized by the software
- Prediction statistics have been calculated for the 981 compounds which could be processed by all tools

SOME EXAMPLES OF ATOMIC DESCRIPTORS



REFERENCES

1. Marosi, A.; Kovacs, Z.; Beni, S.; Kokosi, J.; Noszal, B. Eur. J. Pharm. Sci. 2009, 37, 321-328.