# A SAMPL journey, or “There and Back Again”: From 2 over 5 to 6 and back to square one

Nicolas Tielker,[1] Lukas Eberlein,[1] Stefan Güssregen,[2] Stefan M. Kast[1]

*[1]Fakultät für Chemie und Chemische Biologie, TU Dortmund, Germany*
*[2]Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany*

The SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) series of blind prediction challenges was designed to compare the performance of widely different computational methods for predicting physicochemical properties of drug-like compounds in solution. Over more than a decade our group participated in several SAMPL challenges by employing a statistical-mechanical integral equation approach to solvation thermodynamics [1,2] coupled to a quantum-level model of the solute in the form of the “embedded cluster reference interaction site model” (EC-RISM). [3-7] Such a method allows for efficient calculations of Gibbs energies of solvation of neutral and ionic molecules for a wide range of solvents and solvent mixtures. Optimized with respect to quantitative accuracy, both the electronic relaxation and the excess chemical potential governing the insertion into a solvent are taken into account for predicting the Gibbs energy of the molecule in solution, from which other physicochemical properties can be derived. One part of the latest challenge devoted to small molecule properties, SAMPL6 (2017-2019), focused on acidity constants that are difficult to predict since the compounds often contain functional groups and scaffolds that imply a multitude of tautomeric states (“microstates”). In the second part participants were asked to predict octanol-water partition coefficients, log $P_{\mathrm{ow}}$, for a subset of neutral SAMPL6 compound states.

We developed and optimized a workflow to calculate all the above-mentioned properties of a given compound by determining the Gibbs energies of the most probable conformations for all tautomers of the relevant neutral and, where necessary, ionic forms by EC-RISM calculations for SAMPL6, yielding a root mean square deviation of about 1.0 p$K$ units for acidity constants [6] and 0.5 log units for the subset of molecules for which experimental log $P_{\mathrm{ow}}$ were measured. [7] As we previously participated in the SAMPL2 challenge (2009) [4] for tautomer predictions and the SAMPL5 challenge (2015) [5] for cyclohexane-water distribution coefficients using workflows and methodology available back then, it is timely to revisit these earlier datasets by employing the most advanced EC-RISM-based SAMPL6 technology. The retrospective analysis reveals a number of insights and surprises in terms of expected systematic progress and the impact of error compensation on predictive capabilities.

[1] E. L. Ratkova, D. S. Palmer, M. V. Fedorov, *Chem. Rev.* **2015**, *115*, 6312-6356.
[2] J. Heil, S. M. Kast, *J. Chem. Phys.* **2015**, *142*, 114107.
[3] T. Kloss, J. Heil, S. M. Kast, *J. Phys. Chem. B* **2008**, *112*, 4337-4343.
[4] S. M. Kast, J. Heil, S. Güssregen, K. F. Schmidt, *J. Comput.-Aided Mol. Des.* **2010**, *24,* 343-353.
[5] N. Tielker, D. Tomazic, J. Heil, T. Kloss, S. Ehrhart, S. Güssregen, K. F. Schmidt, S. M. Kast, *J. Comput.-Aided Molec. Des.* **2016**, *30*, 1035-1044.
[6] N. Tielker, L. Eberlein, S. Güssregen, S. M. Kast, *J. Comput.-Aided Mol. Des.* **2018**, *32,* 1151-1163.
[7] N. Tielker, D. Tomazic, L. Eberlein, S. Güssregen, S. M. Kast, *J. Comput.-Aided Mol. Des.* **2020**, https://doi.org/10.1007/s10822-020-00283-4.