Best of Both Worlds: An Expansion of the State of the Art pK_a Model with Data from Three Industrial Partners

Robert Fraczkiewicz[a], Huy Quoc Nguyen[b], Newton Wu[b], Nina Kausch-Busies[c], Sergio Grimbs[c], Kai Sommer[c], Antonius ter Laak[d], Judith Günther[d], Björn Wagner[e], Michael Reutlinger[e].

[a] Simulations Plus, Inc. 42505 10th Street West, Lancaster, CA 93534, USA. [b] Genentech Inc., Discovery Chemistry, 1 DNA Way, South San Francisco, CA 94080, USA.

[c] Bayer AG, Research & Development, Crop Science, 40789 Monheim, Germany. [d] Bayer AG, Research & Development, Pharmaceuticals, Berlin, Germany.

[e] F. Hoffmann-La Roche AG, Roche Pharma Research and Early Development, 4070 Basel, Switzerland.

CONTACT INFORMATION: robert.fraczkiewicz@simulations-plus.com

ABSTRACT

In a unique collaboration between Simulations Plus and several industrial partners, we were able to develop a new version 11.0 of the previously published[1] *in silico* pK_a model, S+pKa, with considerably improved prediction accuracy. The model's training set was vastly expanded by large amounts of experimental data obtained from F. Hoffmann-La Roche AG, Genentech Inc., and the Crop Science division of Bayer AG. The previous v7.0 of S+pKa was trained on data from public sources and the Pharmaceutical division of Bayer AG. The model has shown dramatic improvements in predictive accuracy when externally validated on three new contributor compound sets. Less expected was v11.0's improvement in prediction on new compounds developed at Bayer Pharma after v7.0 was released (2013-2023), even without contributing additional data to v11.0. We illustrate chemical space coverage by chemistries encountered in the five domains, public and industrial, outline model construction, and discuss factors contributing to model's success.

REFERENCES

1. Fraczkiewicz, R.; Lobell, M.; Göller, A. H.; Krenz, U.; Schoenneis, R.; Clark, R. D.; Hillisch, A. *J. Chem. Inf. Model.* **2015**, *55*, 389.
2. Işık, M.; Rustenburg, A. S.; Rizzi, A.; Gunner, M. R.; Mobley, D. L.; Chodera, J. D. *J. Comput.-Aided Mol. Des.* **2021**, *35*, 131.
3. ADMET Predictor(R) *v 11.0*; Simulations Plus, Inc.: Lancaster, CA, USA, **2023**.
4. 4,5-Diamino-6-hydroxypyrimidine. *PubChem, National Library of Medicine* [Online Early Access]. Published Online: 2023. https://pubchem.ncbi.nlm.nih.gov/compound/135436550.

METHODS

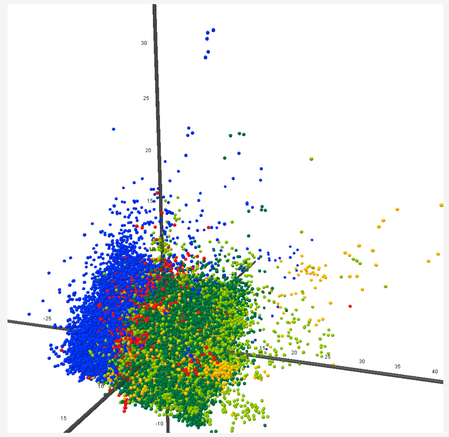
Model algorithm = thermodynamics informed neural network ensemble.[1] All ionization microstates are included.

Data available for model building and internal testing = 50,514 compounds with 70,669 pK_a values.

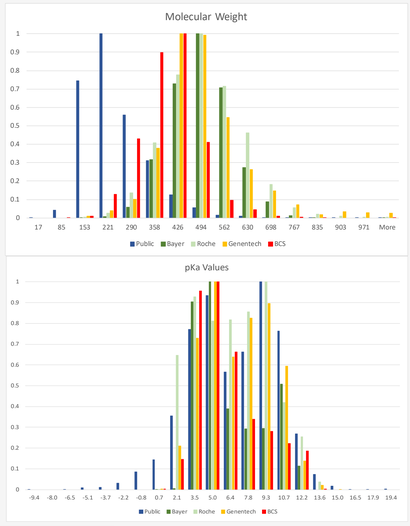
1. Public domain = 10,861 compounds with 13,579 pK_a values. Additional 191 compounds were used in training the carbon protonation submodel (Carbobases).
2. Bayer Pharma = 16,300 compounds with 19,842 pK_a values.
3. Roche = 17,172 compounds with 28,731 pK_a values.
4. Genentech = 2,173 compounds with 4,045 pK_a values.
5. Bayer CropScience = 4,008 compounds with 4,372 pK_a values.

RESULTS

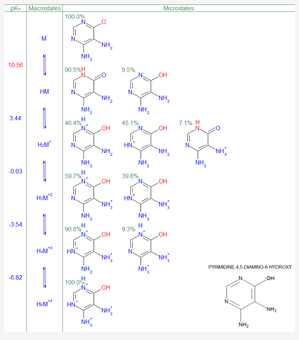
Projection of the chemical space defined by the first three Principal Components of a data matrix whose column are ~150 molecular descriptors calculated by ADMET Predictor. Coloring of data points: blue = Public, dark green = Bayer Pharma, light green = Roche, yellow = Genentech, red = Bayer CropScience. The graph was created by the Miner3D module of ADMET Predictor.[3]



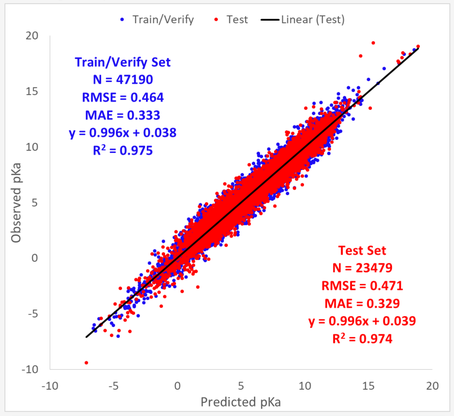
Relative frequency histograms of molecular weight in daltons and observed pK_a values for Public (blue), Bayer Pharma (dark green), Roche (light green), Genentech (yellow), and Bayer CropScience (red) data



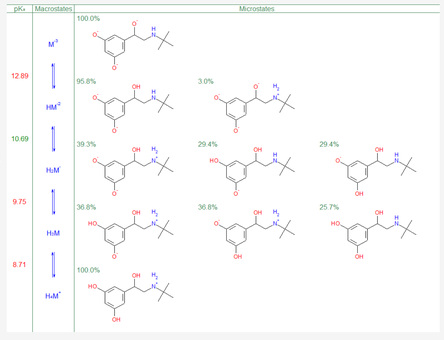
Apparent pK_a (left column) and ionization microstates (right column) predicted with v11.0 for 4,5-diamino-6-hydroxy-pyrimidine. Even though the input tautomer exactly corresponded to the compound's name, the S+pKa model recognized an ortho-pyridone tautomer as the one dominating[4] the neutral macrostate.



Performance graph for the final version 11.0 of the S+pKa model not including Carbobases. Only the subsets labeled Train/Verify was used to build the model's Artificial Neural Network Ensembles (blue points). The remaining subset labeled Test (red points) was set aside for the purpose of selecting the best combination of 9 submodels out of hundreds of prototypes. Predictive statistics: MAE = mean absolute error, RMSE = root-mean-square error, and R^2 = determination coefficient. Linear equations, y = ax + b, illustrate best fit lines to the respective subsets of points, although only Test line is shown.

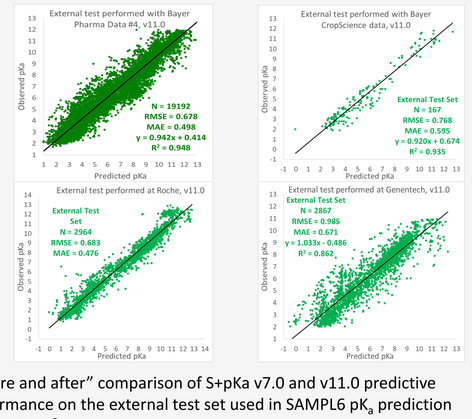


Apparent pK_a (left column), macrostate transitions (middle column), and ionization microstates (right column), predicted with v11.0 for Terbutaline. Green percentage numbers are relative contributions (abundances) of microstates to their respective macrostates. Six microstates with less than 1% contributions were omitted for display clarity.

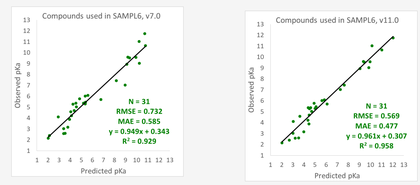


RESULTS

Results of external testing at partner sites.



"Before and after" comparison of S+pKa v7.0 and v11.0 predictive performance on the external test set used in SAMPL6 pK_a prediction competition.[2]



Summary of performance improvements between v7.0 and v11.0 of the S+pKa model. RMSE = root mean square error, MAE = mean absolute error. Training/Verification and Internal Test subsets have been described in the Data Sets section. External validations were performed at respective companies.

		RMSE		MAE	
Data Set	Number of pK_a	v7.0	v11.0	v7.0	v11.0
Overall Train/Verify	47190	0.714	0.464	0.505	0.333
Overall Test	23479	0.662	0.471	0.465	0.329
Roche Train/Verify	19137	0.802	0.409	0.599	0.29
Roche Test	9594	0.775	0.399	0.576	0.278
Genentech Train/Verify	3105	0.793	0.522	0.601	0.383
Genentech Test	940	0.782	0.519	0.602	0.388
Bayer CropScience Train/Verify	3342	1.152	0.842	0.559	0.423
Bayer CropScience Test	1030	1.099	0.773	0.536	0.395
Roche External Test	2964	0.896	0.683	0.706	0.476
Genentech External Test	2867	1.133	0.985	0.805	0.671
Bayer CropScience External Test	167	1.132	0.768	0.858	0.595
Bayer Pharma External Test 1	5642	0.573	0.586	0.411	0.427
Bayer Pharma External Test 2	9149	0.690	0.687	0.508	0.503
Bayer Pharma External Test 3	16363	0.673	0.624	0.507	0.470
Bayer Pharma External Test 4	19192	0.767	0.678	0.541	0.498
SAMPL6	31	0.732	0.569	0.585	0.477

www.simulations-plus.com