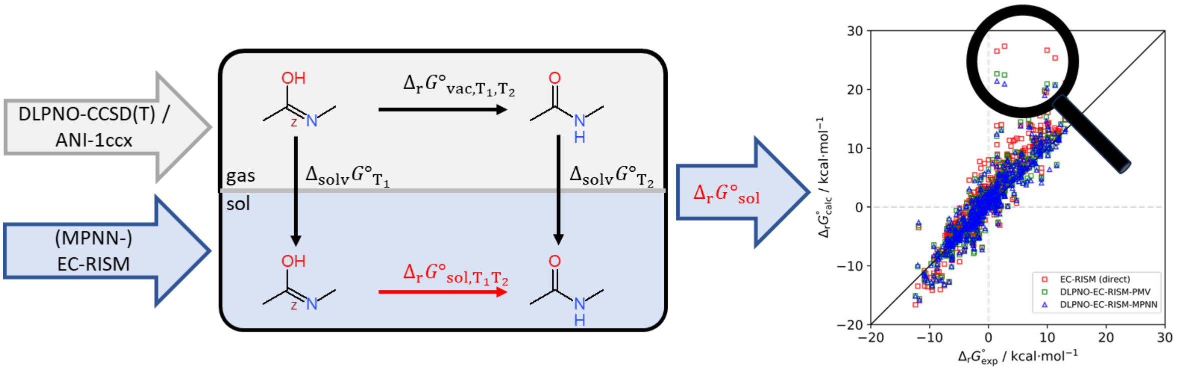
Tautomerization prediction as a testbed for theory-based experimental uncertainty analysis

Michael Strobl, Nicolas Tielker, Christian Chodun, Lukas Eberlein, Yannic Alber,
Julia Jasper-Eberlein, Stefan M. Kast

*Fakultät für Chemie und Chemische Biologie, Technische Universität Dortmund,
44227 Dortmund, Germany*

Aqueous tautomer equilibria are highly relevant to a wide range of chemical and biological issues, e.g. DNA base-pairing or biological activity of drugs. [1,2] Despite the importance of these processes, experimental data is scarce and potentially unreliable due to the difficulties associated with the measurements. Complementary theoretical approaches aim at reliability and predictive power, but developing such models on the basis of unreliable datasets is difficult.

Here, different approaches to calculate free energies of tautomerization are applied to several tautomer equilibria datasets available, the SAMPL2 blind prediction challenge compounds, [3-5] the Tautomer Database [6] and the Tautobase. [1] The first ansatz is a direct approach derived solely from solution-phase properties provided by the Embedded Cluster Reference Interaction Site Model (EC-RISM). [7] In the second approach, an indirect thermodynamic route is chosen, which is complemented by gas phase free energies derived from either state-of-the-art DLPNO-CCSD(T) [8] calculations or quantum-based machine learning models like ANI-1ccx. [9] In this context, application of a machine learning-based optimization of EC-RISM is also investigated.



Subsequently, our best performing model combinations are compared with the literature, which allows us to construct a consensus dataset and perform a statistical analysis with particular emphasis on curating the reference datasets. Data points where the various theoretical approaches agree on a certain value range, but show a large deviation from the experimental reference, are then analyzed in more detail. Hence, we are able to identify suspicious database entries that may be based on problematic measurements or incorrect annotations. As a key result, an ordered set of tautomer pairs with increasing experimental uncertainty is produced, measured by increasing consensus prediction error. This curated dataset will more faithfully allow for training and evaluating novel computational methods.

[1] O. Wahl, T. Sander, *J. Chem. Inf. Model.*, **2020**, *60*, 1085-1089.
[2] Y. C. Martin, *Drug Discov. Today Technol.*, **2018**, *27*, 59-64.
[3] M. T. Geballe, …, P. J. Taylor, *J. Comput. Aided Mol. Des.*, **2010**, *24*, 259-279.
[4] S. M. Kast, …, K. F. Schmidt, *J. Comput. Aided Mol. Des.*, **2010**, *24*, 343-353.
[5] N. Tielker, L. Eberlein, …, S. M. Kast, *J. Comput. Aided Mol. Des.*, **2021**, *35*, 453-472.
[6] D. K. Dhaked, L. Guasch, M. C. Nicklaus, *J. Chem. Inf. Model,.* **2020**, *60*, 1090-1100.
[7] T. Kloss, J. Heil, S. M. Kast, *J. Phys. Chem. B*, **2008**, *112*, 4343-4337.
[8] F. Pavošević, …, F. Neese, E. F. Valeev, *J. Chem. Phys.*, **2017**, *146*, 174108.
[9] J. S. Smith, …, A. E. Roitberg, *Nat. Commun.*, **2019**, *10*, 2903.