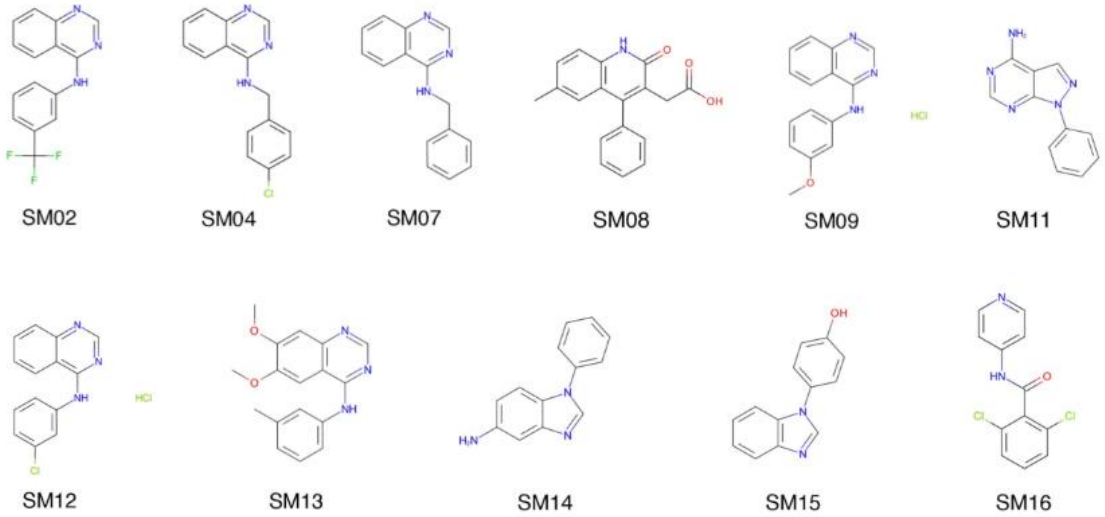
3DEXPERIENCE®

COSMO-RS based predictions for the SAMPL6 logP challenge

Christoph Loschen, Jens Reinisch, Andreas Klamt

SAMPL6

Part II: Blind challenge for octanol-water partition coefficients (logP) of 11 drug-like molecules / fragments



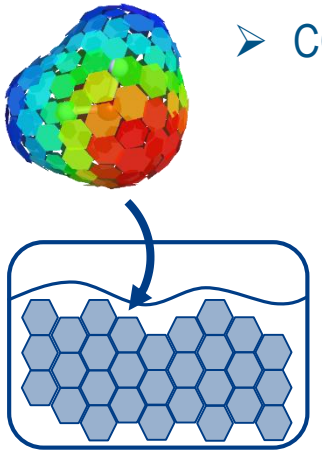
SAMPL6 blind challenge

Methods used:

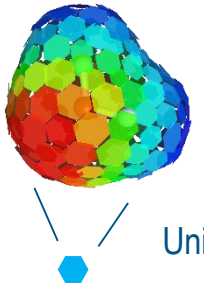
- QSPR & Machine learning
- Molecular Dynamics
- Implicit solvation models (SMD)
- 3D-RISM
- **COSMO-RS**

COSMO-RS

In a nutshell



- COSMO: implicit solvation model via DFT



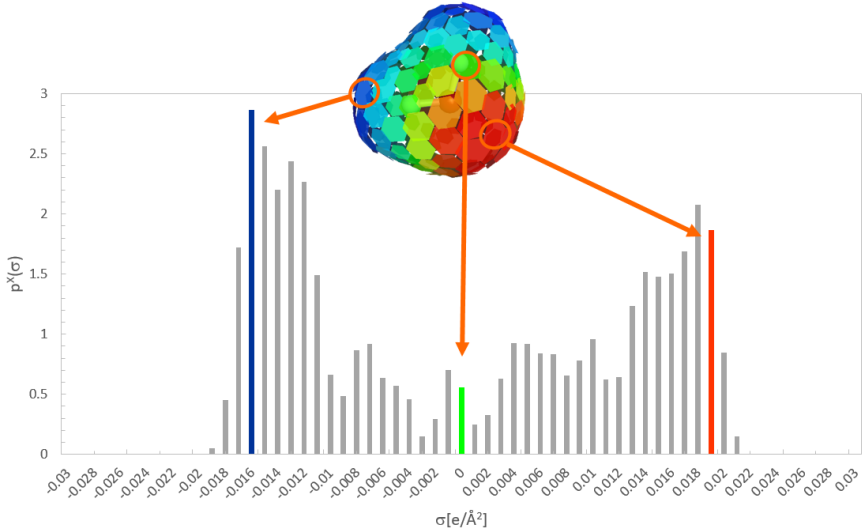
- COSMO-RS: statistical thermodynamics
  - Misfit/Electrostatic: $E_{MF}(\sigma,\sigma') = a_{contact} c_{MF} (\sigma + \sigma')^2$
  - Hydrogen bonds: $E_{HB}(\sigma,\sigma') \cong a_{contact} c_{HB} \min(0, \sigma\sigma' - \sigma_{HB}^2)^2$
  - Van der Waals: $E_{vdW} = a_{contact} (\tau_{vdW} + \tau'_{vdW})^2$
  - Combinatorial/entropy term

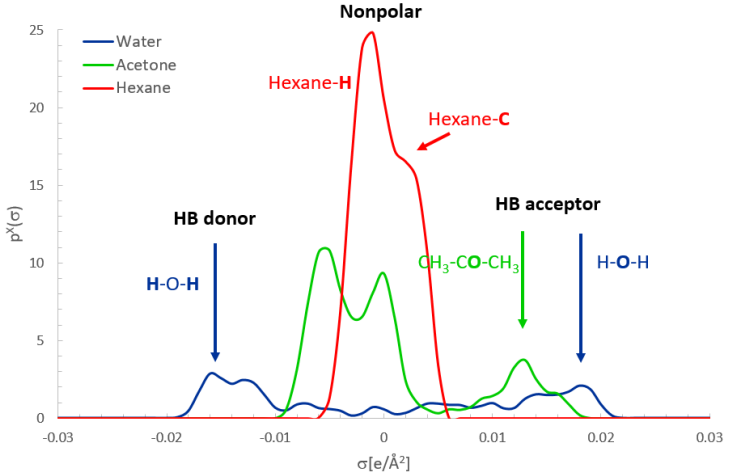
Intermolecular interactions are based on σ surface segments.

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

COSMO-RS

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



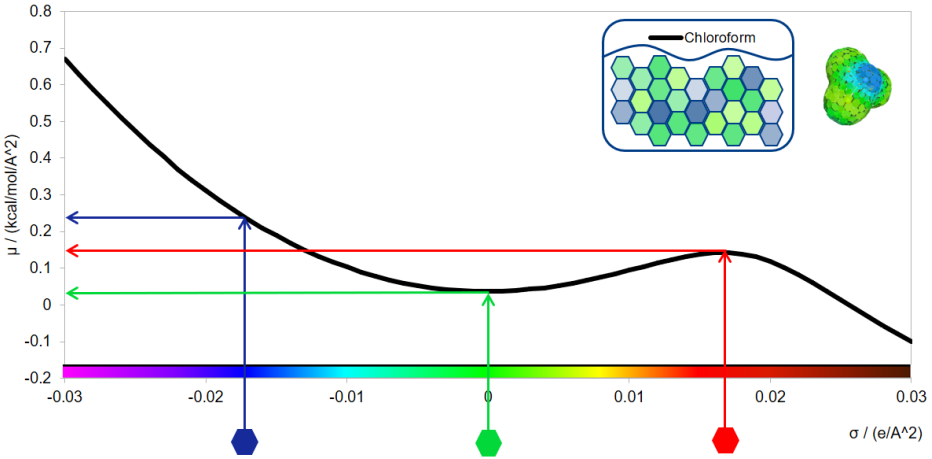


$$E(\sigma,\sigma') = E_{MF}(\sigma,\sigma') + E_{HB}(\sigma,\sigma') + E_{vdW}(\sigma,\sigma')$$

$$\mu_s(\sigma) = -kTln\int p_s(\sigma')exp\left(-\frac{E(\sigma,\sigma') - \mu_s(\sigma')}{kT}\right)d\sigma'$$

COSMO-RS

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



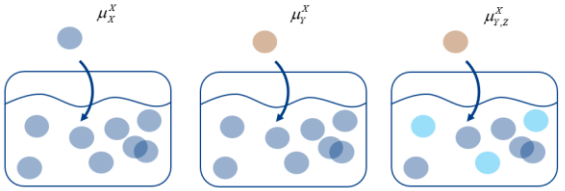
From $\mu_s(\sigma)$ one obtains the chemical potential of a substance in solution μ_s and **all** related properties:

$$\mu_s = \int d\sigma\, p(\sigma)\, \mu_s(\sigma) + RTln(x\gamma_{comb})$$

* 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-RS

Property prediction via equilibrium chemical potentials



$$\mu_{Phase\ A}^{compound\ X} = \mu_{Phase\ B}^{compound\ X}$$

- Activity coefficient: $\ln \gamma_Y^X = \left(\mu_Y^{*X} - \mu_X^X \right) / RT$
- Solubility: $\ln(x_{solv}) = \left[\mu_{cryst.}^X - \mu_{solv.}^{*X} \right] / RT$
- Partition coefficient: $\log P_{OW} = \log_{10} \left[\exp\left\{ \left(\mu_W^{*X} - \mu_O^{*X} \right) / RT \right\} \frac{c_W}{c_O} \right]$

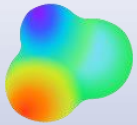
$\mu_{phase,1}^{*x} = \mu_{phase,1}^{x} - RT \ln(x)$: pseudo-chemical potential (Naim, A. B. Solvation Thermodynamics; Plenum Press: New York, NY, 1987.)

COSMO*therm* workflow

Conformational sampling in the *liquid* phase: COSMO*conf* [1,2]

- Initial 3D Structure generation
- 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*



- 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

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.

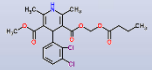
Conformations

COSMO*conf* workflow overview

- 3D structure and initial conformer generation
- Clustering by structure and conformer reduction by relative energies
- BP/TZVP COSMO optimization (TM)
- Clustering by chemical potential and conformer reduction
- BP/TZVPD COSMO single point (TM)
- Relevant conformer selection via iterative chemical potential computation in a diverse solvent set

COSMO*quick* workflow

Molecule perception



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

Generation of approximated σ-surface

- Database of about 200,000 COSMO files
- Use fingerprints to find best sub-structure matches 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

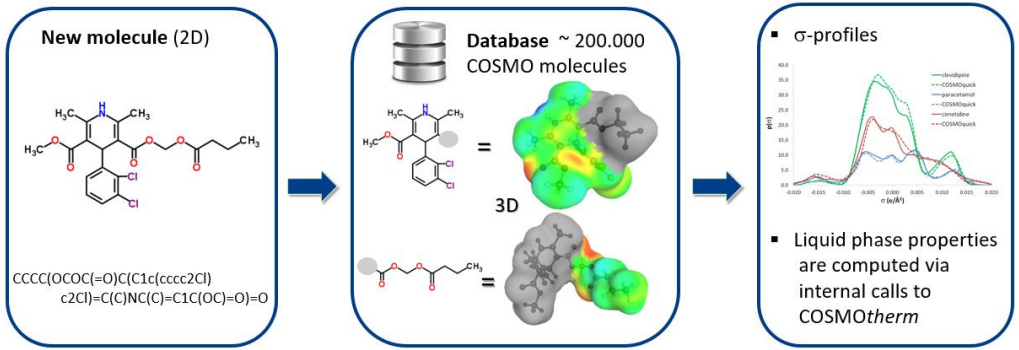
- Deviations from experiment are fitted via machine learning
- For new compounds correction term is predicted
- Model trained via XGBoost library[2]

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.

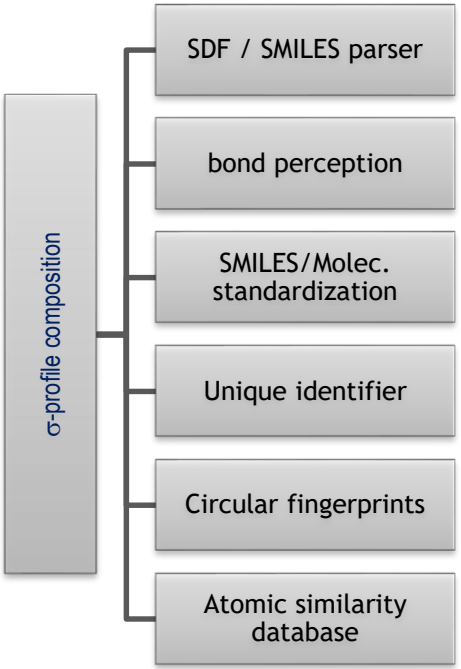
COSMO*quick:* instant σ-profile composition

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

Assumption: σ-profiles of compounds are somewhat additive!



cheminformatics backend



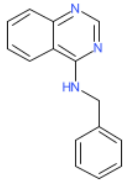
*Hornig, M. & Klamt, A. *J Chem Inf Model,* **2005**, *45*, 1169-1177.

COSMO*quick:* instant σ-profile composition

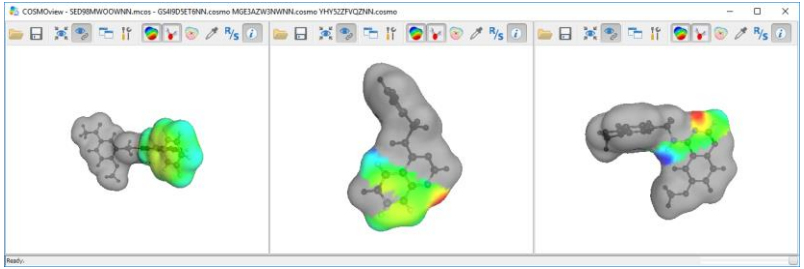
Compound	N,fragments	<quality>
SM02	5	4.9
SM04	4	5.2
SM07	3	6.2
SM08	4	**3.5**
SM09	6	4.9
SM11	0.0	9.0
SM12	4	6.3
SM13	2	6.5
SM14	3	4.3
SM15	4	**4.0**
SM16	2	5.6

1: lowest simlarity (single atom match)
9: highest similarity (full match – 8 shells)

Example: SM07



Atomic weight strings:
w={0000000000000000100000111111100000000000001001111111000}
w={1111000000000011011000000010}
w={0000000111100000000000000000001100000000}



- Compounds are not well represented except SM11
- significant fragmentation effect expected!
- Only a fraction of a second needed for σ-profile

COSMO*quick*

ML correction term:

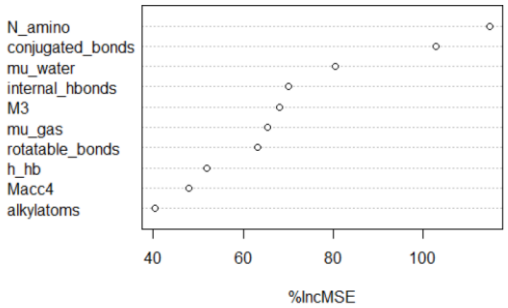
$$\log P = \log P_{TZVP} + \Delta \log P_{ML-corr}$$

COSMOtherm & approx. σ

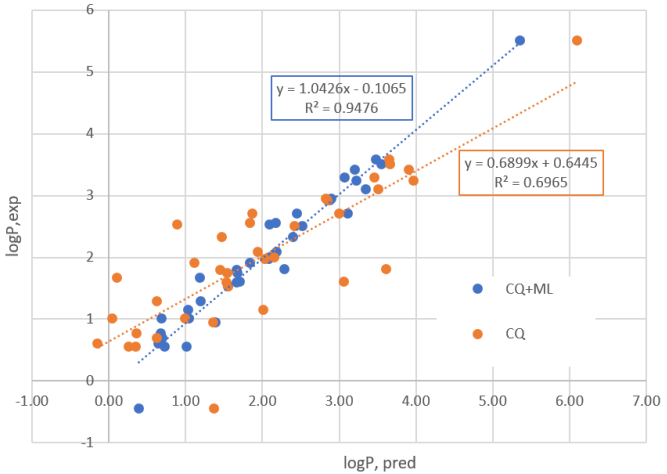
XGBoost

data type	n	source
Training & crossval.	10964	PHYSPROP
Test set I	40	Slater et al. J Pharm Sci 1994, 83,1280.
Test set II	37	PHYSPROP subset via substructure search

variable(i.e. descriptor) importance



test set II results



Gradient Tree Boosting

XGBoost library: https://xgboost.readthedocs.io/en/latest/index.html

- State-of-the-art in machine learning for tabular data
- Implementation of stochastic gradient boosting*

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Predicted values | learning rate | decision tree

$$\gamma_m = \arg\min \sum_{i=1}^{n} L(y_i, F_{m-1}(x) + \gamma_m h_m(x))$$

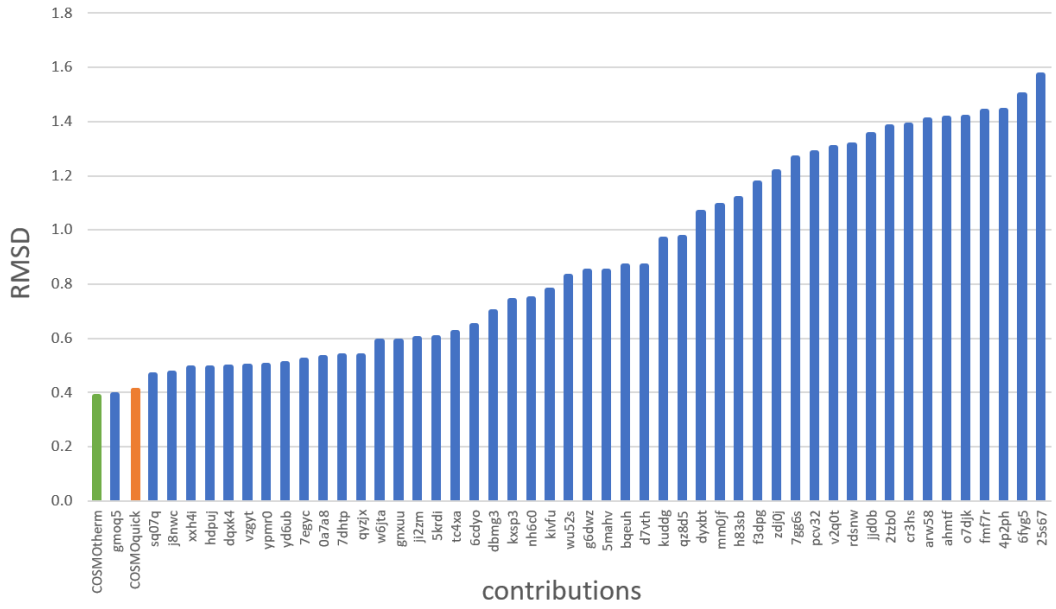
Loss function, e.g. squared error

Winning list in ML competitions: https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions

*Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378. http://dx.doi.org/10.1016/S0167-9473(01)00065-2

Final Results

COSMO-RS based methods



1st in RMSD
COSMO*therm*

3rd in RMSD
COSMO*quick*

- QSPR
- Deep Learning
- SMD
- Molecular Dynamics
- 3D-RISM

Results

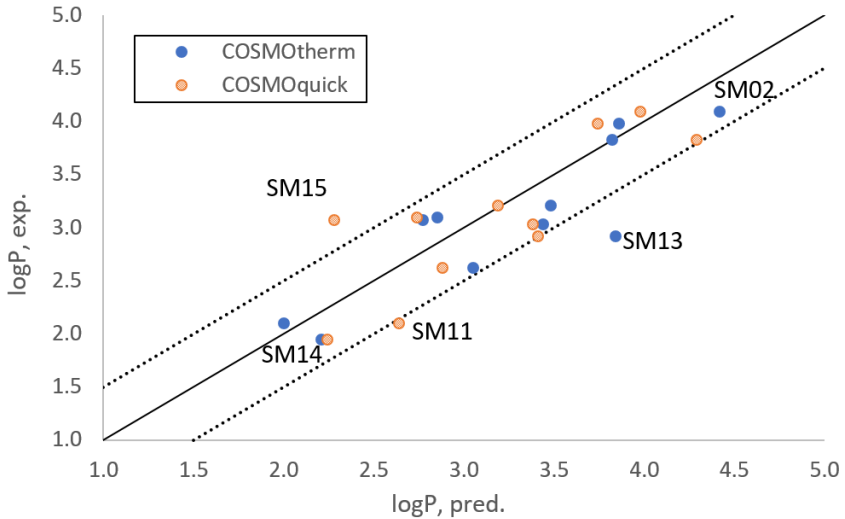
Comparison of COSMO-RS Submissions

id	σ-surface from	Method	level	RMSE
hmz0n	COSMO files	Turbomole	FINE19	0.38
3vqbi	SMILES	COSMOquick	TZVP+ML	0.41
(not submitted)	COSMO files	COSMOquick	TZVP+ML	0.35

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

Results

Comparison of COSMO*therm* Submissions vs experiment

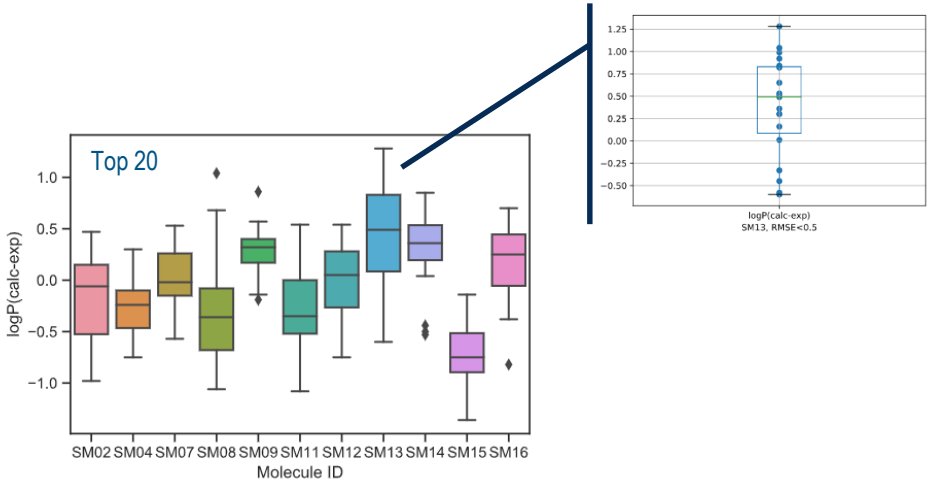


Correlations between COSMO*therm* based submissions: R^2=0.76

SM15 (outlier COSMO*quick*) : bad database representation

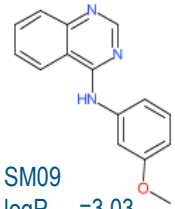
SM13 (Outlier COSMO*therm*):?

SM13



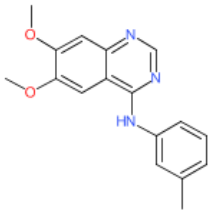
Deviations from experiment for logP predictions (logP(calc)-logP(exp)) of the best 20 submissions.

Consistency check of SM13 via group contributions* ($logP_{exp}$=2.92):

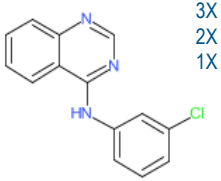


SM09
$logP_{exptl}$=3.03

2X -CH3	[aliphatic carbon]	0.55
1X -O-	[oxygen, one aromatic attach]	-0.47
1X Ring reaction ->	1,2-dialkyloxy	-0.52

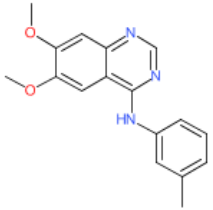


SM13
3.14



SM12
$logP_{exptl}$=3.83

-1X -CL	[chlorine, aromatic attach]	-0.65
3X -CH3	[aliphatic carbon]	0.55
2X -O-	[oxygen, one aromatic attach]	-0.47
1X Ring reaction ->	1,2-dialkyloxy	-0.52



SM13
3.38

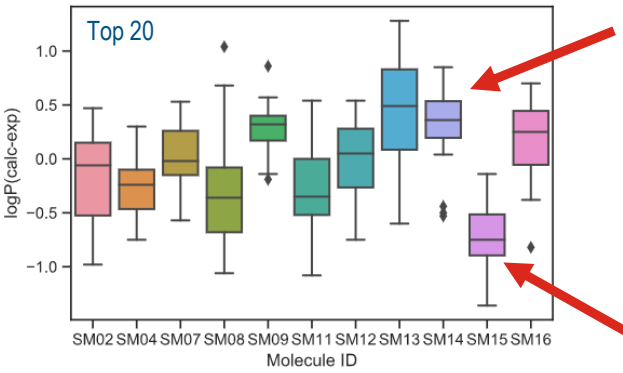
- There is a clear trend for an SM13 logP overestimation
- Finite dilution effect, i.e. aggregation in aqueous solution ?

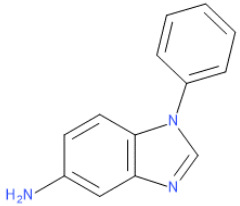
*Increments from Kowwin[tm]

SM15

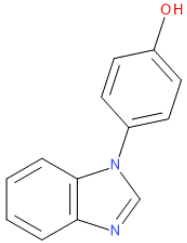
Most models predict logP too low, i.e.

- overestimate the water solubility





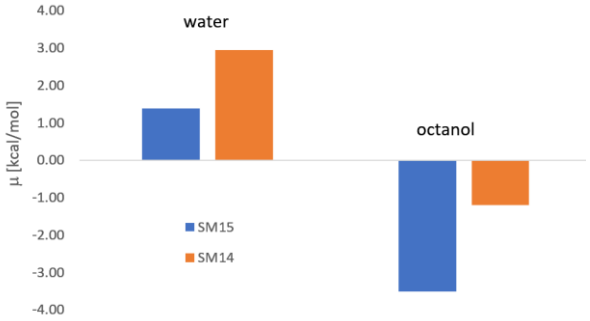
SM14 $logP_{exp}$=1.95



SM15 $logP_{exp}$=3.07

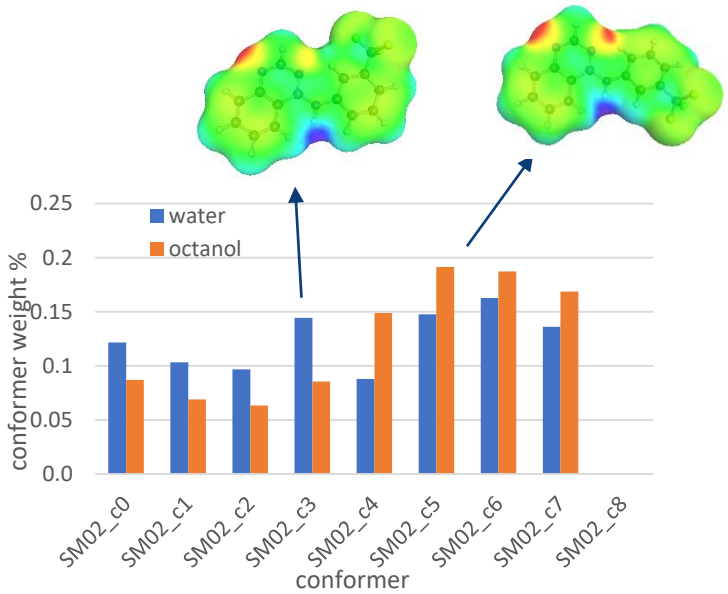
ID	logP,exp	logP,mean	logP,CT
SM15	3.1	2.3	2.8
SM14	2.0	2.2	2.2

COSMO-RS chemical potentials:



- The main reason for the high logP is the low chemical potential in octanol!

Conformational Effects



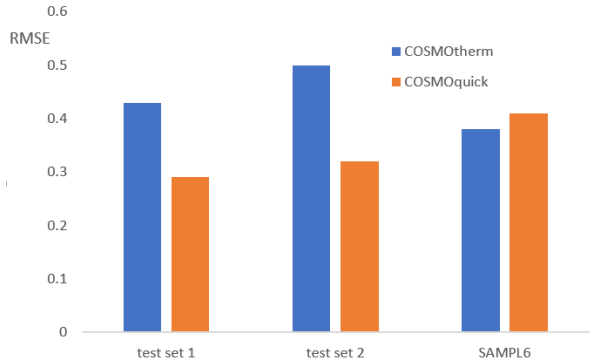
Root mean squared deviation (RMSD) from SAMPL6 experimental data using different conformer generation methods.

method	Conformer sampling	RMSD
COSMOtherm	random conformers	0.45
COSMOtherm	COSMOconf	0.38

- Conformations are Boltzmann weighted, iterative computation of chemical potentials
- Set of rather rigid compounds

Test set data

And comparison with SAMPL6 results



Decrease of COSMOquick/ML performance:

- Database representation of SAMPL6 substructures

Improvement of COSMOtherm performance:

- Test set (literature) data not as clean as SAMPL6 data!

Need for more accurate experimental data!

SAMPL6 logP

Summary & Learnings

- Need for more such high quality experiments!
- Difficulties in finding useful, i.e. predictive testsets
- Importance of conformational effects (even for less flexible compounds) or suitable descriptors
- SAMPL6 compounds are somewhat underrepresented in COSMO*quick* σ-*surface* database
- 3 of top 4 entries used stochastic gradient boosting (XGBoost)

Many thanks for your attention!

contact:
christoph.loschen@3ds.com