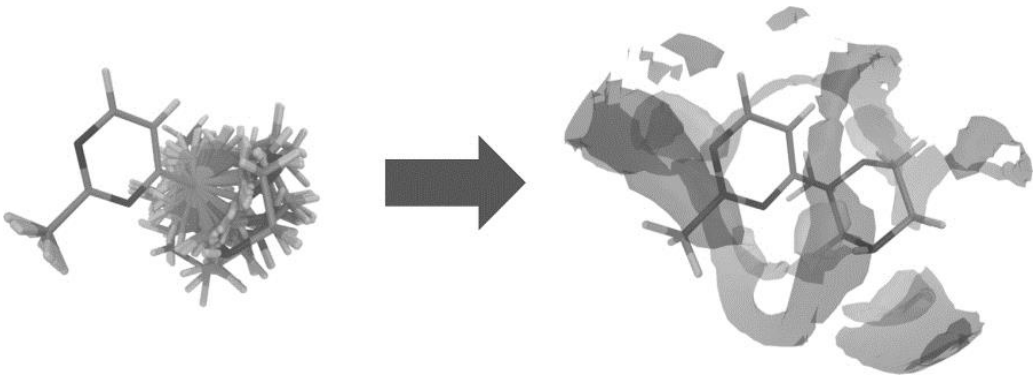
The SAMPL5 challenge for embedded-cluster integral equation theory: solvation free energies, aqueous pK_a and cyclohexane-water log *D*

Nicolas Tielker,[1] Daniel Tomazic,[1] Jochen Heil,[1] Thomas Kloss,[2] Sebastian Ehrhart,[3] Stefan Güssregen,[4] K. Friedemann Schmidt,[4] Stefan M. Kast[1]

[1]*Fakultät für Chemie und Chemische Biologie, TU Dortmund, Germany*
[2]*IPhT, L'Orme des Merisiers, CEA-Saclay, Gif-sur-Yvette, France*
[3]*CERN, Geneva, Switzerland*
[4]*Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany*

Reliable yet fast prediction of physicochemical properties of drug-like compounds requires proper theories, as for instance provided by the integral equation approach to fluid phase thermodynamics [1]. Such a method allows for efficient calculations of free energies of solvation or partition coefficients between immiscible or partly miscible phases like water and cyclohexane. To accurately model the solvation of small molecules we here combine such a statistical-mechanical description of the solvent with a quantum-level description of the solute in the form of the "embedded cluster reference interaction site model" (EC-RISM). This combination, optimized with respect to quantitative accuracy, takes both the electronic relaxation and the excess chemical potential governing the insertion into a solvent into account for predicting the free energy of solvation [2]. It is therefore possible to address challenging problems related to drug discovery.



The partition coefficients of drug-like molecules are difficult to predict since these species often contain functional groups and scaffolds that can result in more than one tautomeric and ionization state at physiological pH. Furthermore, the molecules' conformational ensembles must be considered for each solvent. To capture the state of such compounds in solution we exhaustively sample their conformational, tautomeric, and ionization states followed by calculating their free energies of solvation. This allows for the accurate prediction of both, partition coefficients (log *P*) and the pH-dependent distribution coefficients for ionizable compounds (log *D*), as demonstrated for the dataset of the SAMPL5 blind prediction challenge [3]. The compounds cover a wide range of chemistries, therefore offering a realistic testbed for a methodology that has not been parametrized using the target properties.

[1] E. L. Ratkova, D. S. Palmer, M. V. Fedorov, *Chem. Rev.*, **2015**, *115*, 6312-6356.
[2] T. Kloss, J. Heil, S. M. Kast, *J. Phys. Chem. B*, **2008**, *112*, 4337-4343.
[3] N. Tielker, D. Tomazic, J. Heil, T. Kloss, S. Ehrhart, S. Güssregen, K. F. Schmidt, S. M. Kast, *J. Comput.-Aided Mol. Des.*, **2016**, *30*, 1035-1044.