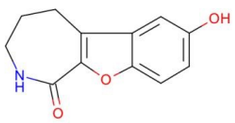
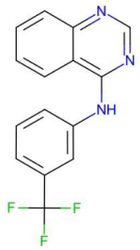
pKa prediction on the SAMPL6 dataset
— Lessons learned —

Bogdan I. Iorga, Edithe Selwa
Institut de Chimie des Substances Naturelles, CNRS
Gif-sur-Yvette, France

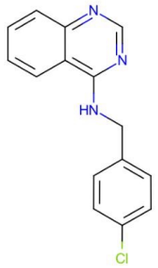
Oliver Beckstein, Ian Kenney
Arizona State University
Phoenix, AZ

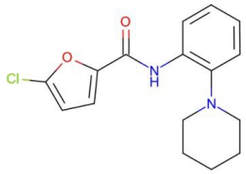
SAMPL6 pKa challenge dataset and requirements

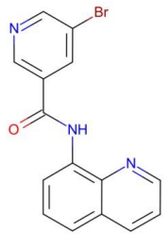

SM01

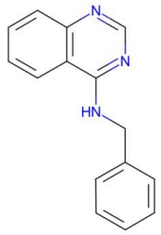

SM02

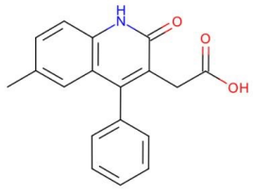

SM03

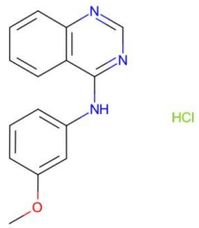

SM04

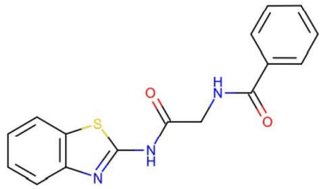

SM05

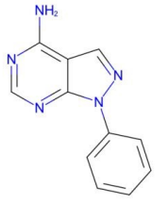

SM06

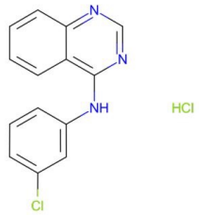

SM07

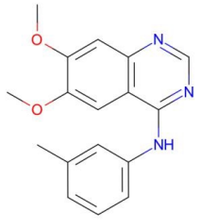

SM08

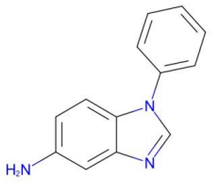

SM09

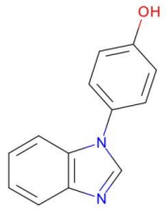

SM10

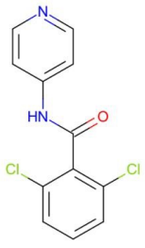

SM11

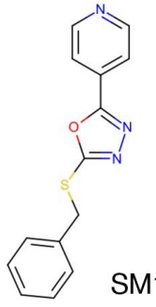

SM12

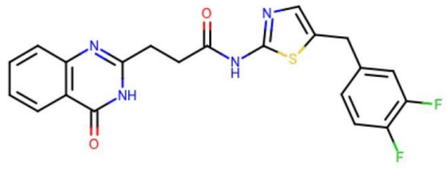

SM13

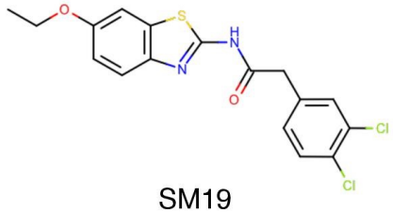

SM14

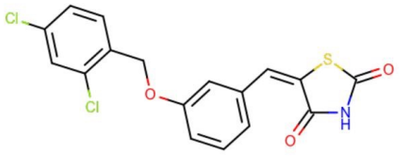

SM15

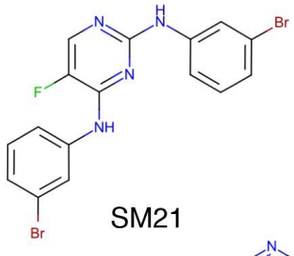

SM16

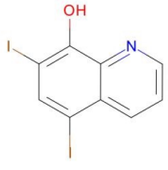

SM17

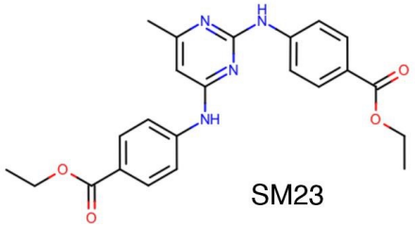

SM18

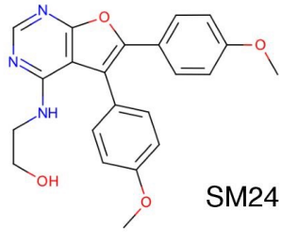

SM19


SM20


SM21


SM22


SM23


SM24

Microstates provided:

- different protonation states and tautomers for 24 compounds (SMILES)

Requirements:

- **Prediction Type I -** microscopic pKas
- **Prediction Type II -** microstate populations as a function of pH
- **Prediction Type III -** macroscopic pKas

Choice of the method (I): Constant pH molecular dynamics (CpHMD)

- Our previous participations to SAMPL challenges [1-3]: free energy calculations using MD simulations, with MDPOW (https://github.com/Becksteinlab/MDPOW) and OPLS-AA force field

- At first sight, CpHMD looks very interesting

☑ provides easily populations for microstates

☑ implemented and thoroughly tested in AMBER for protein residues

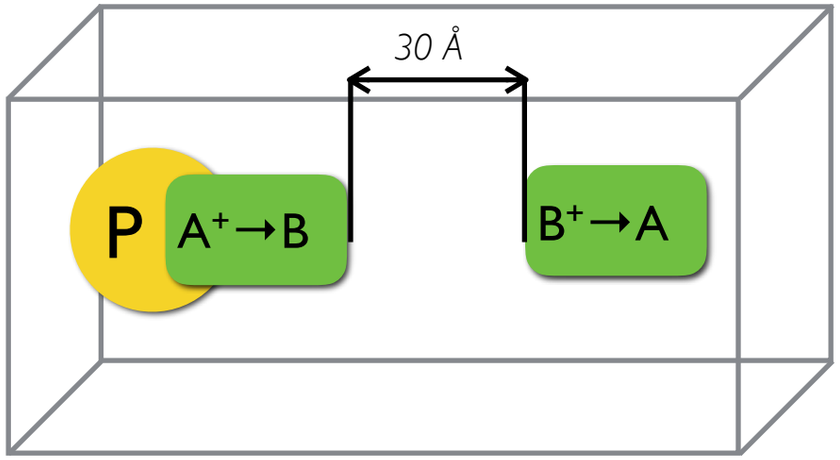
○ CpHMD-related parametrization of all microstates in GAFF/AMBER: huge endeavor and missing expertise in our team

○ need for a reference compound with known experimental pKa value: blocking

1. Beckstein, O.; Iorga, B.I. *JCAMD* **2012,** *26*, 635-645.
2. Beckstein, O.; Fourrier, A.; Iorga, B.I. *JCAMD* **2014,** *28*, 265-276.
3. Kenney, I.M.; Beckstein, O.; Iorga, B.I. *JCAMD* **2016,** *30*, 1045-1058.

Choice of the method (II): Free energy calculations using non-equilibrium (fast-growth)

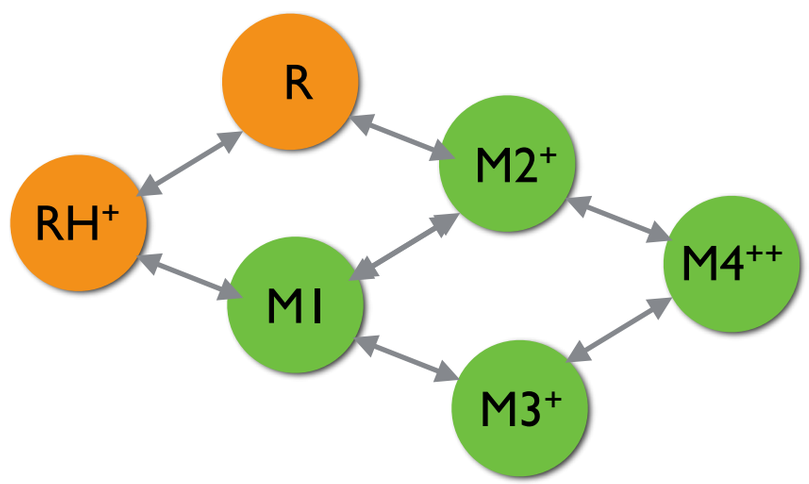
- Protocol developed in our group based on PMX [4] and used in our participation to D3R GC2 [5] and GC3 for computing relative protein-ligand affinities
- Possibility of computing relative free energies for transformations involving a change in the overall charge of the molecule

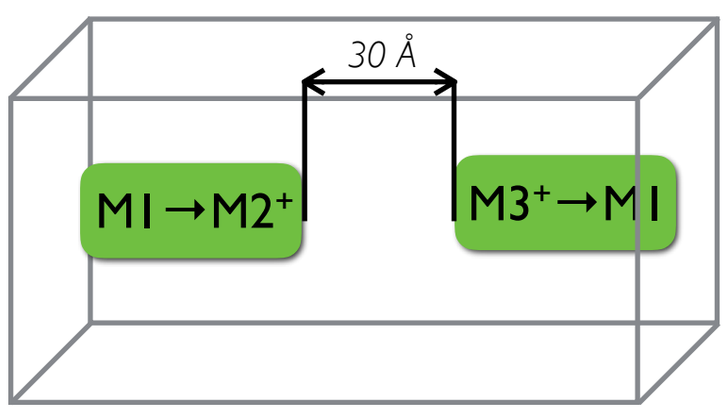


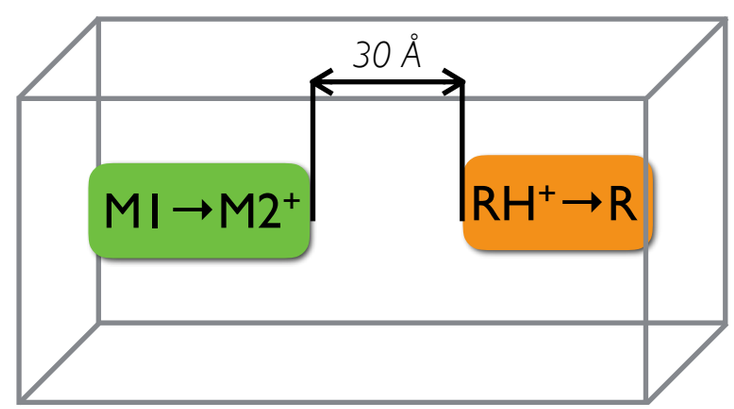
4. Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B.L. *J. Comput. Chem.* **2015**, *36*, 348-354.
5. Selwa, E.; Elisée, E.; Zavala, A.; Iorga, B.I. *JCAMD* **2018**, *32*, 273-286.

Free energy calculations using fast-growth molecular dynamics — pKa prediction —

- **relative pKa values for all microstates** – coupling with a SINGLE reference compound with known experimental value provides absolute pKa values for all microstates – general method

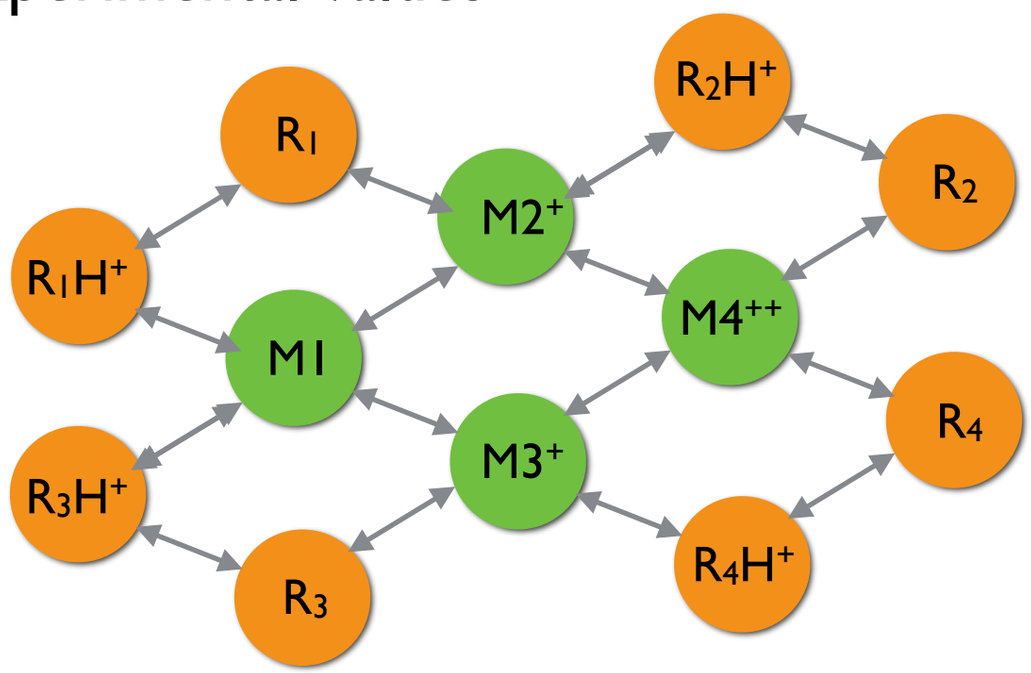


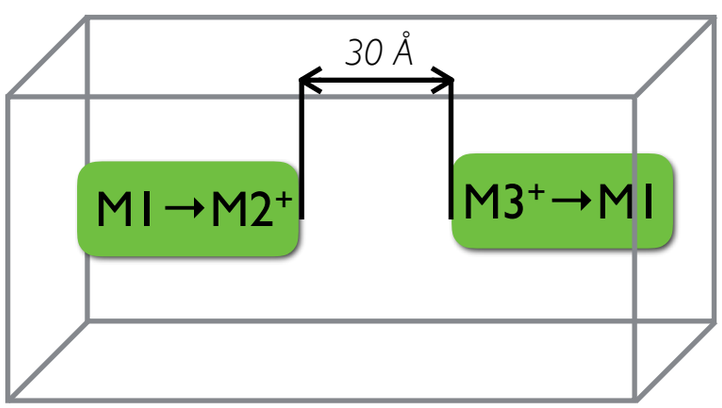


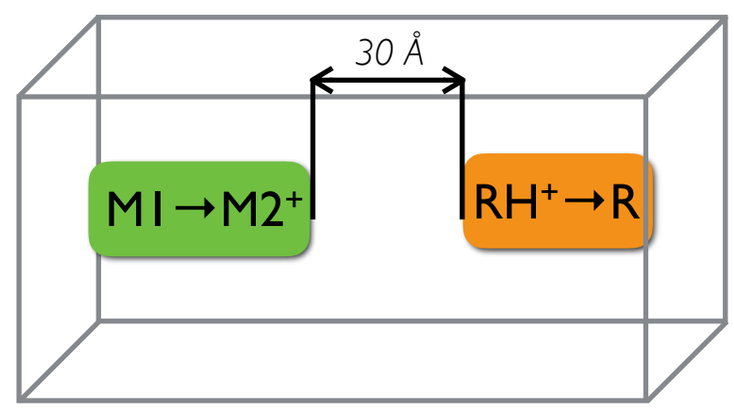


Free energy calculations using fast-growth molecular dynamics — pKa prediction —

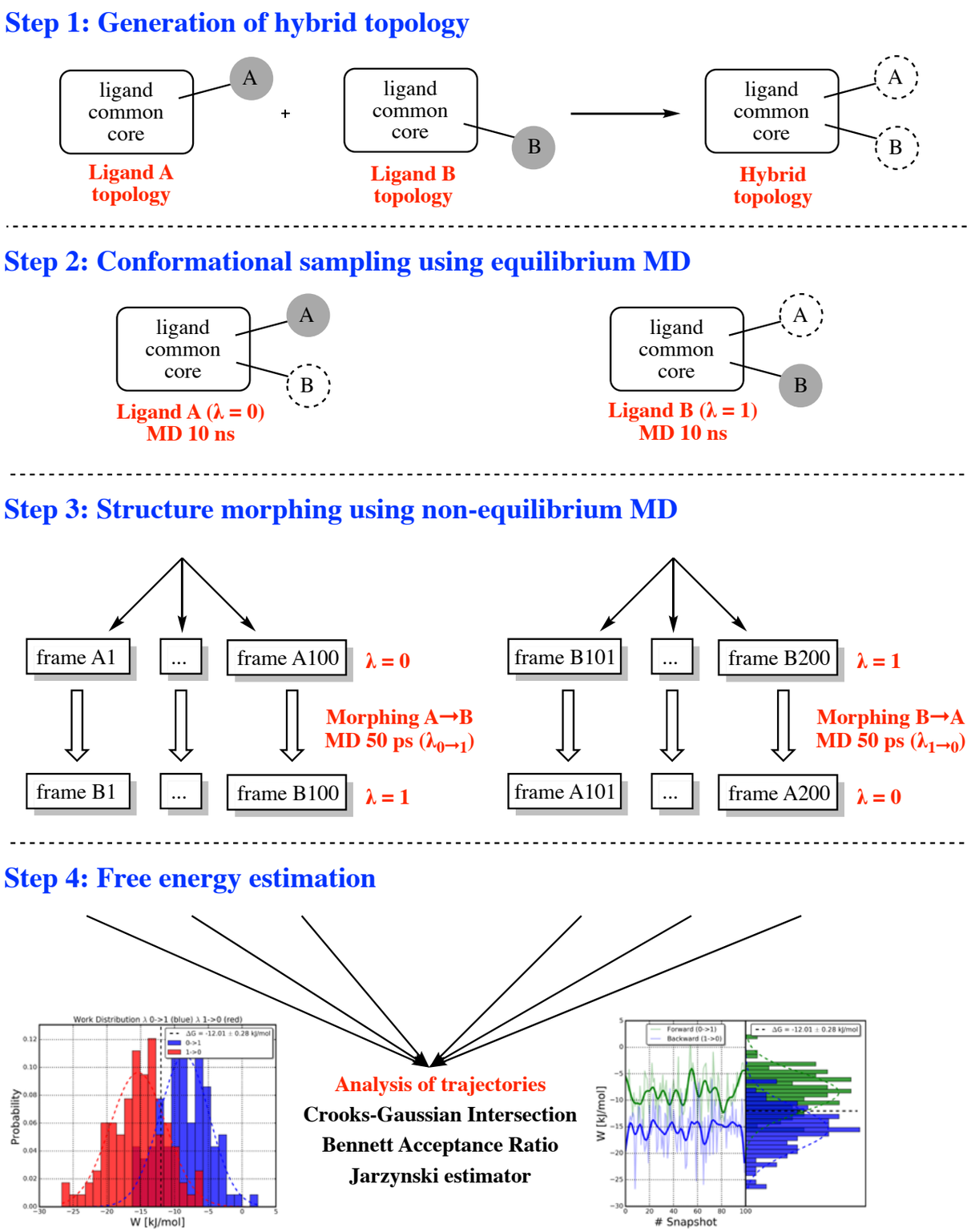
- **absolute pKa values for all microstates** – coupling every microstate with a reference compound with known experimental value – simple systems – probably better in terms of error propagation, but depending more on the accuracy of pKa experimental values







Protocol for non-equilibrium (fast-growth) simulations



- **Force field:** mainly OPLS-AA and GAFF/AMBER
- **Ligand parametrization:** MOL2FF (OPLS-AA) and AmberTools (GAFF/AMBER)

Ligandbook

Domański, J.; Beckstein, O.; Iorga, B. I., Ligandbook – an online repository for small and drug-like molecule force field parameters. *Bioinformatics* **2017**, *33*, 1747-1749 (https://ligandbook.org).

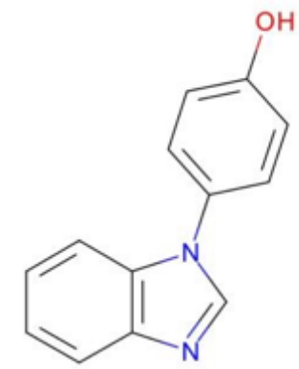
- **Generation of hybrid topologies:** in house scripts
- **MD simulations:** Gromacs
- **Analysis:** PMX [4]

4. Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B.L. *J. Comput. Chem.* **2015**, *36*, 348-354.

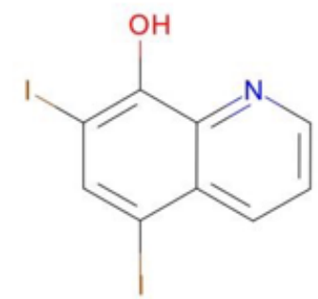
Selwa, E.; Elisée, E.; Zavala, A.; Iorga, B.I. *JCAMD* **2018**, *32*, 273-286.

Free energy calculations using fast-growth molecular dynamics — intrinsic pKa prediction error —

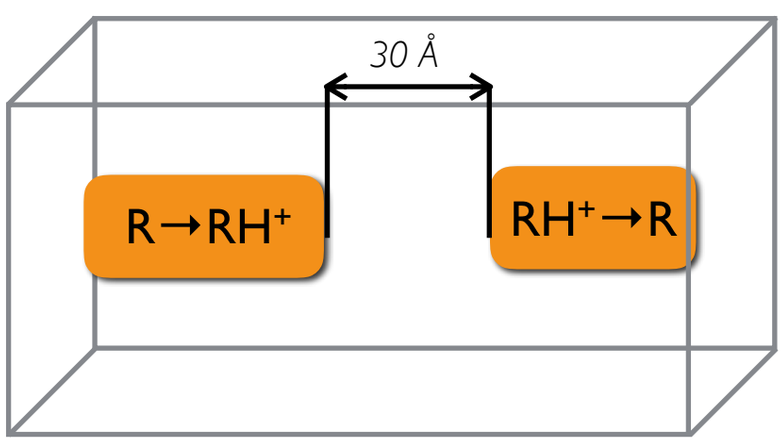
Transformation	Force field	ΔΔG (kJ/mol)
phenolate/phenol	**OPLS-AA**	**–4.65 ± 1.46**
benzimidazolium/benzimidazole	**OPLS-AA**	**0.61 ± 0.48**
SM15_micro001/SM15_micro004	**OPLS-AA**	**–3.56 ± 0.45**
SM15_micro003/SM15_micro002	**OPLS-AA**	**32.77 ± 0.43**
SM20_micro003/SM20_micro004	**OPLS-AA**	**2.50 ± 1.06**
SM22_micro001/SM22_micro004	**OPLS-AA**	**15.57 ± 2.75**
SM22_micro002/SM22_micro001	**OPLS-AA**	**0.05 ± 0.55**
SM22_micro002/SM22_micro003	**OPLS-AA**	**–2.63 ± 1.20**
SM22_micro003/SM22_micro004	**OPLS-AA**	**5.39 ± 1.88**

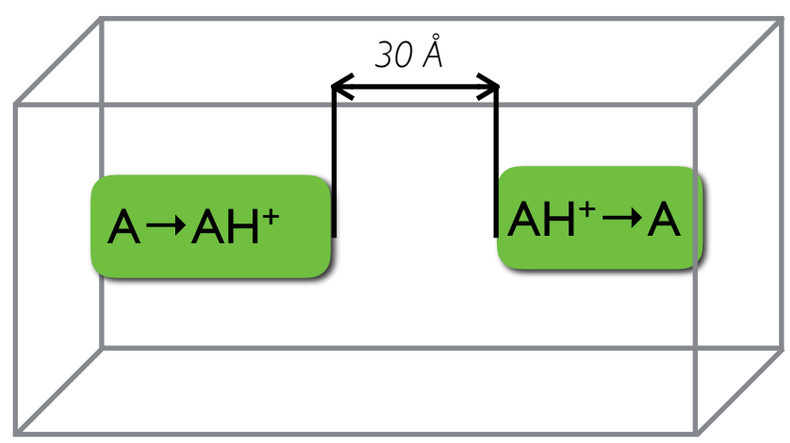


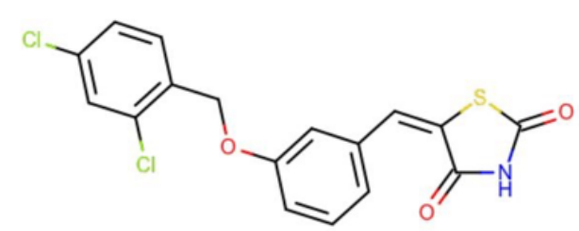
SM15



SM22





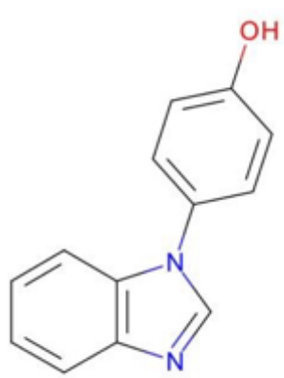


SM20

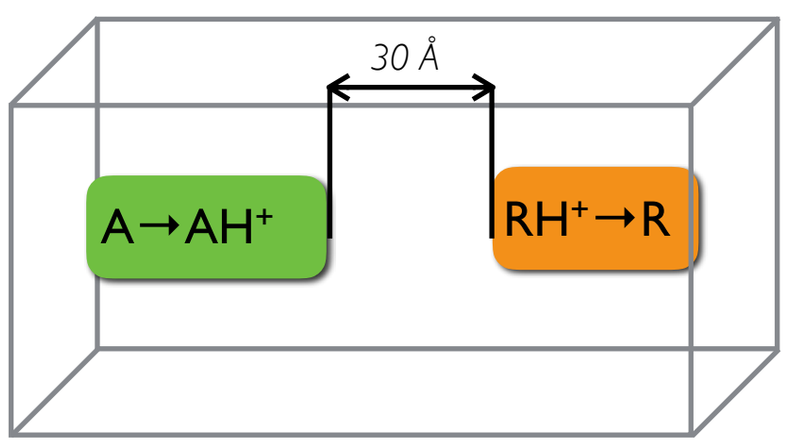
Free energy calculations using fast-growth molecular dynamics — microscopic pKa prediction —

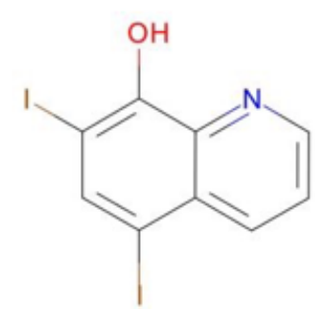
Transformation	Reference	Force field	ΔΔG (kJ/mol)
SM15_micro002/SM15_micro004	phenolate/phenol	**GAFF/AMBER**	**227.06 ± 0.51**
SM15_micro003/SM15_micro001	phenolate/phenol	**GAFF/AMBER**	**370.85 ± 0.51**
SM22_micro001/SM22_micro004	phenolate/phenol	**GAFF/AMBER**	**400.28 ± 0.44**
SM22_micro002/SM22_micro001	pyridine/pyridinium	**GAFF/AMBER**	**442.05 ± 0.39**
SM22_micro002/SM22_micro003	phenolate/phenol	**GAFF/AMBER**	**241.93 ± 0.58**
SM22_micro003/SM22_micro004	pyridine/pyridinium	**GAFF/AMBER**	**623.25 ± 0.39**

Predictions not submitted.



SM15



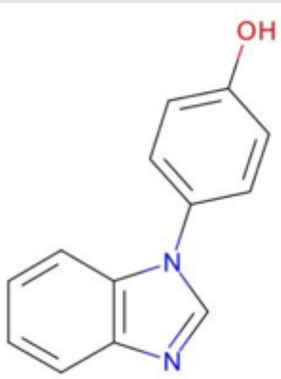


SM22

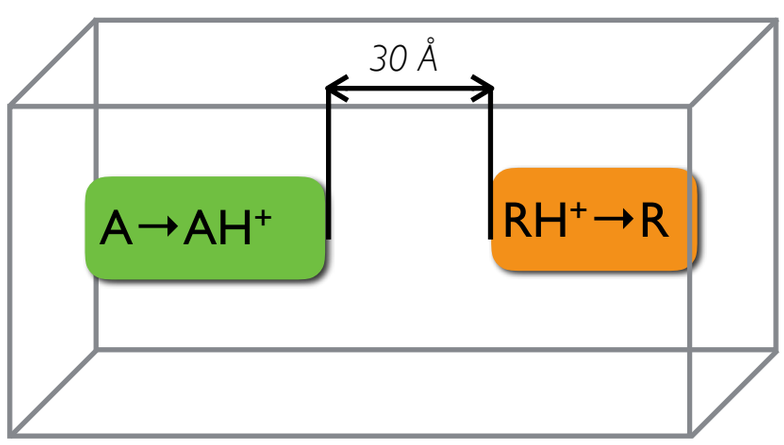
Free energy calculations using fast-growth molecular dynamics — microscopic pKa prediction —

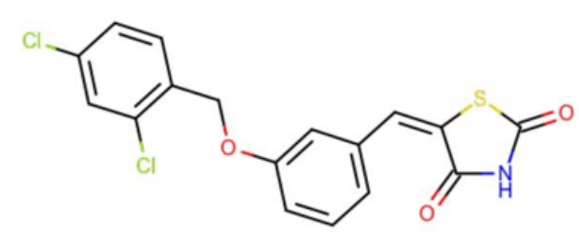
Transformation	Reference	Force field	ΔΔG (kJ/mol)
SM15_micro002/SM15_micro004	phenolate/phenol	**OPLS-AA**	**–8.45 ± 2.15**
SM15_micro003/SM15_micro001	phenolate/phenol	**OPLS-AA**	**–10.03 ± 2.26**
SM15_micro001/SM15_micro004	benzimidazolium/benzimidazole	**OPLS-AA**	**100.81 ± 0.40**
SM15_micro003/SM15_micro002	benzimidazolium/benzimidazole	**OPLS-AA**	**116.08 ± 0.42**
SM20_micro003/SM20_micro004	phthalimidate/phthalimide	**OPLS-AA**	**911.67 ± 1.28**
SM22_micro001/SM22_micro004	phenolate/phenol	**OPLS-AA**	**204.15 ± 2.18**
SM22_micro002/SM22_micro001	pyridine/pyridinium	**OPLS-AA**	**227.31 ± 0.47**
SM22_micro002/SM22_micro003	phenolate/phenol	**OPLS-AA**	**141.45 ± 1.36**
SM22_micro003/SM22_micro004	pyridine/pyridinium	**OPLS-AA**	**289.16 ± 0.83**

Predictions not submitted.

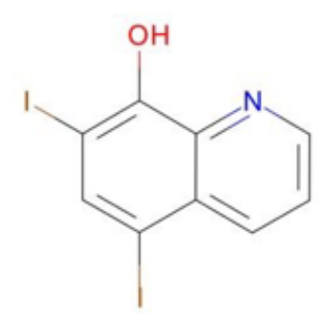


SM15





SM20



SM22

Choice of the method (III): DFT calculations with correction for systematic errors

- Protocol described in [6] using Gaussian09:
  - geometry optimization B3LYP/6-311+G(d,p) 5d : $E_{(g)}$
  - vibrational frequency calculation B3LYP/6-311+G(d,p) 5d: $G^0_{(g)}$
  - single-point energy calculation CPCM: $E_{(g)} + \Delta G^0_{solv}$
  - $G^*(H^+_{(s)})$: –272.2 kcal/mol (literature value)

$$pK_a = \Delta G^*_a / RT\ln(10) \qquad \Delta G^*_a = G^*\left(A^-_{(s)}\right) + G^*\left(H^+_{(s)}\right) - G^*\left(HA_{(s)}\right)$$

- Values computed for a reference dataset (38 simple inorganic and organic compounds with known experimental pKa values) and for all microstates.
- Computing time: between <1 min and 7 hours on a 8-core machine
- For reference compounds:

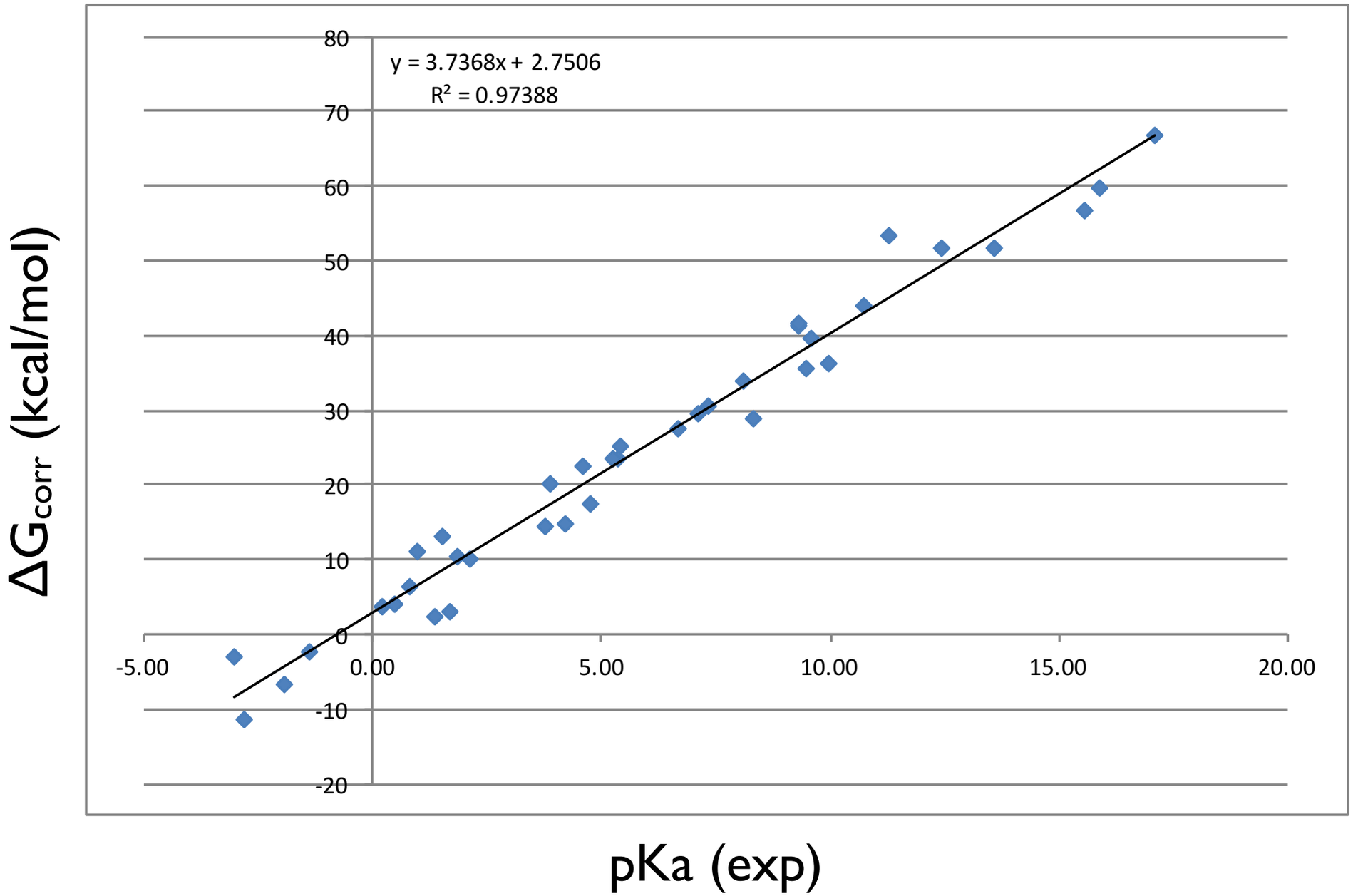
$$\Delta G_{corr} = RT\ln(10)\ [\ pKa(exp) - pKa(calc)]$$

6. Muckerman, J.T.,;Skone, J.H.; Ning, M.; Wasada-Tsutsui, Y. *Biochim. Biophys. Acta* **2013**, *1827*, 882-891.

Reference dataset

Acid (neutral)	Base (negative)	pKa	Acid (positive)	Base (neutral)	pKa
nitric acid	nitrate	-1.40	anilinium	aniline	4.62
methanol	methanolate	15.54	4-nitro anilinium	4-nitro aniline	0.98
trifluoroethanol	trifluoroethanolate	12.43	4-bromo anilinium	4-bromo aniline	3.89
hexafluoroisopropanol	hexafluoroisopropanolate	9.30	4-methoxy anilinium	4-methoxy aniline	5.36
nonafluorotertbutanol	nonafluorotertbutanolate	5.40	2,5-dichloro anilinium	2,5-dichloro aniline	1.53
acetic acid	acetate	4.76	pyridinium	pyridine	5.24
nitro acetic acid	nitro acetate	1.68	3-nitro pyridinium	3-nitro pyridine	0.81
phthalimide	phthalimidine	8.30	2-chloro pyridinium	2-chloro pyridine	0.49
uracil	uracil anion	9.50	4-cyano pyridinium	4-cyano pyridine	1.86
sulfuric acid	hydrosulfate	-3.00	2,4,6-collidinium	2,4,6-collidine	7.33
benzenesulfonic acid	benzenesulfonate	-2.80	2,6-dimethyl pyridinium	2,6-dimethyl pyridine	6.70
methanesulfonic acid	methanesulfonate	-1.90	4-dimethylamino pyridinium	4-dimethylamino pyridine	9.60
trifluoroacetic acid	trifluoroacetate	0.23	benzylammonium	benzylamine	9.30
formic acid	formate	3.77	triethylammonium	triethylamine	10.72
benzoic acid	benzoate	4.21	pyrrolidinium	pyrrolidine	11.27
ethanol	ethanolate	15.90	guanidinium	guanidine	13.60
isopropanol	isopropanolate	17.10	hydrazinium	hydrazine	8.12
oxalic acid	oxalate	1.38			
phosphoric acid	dihydrophosphate	2.15			
phenol	phenolate	9.95			
4-nitro phenol	4-nitro phenolate	7.14			

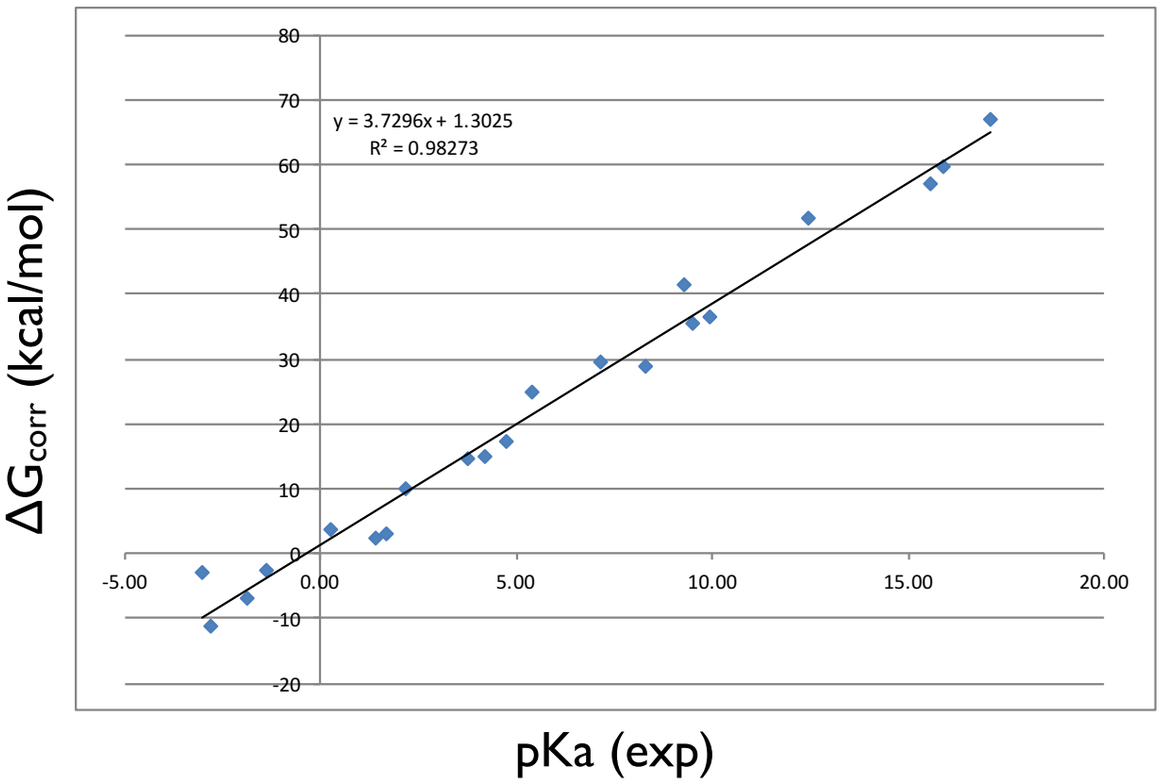
Reference dataset: global linear regression — correction for a systematic error —



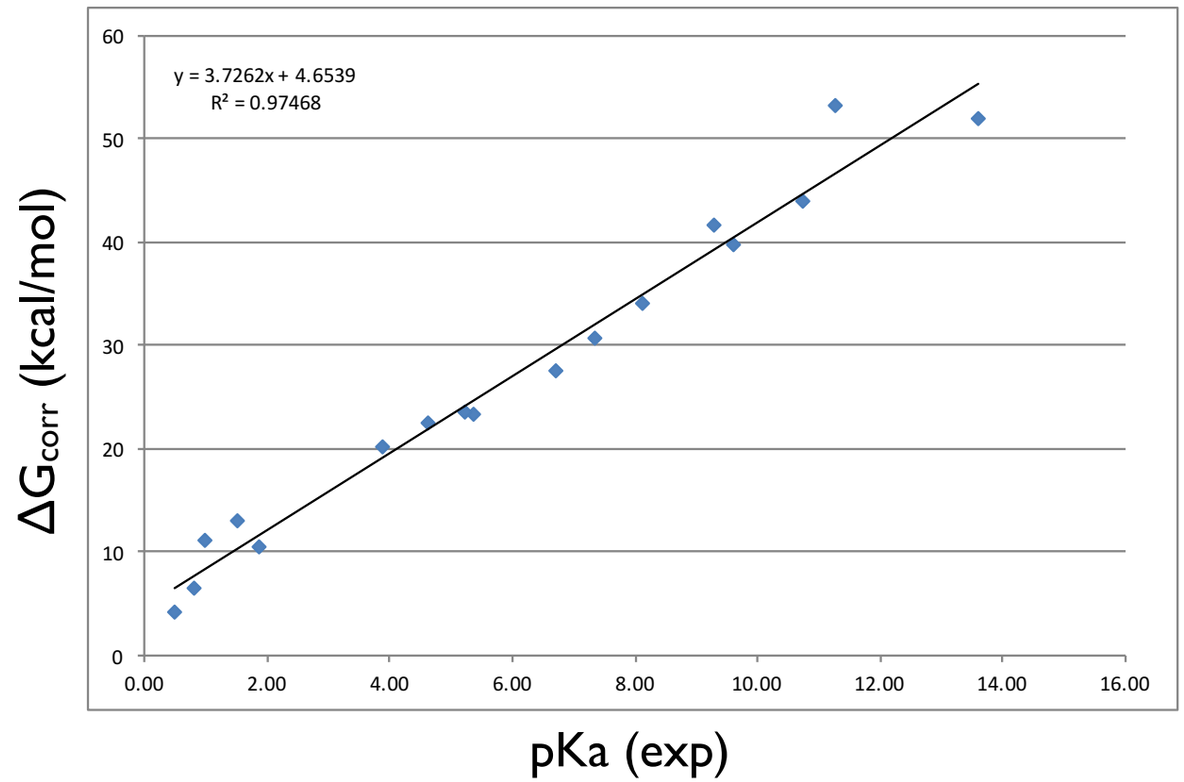
$y = 3.7368x + 2.7506$ $R^2 = 0.97388$

Reference dataset: linear regression for two separate datasets — correction for a systematic error —

Neutral acids



Positively charged acids



$y = 3.7296x + 1.3025$ $R^2 = 0.98273$

$y = 3.7262x + 4.6539$ $R^2 = 0.97468$

$y = 3.7368x + 2.7506$ $R^2 = 0.97388$

pKa prediction for the SAMPL6 dataset

- The same protocol applied to all microstates
- Two types of corrections applied: either global or according to the charge of the acid form ⇒ two separate submissions
- Global correction expected to be worse than separate corrections
- Type I predictions submitted for all compounds
- Type II and type III predictions submitted only for SM15, SM20 and SM22.

Compound	pKa experimental	pKa predicted (global correction)	pKa predicted (separate corrections)
SM15	4.70 ± 0.01	5.21	6.14
	8.94 ± 0.01	11.41	10.64
SM20	5.70 ± 0.03	8.34	7.58
SM22	2.40 ± 0.02	1.31	2.02
	7.43 ± 0.01	7.93	7.41
RMSE		**1.72 (35bdm)**	**1.32 (p0jba)**

Macroscopic pKa from microstates

- Input: relative free energies between micro states $1 \leq k \leq M$

$$\Delta G_{kl} = G_k - G_l$$

Note: we discuss free energies, but we can easily convert

$$\beta = \frac{1}{kT} \qquad \mathrm{p}K_\mathrm{a} = \beta \Delta G / \ln 10$$

- Free energy of micro state *k* (relative to arbitrary reference state *k*=1):

$$\Delta G_k = G_k - G_1$$

- ΔG_k can be computed from sums of ΔG_{mn} that connect state 1 and state *k*.

- "Macroscopic" variable: number of protons *N*, with N_k the number of protons in state *k.*

- $\delta_{Nk,N}$ is an "indicator" to pick out states with *N* protons

Macroscopic pKa from microstates

- Free energy difference between two macro states *N* and *N*-1 (i.e., adding one proton)

$$\beta \Delta G_{N,N-1} = -\ln \left[\frac{\sum_{k=1}^{M} \exp(-\beta \Delta G_k) \delta_{N_k,N}}{\sum_{k=1}^{M} \exp(-\beta \Delta G_k) \delta_{N_k,N-1}} \right]$$

- Free energy difference for protonation reaction (taking into account hydration free energy of the proton)

$$\mathrm{A}^- + \mathrm{H}^+ \rightleftharpoons \mathrm{HA}$$

$$\Delta G = \Delta G_{N,N-1} - \Delta G_{\mathrm{hyd}}(\mathrm{H}^+)$$

- Effective (or “macroscopic” pKa):

$$\mathrm{p}K_\mathrm{a} = \beta \Delta G / \ln 10$$

Concluding remarks

- Approaches involving QM and QSPR methods seem to be currently the best options for computing pKa values, in terms of speed and accuracy (method dependent).
- MD simulations (free energy calculations) represent an interesting alternative, for the moment hampered by difficulties in parametrization of species involved, the computing time required and the quality of force field parameters.
- In the tradition of SAMPL challenges, after hydration free energies and partition coefficients, pKas have a huge potential in the improvement of currently available force fields.

Acknowledgements

D3R/SAMPL organizers

ICSN, CNRS, Gif-sur-Yvette, France

Edithe SELWA

Agustin ZAVALA

Eddy ELISEE

Ludovic CHAPUT

Hristo NEDEV

Pascal RETAILLEAU

Arizona State University

Oliver BECKSTEIN

Ian KENNEY

Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Bert DE GROOT (PMX)

LabEx LERMIT
Laboratory of Excellence in Research on Medication and Innovative Therapeutics
ANR-10-LABX-33

Campus Paris Saclay
FONDATION DE COOPERATION SCIENTIFIQUE

ICSN Institut de Chimie des Substances Naturelles

îledeFrance

SATT PARIS-SACLAY

jpiamr
DesInMBL
Structure-guided design of pan-inhibitors of metallo-β-lactamases