COSMOlogic
Predicting Solutions

COSMO-RS based predictions for the SAMPL6 logP challenge

Christoph Loschen, Jens Reinisch, Andreas Klamt
COSMO*logic* GmbH & Co. KG – Dassault Systèmes

loschen@cosmologic.de

SAMPL6 logP blind challenge

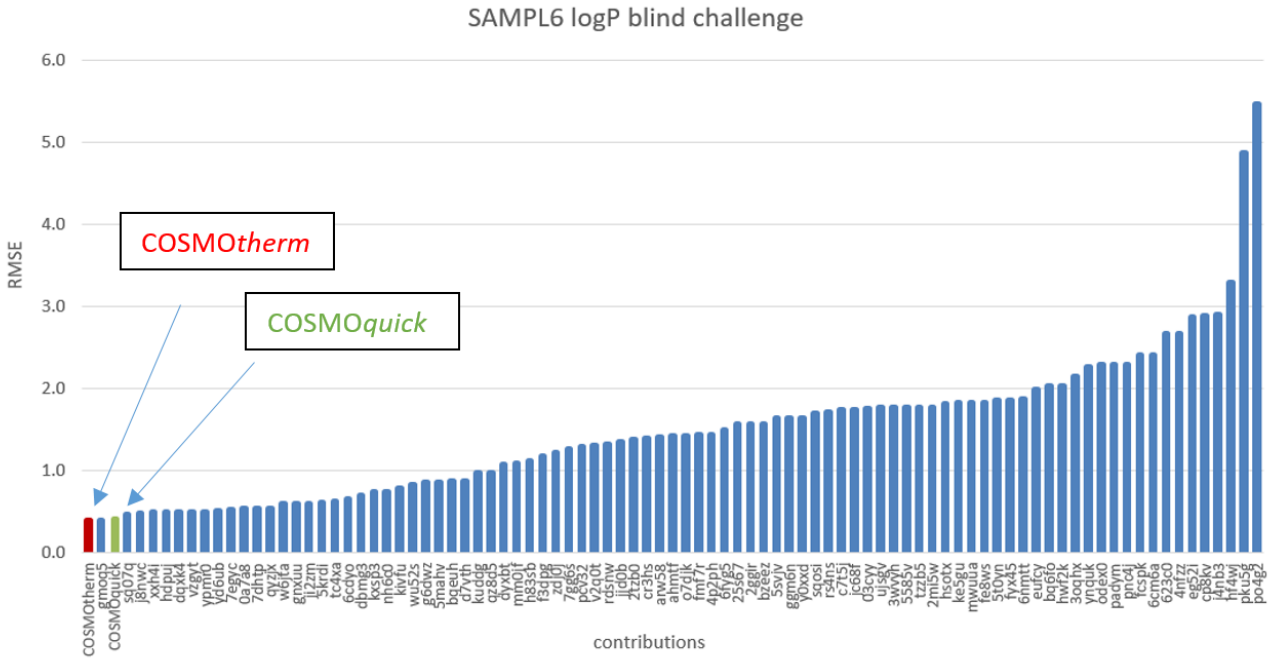
COSMOlogic
Predicting Solutions

Methods used

- QSPR/machine learning
- Deep Learning
- Molecular Dynamics
- Implicit solvation models (SMD)
- 3D-RISM
- COSMO-RS

Results

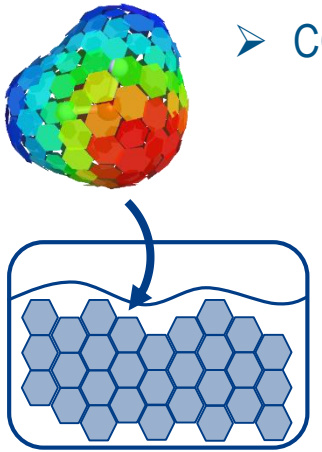
COSMO-RS based methods



COSMO-RS

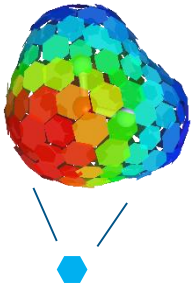
COSMOlogic
Predicting Solutions

In a nutshell



- COSMO: implicit solvation model via DFT
- COSMO-RS: statistical thermodynamics
  - Electrostatic: $E \sim (\sigma + \sigma')^2$
  - Hydrogen bonds: $E \sim (\sigma \times \sigma')$
  - Van der Waals: $E \sim$ area
  - Combinatorial term: E (Shape)

σ-surface



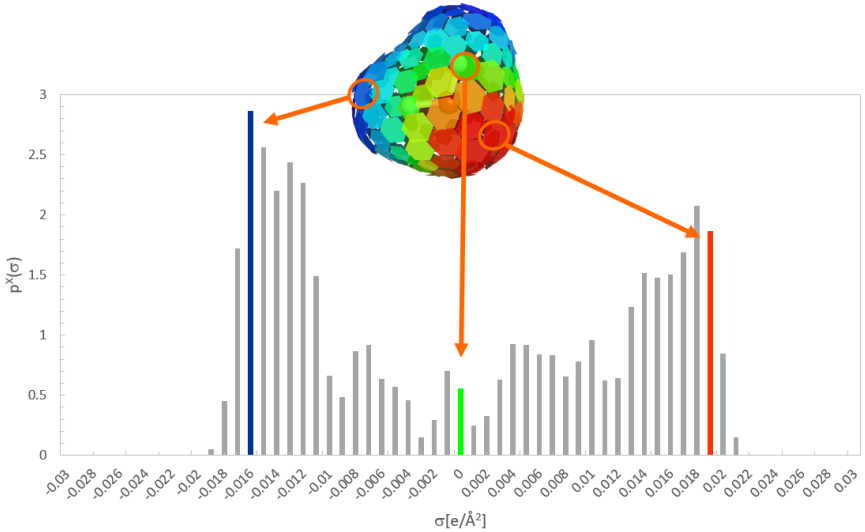
Intermolecular interactions are based on σ surface segments

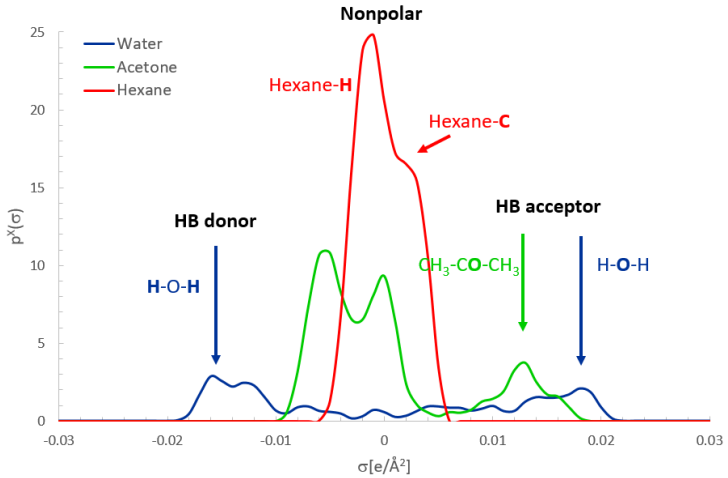
Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224-2235.
Klamt, A. *WIREs: Comput. Mol. Sci.* **2011**, *1*, 699-70

COSMO-RS

COSMOlogic
Predicting Solutions

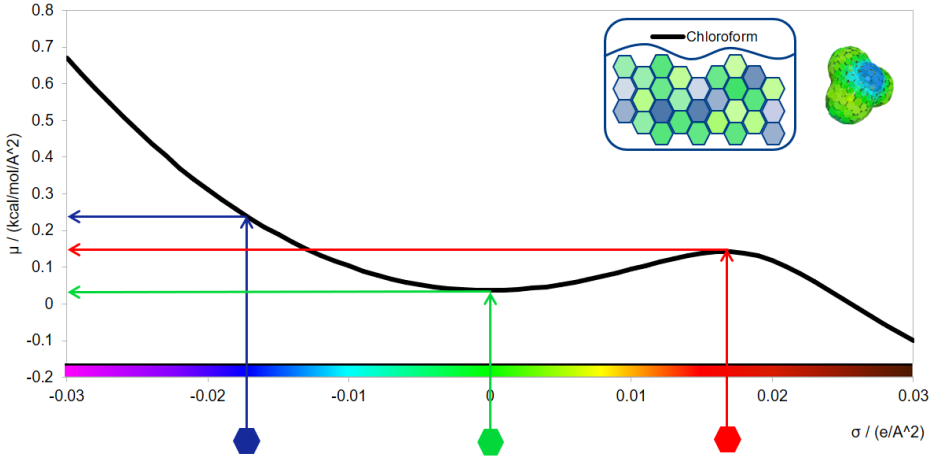
σ-profile $p(\sigma)$: a histogram of charged surface segments of a molecule





COSMO-RS

The σ-potential $\mu_s(\sigma)$ is a characteristic function of a system at a given T.



From $\mu_s(\sigma)$ one obtains the chemical potential of a substance in solution μ_s and **all** related properties:

$$\mu_S = \int d\sigma\, p\,(\sigma)\mu_S(\sigma) + RT \ln\{x\gamma_{comb.,S}\}$$

* F. Eckert and A. Klamt, *AIChE Journal*, 48 (**2002**) 369-385; A. Klamt and F. Eckert, *Fluid Phase Equilibria*, 172 (**2000**) 43-72; A. Klamt, V. Jonas, T. Buerger and J. C. W. Lohrenz, *J. Phys. Chem. A*, 102 (**1998**) 5074; A. Klamt, J. Phys. Chem., 99 (**1995**) 2224.

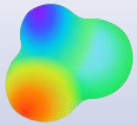
COSMO*therm* workflow

COSMOlogic
Predicting Solutions

Conformational sampling in liquid phase: COSMO*conf* v4.3

- Structure generation (Balloon & RDKit)
- Iterative conformer reduction according to energy and clustering of structures and (liquid) chemical potentials
- COSMO-Levels used: BP/SV(P) , BP/TZVP & BP/TZVPD (Turbomole v7.3)[3]
- Identification of optimal conformational set

Liquid Phase Statistical Thermodynamics COSMO-RS: COSMO*therm* v19



- COSMO-RS based computation of chemical potentials in water & octanol[4-6]
- Consideration of conformational effects
- Assuming wet octanol (27.4% mf water)
- Parameterization: BP_TZVPD_FINE_19

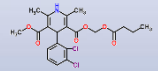
$$\log_{10}(P) = \log_{10}\left(\frac{c_1}{c_2}\right) = \left(\mu_2^x - \mu_1^x\right) / RT\ln(10) + \log_{10}(V_2 / V_1)$$

1. COSMOconf 4.3; COSMOlogic GmbH & Co. KG; http://www.cosmologic.de: Leverkusen, Germany, 2018.
2. Klamt, A.; Eckert, F.; Diedenhofen, J. Phys. Chem. B 2009, 113 (14), 4508–4510.
3. TURBOMOLE V7.3; University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com: Karlsruhe, Germany, 2018.
4. Klamt, A. J. Phys. Chem. 1995, 99 (7), 2224–2235.
5. Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. J. Phys. Chem. A 1998, 102 (26), 5074–5085.
6. COSMOtherm, Release 19; COSMOlogic GmbH & Co. KG; http://www.cosmologic.de: Leverkusen, Germany, 2019.

COSMO*quick* workflow

COSMOlogic
Predicting Solutions

Molecule perception



- Input: SMILES, xyz, SDF
- Bond perception
- Normalization
- Circular fingerprints
- QSPR descriptors

Generation of approximated COSMO files

- Database of about 200,000 COSMO files
- Use fingerprints to find substructures in database
- Construct COSMO surfaces

COSMO*therm* property calculation

- Compute σ-potentials
- Liquid phase properties: chemical potentials, activity coefficients, partition coefficients etc.

Empirical correction term via ML

- Deviations from experiment are fitted via machine learning
- For new compounds correction term is predicted
- Model trained via XGBoost library[3]

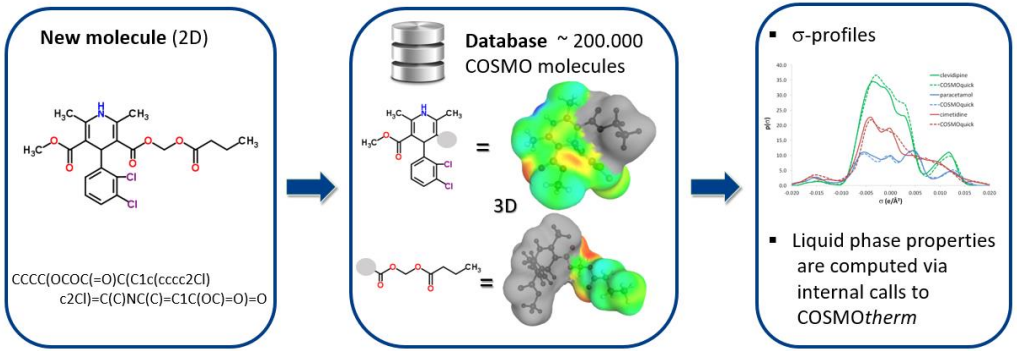
1. COSMOquick 1.7; COSMOlogic GmbH & Co. KG; http://www.cosmologic.de: Leverkusen, Germany, 2018.
2. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM, 2016; pp 785–794.

COSMO*quick:* instant σ-profile composition

COSMOlogic Predicting Solutions

Idea: Compose larger molecules from a database of pre-calculated molecules

Assumption: σ-profiles of compounds are somewhat additive!



cheminformatics backend

σ-profile composition

- SDF / SMILES parser
- bond perception*
- SMILES/Molec. standardization
- Unique identifier
- Circular fingerprints
- Atomic similarity database

Hornig, M. & Klamt, A. *J Chem Inf Model,* **2005**, *45*, 1169-1177

Loschen, C. & Klamt, A. *Ind. Eng. Chem. Res.* **2012**, *51*, 14303-14308.

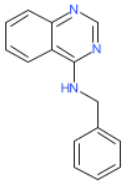
COSMO*quick:* instant σ-profile composition

COSMOlogic
Predicting Solutions

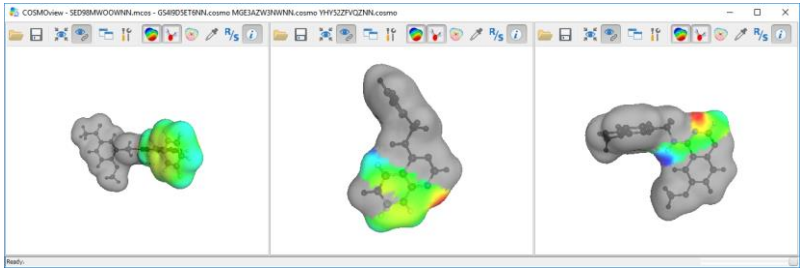
Compound	N,fragments	<similarity>
SM02	5	4.9
SM04	4	5.2
SM07	3	6.2
SM08	4	3.5
SM09	6	4.9
SM11	0.0	9.0
SM12	4	6.3
SM13	2	6.5
SM14	3	4.3
SM15	4	4.0
SM16	2	5.6

1: lowest simlarity (single atom match)
9: highest similarity (full match)

Example: SM07



Atomic weight strings:
w={0000000000000000010000011111110000000000000010011111111000}
w={1111000000000011011000000010}
w={0000000111100000000000000000011000000000}



- significant fragmentation effect expected!
- Only a fraction of a second needed for σ-profile

COSMO*quick*

COSMOlogic
Predicting Solutions

ML correction term:

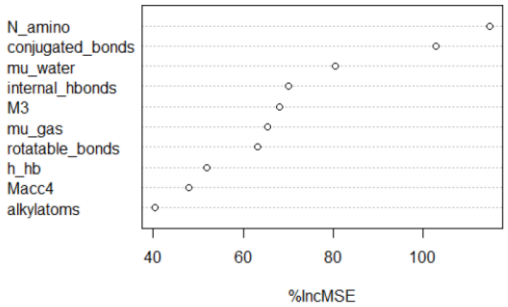
$$\log P = \log P_{TZVP} + \Delta \log P_{ML-corr}$$

COSMOtherm & approx. σ

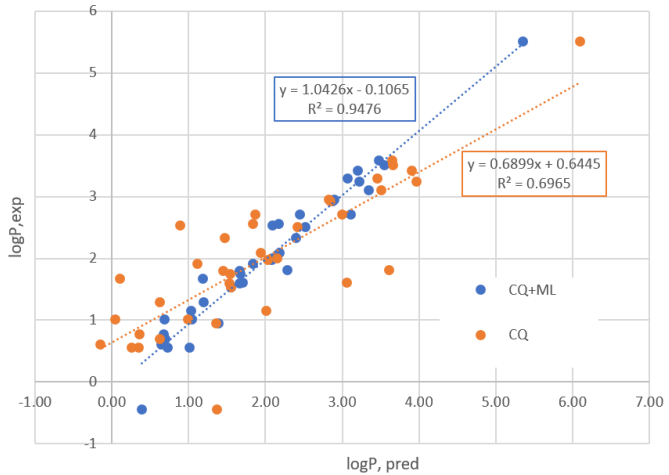
XGBoost

data type	n	source
Training & crossval.	10964	PHYSPROP
Test	37	PHYSPROP subset via substructure search

variable(i.e. descriptor) importance



test set results (RMSE=0.46)

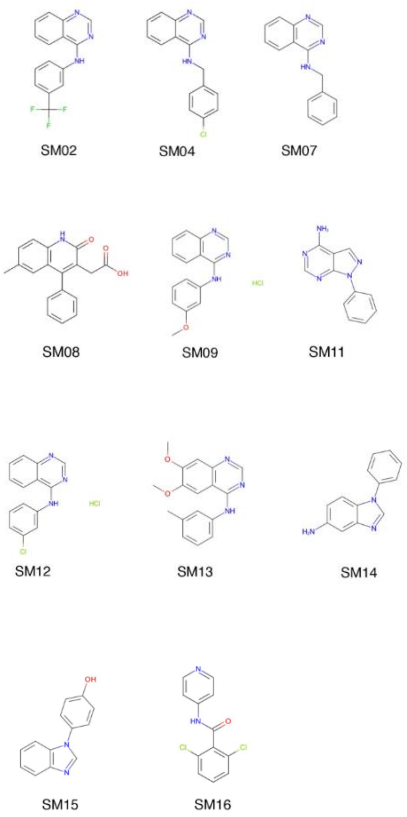


Results

Comparison of COSMO*therm* submissions

compound	logP,exp	COSMOtherm	COSMOquick
SM02	4.09	4.42	3.98
SM04	3.98	3.86	3.74
SM07	3.21	3.48	3.19
SM08	3.1	2.85	2.74
SM09	3.03	3.44	3.38
SM11	2.1	2.00	2.64
SM12	3.83	3.82	4.29
SM13	2.92	**3.84**	3.41
SM14	1.95	2.21	2.24
SM15	3.07	2.77	**2.28**
SM16	2.62	3.05	2.88

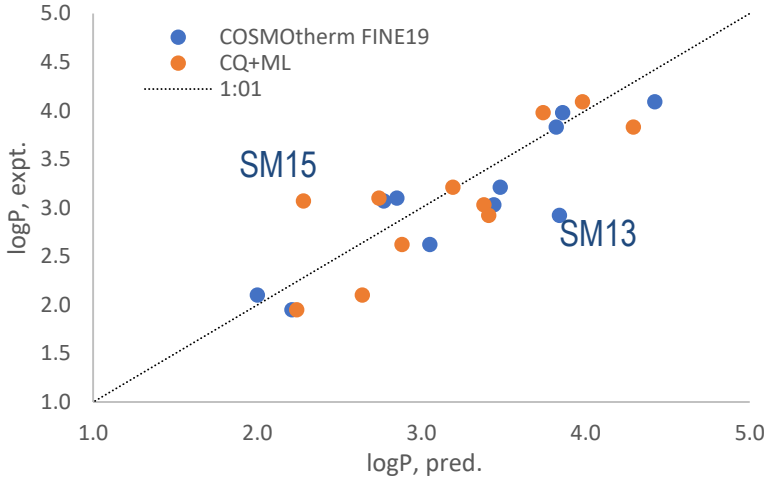
id	σ-surface from	Method	n , fragments	level	RMSE
hmz0n	COSMO files	Turbomole	-	FINE19	0.38
3vqbi	SMILES	COSMOfrag	37	TZVP+ML	0.41
	COSMO files	Turbomole	-	TZVP+ML	**0.35**
0.5 * hmz0n+0.5 *3vqbi	COSMO files / SMILES	consensus	-	FINE19 TZVP+ML	**0.34**



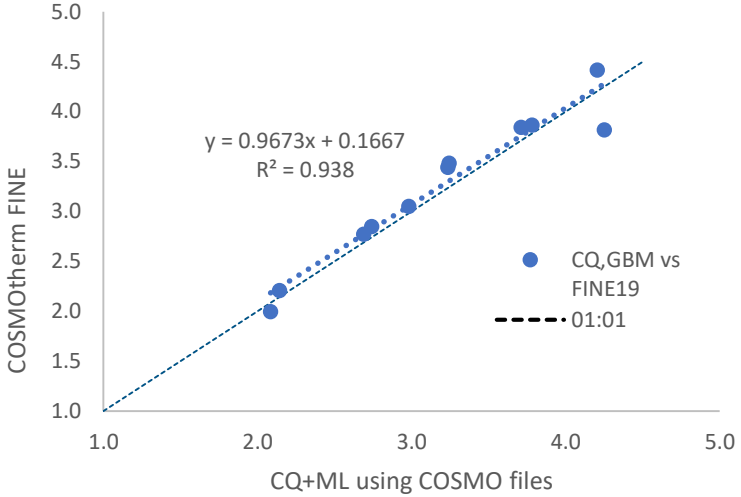
Results

Comparison of COSMO*therm* Submissions

versus experiment



with each other without fragmentation



SAMPL6

Learnings

- Need for systematic tautomer & deprotonation workflow (pKa part)
- TZVP (intermediate level) shows systematic problems with this compounds class
- Probably more potential for ML based approaches
- Monitor fragmentation error (COSMO*quick*)

Many thanks to the SAMPL6 organizers for setting up the challenge!