HAL
archives-ouvertes.fr

SAMPL6: calculation of macroscopic pKa values from ab initio quantum mechanical free energies

Edithe Selwa, Ian Kenney, Oliver Beckstein, Bogdan Iorga

▶ To cite this version:

Edithe Selwa, Ian Kenney, Oliver Beckstein, Bogdan Iorga. SAMPL6: calculation of macroscopic pKa values from ab initio quantum mechanical free energies. Journal of Computer-Aided Molecular Design, Springer Verlag, 2018, 32 (10), pp.1203-1216. 10.1007/s10822-018-0138-6 . hal-02119827

HAL Id: hal-02119827
https://hal.archives-ouvertes.fr/hal-02119827

Submitted on 23 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Manuscript LateX

Click here to download Manuscript manuscript-sampl6-revision.tex

Click here to view linked References

Journal of Computer-Aided Molecular Design manuscript No.
(will be inserted by the editor)

SAMPL6: Calculation of macroscopic pK_a values from *ab initio* quantum mechanical free energies

Edithe Selwa · Ian M. Kenney · Oliver Beckstein · Bogdan I. Iorga

Received: date / Accepted: date

Abstract Macroscopic pK_a values were calculated for all compounds in the SAMPL6 blind prediction challenge, based on quantum chemical calculations with a continuum solvation model and a linear correction derived from a small training set. Microscopic pK_a values were derived from the gas-phase free energy difference between protonated and deprotonated forms together with the Conductor-like Polarizable Continuum Solvation Model and the experimental solvation free energy of the proton. pH-dependent microstate free energies were obtained from the microscopic pK_as with a maximum likelihood estimator and appropriately summed to yield macroscopic pK_a values or microstate populations as function of pH. We assessed the accuracy of three approaches to calculate the microscopic pK_as: direct use of the quantum mechanical

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01GM118772 (to OB). BII was supported in part by grants ANR-10-LABX-33 (LabEx LERMIT) and ANR-14-JAMR-0002-03 (JPIAMR) from the French National Research Agency (ANR), and by a grant DIM MAL-INF from the Région Ile-de-France.

E. Selwa
Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Université Paris-Saclay, Labex LERMIT,
1 Avenue de la Terrasse, 91198 Gif-sur-Yvette, France

I. M. Kenney
Department of Physics, Arizona State University, P.O. Box 871504, Tempe, AZ 85287-1504, USA

O. Beckstein
Department of Physics and Center for Biological Physics, Arizona State University, P.O. Box 871504,
Tempe, AZ 85287-1504, USA
Tel.: +1 480 727 9765
Fax: +1 480 965-4669
E-mail: oliver.beckstein@asu.edu

B. I. Iorga
Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Université Paris-Saclay, Labex LERMIT,
1 Avenue de la Terrasse, 91198 Gif-sur-Yvette, France
Tel.: +33 1 69 82 30 94
Fax: +33 1 69 07 72 47
E-mail: bogdan.iorga@cnrs.fr

E. Selwa and I. M. Kenney contributed equally to this work.

free energy differences and correction of the direct values for short-comings in the QM solvation model with two different linear models that we independently derived from a small training set of 38 compounds with known pK_a. The predictions that were corrected with the linear models had much better accuracy [root-mean-square error (RMSE) 2.04 and 1.95 pK_a units] than the direct calculation (RMSE 3.74). Statistical measures indicate that some systematic errors remain, likely due to differences in the SAMPL6 data set and the small training set with respect to their interactions with water. Overall, the current approach provides a viable physics-based route to estimate macroscopic pK_a values for novel compounds with reasonable accuracy.

Keywords pK_a · pH · quantum chemistry · SAMPL challenge

1 Introduction

The SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) challenges allow the molecular modeling community to assess, in “blind” conditions, the accuracy and efficiency of current computational chemistry methods and tools, leading to continuous improvements of the available computational methods. The previous SAMPL challenges [1–5] involved hydration free energy calculations, with the exception of the last edition, SAMPL5, which was dedicated to the prediction of distribution coefficients [6]. Our past participations in SAMPL challenges [7–9] represented unique opportunities for us to test our approaches and to develop and improve new computational tools. In 2018, the SAMPL6 challenge focused on the prediction of microscopic and macroscopic pK_a values for fragment-like organic compounds.

The equilibrium acid dissociation reaction in aqueous solution

$$\mathrm{HA}_{(s)} \rightleftharpoons \mathrm{H}^+_{(s)} + \mathrm{A}^-_{(s)} \tag{1}$$

with acid dissociation constant $K_a = [\mathrm{A}^-][\mathrm{H}^+]/[\mathrm{HA}]$ is of broad importance in biological systems, in synthetic chemistry, and pharmacology [10–14]. The pK_a, defined as

$$\mathrm{p}K_a = -\log_{10}\frac{K_a}{c_0} \tag{2}$$

for the standard state concentration $c_0 = 1\,\mathrm{mol/l}$, measures thermodynamic acidity. The theoretical prediction of pK_a values is an ongoing challenge [15]. In the narrow realm of protein biochemistry, good progress has been made in calculating the physiologically important *changes* in pK_as of standard amino acid residues in different environments with accuracies better than 1 pK_a unit [12], especially with constant pH molecular dynamics simulations [16–19], which have been applied to study a wide range of phenomena [20–22]. *Absolute* pK_a calculations of arbitrary molecules using physics-based quantum chemistry approaches (as opposed to machine learning (ML) ones) have been more challenging and accuracy of 1 pK_a unit has been difficult to achieve consistently [15, 23] whereas a range of methods can achieve “chemical accuracy” (defined as 2.5 pK_a units by Ho and Coote [15]). The clear advantage of *ab initio* approaches is that they can be applied to any novel compound. Here we report on pK_a calculations of the 24 compounds in the SAMPL6 challenge (Fig. 1) with a quantum-chemical approach originally developed by Muckerman et al [24]. The SAMPL6 compounds are, however, chemically more complex and 23 contain multiple titratable protons and, in some cases, tautomers so that *macroscopic* pK_a have to be calculated.

The calculation of microscopic pK_as, i.e., the free energy difference for the deprotonation reaction Eq. 1 at the standard state (concentration 1 mol/l and temperature $T = 298.15$ K, indicated by the superscript "*")

$$\mathrm{p}K_\mathrm{a} = \frac{\Delta G^*_{(\mathrm{s})}}{RT\ln 10}, \tag{3}$$

is straightforward using quantum chemical gas-phase calculations. However, it is well-known [15, 23] that direct calculations lead to large errors in the calculated pK_as, mainly due to the poor continuum solvation models that have to be employed in order to obtain free energies in solution. One approach to correct for these systematic errors is to generate a model to correct the raw quantum chemical free energies [24]. We generated linear models from a training set with 38 simple compounds with experimentally known pK_a (Fig. 2 and 3). We fit a global model to all the data (the global linear model) and we split the training set with a simple classifier, namely the charge of the acid, yielding a piecewise linear model with separate linear functions for neutral and cationic acids. We calculated the macroscopic pK_as for all 24 SAMPL6 compounds and compared the accuracy of the three approaches [QM computed (raw), linear fit global, and linear fit piecewise].

2 Methods

Following Muckerman et al [24], our strategy was to compute gas-phase free energy differences

$$\Delta G^\circ_{(\mathrm{g})} = G^\circ(\mathrm{A}^-_{(\mathrm{g})}) + G^\circ(\mathrm{H}^+_{(\mathrm{g})}) - G^\circ(\mathrm{HA}_{(\mathrm{g})}) \tag{4}$$

(denoted as standard state free energies at 1 atm pressure and 298.15 K) for the deprotonation reaction for all titratable protons,

$$\mathrm{HA}_{(\mathrm{g})} \rightleftharpoons \mathrm{H}^+_{(\mathrm{g})} + \mathrm{A}^-_{(\mathrm{g})}. \tag{5}$$

To obtain solution free energy differences corresponding to Eq. 1,

$$\Delta G^*_{(\mathrm{s})} = G^*(\mathrm{A}^-_{(\mathrm{s})}) + G^*(\mathrm{H}^+_{(\mathrm{s})}) - G^*(\mathrm{HA}_{(\mathrm{s})}) \tag{6}$$

(where the standard state refers to 1 mol/l), a solvation free energy contribution $\Delta G^\circ_\mathrm{solv}$ is added to the gas-phase free energies of the acid HA and the base A^- from Eq. 4,

$$G^*_{(\mathrm{s})} = G^\circ_{(\mathrm{g})} + \Delta G^{\circ\to *} + \Delta G^\circ_\mathrm{solv} = G^\circ_{(\mathrm{g})} + \Delta G^*_\mathrm{solv} \tag{7}$$

with $\Delta G^{\circ\to *} = 1.894$ kcal/mol accounting for the change in standard state in the gas phase. The free energy of the proton in the gas phase is calculated analytically in the ideal gas limit (the Sackur-Tetrode equation [25]), $G^\circ(\mathrm{H}^+_{(\mathrm{g})}) = -6.28$ kcal/mol, and for the solvation free energy of the proton we chose the same value as Muckerman

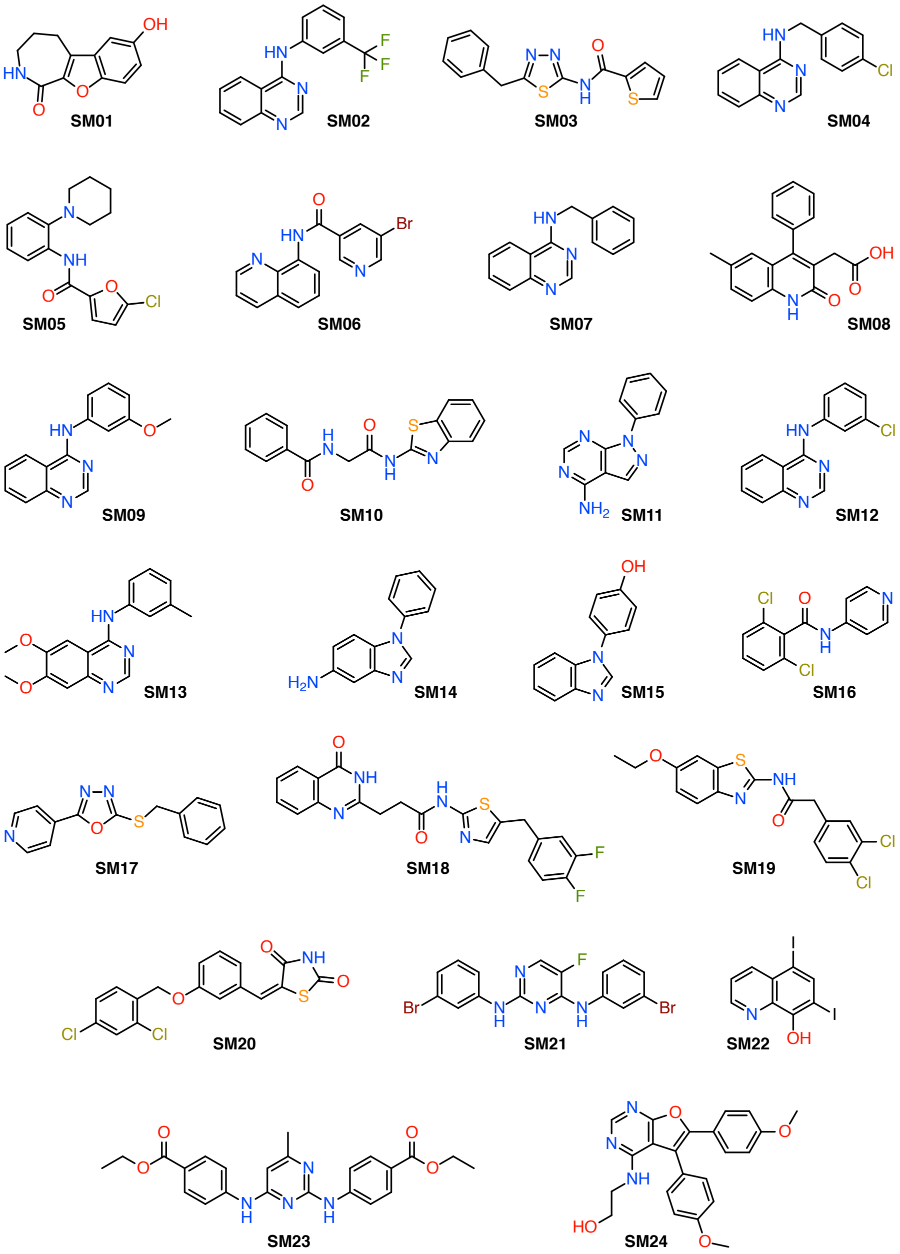


Fig. 1 Chemical structures of the SAMPL6 data set. **SM20** is the only compound that contains a single titratable proton; all other compounds contain multiple titratable protons and, in some cases, tautomers.

et al [24], $G^*(\mathrm{HA}_{(s)}) = -272.2$ kcal/mol although other values are also discussed in the literature [15, 26]. With $\Delta G^*_{(s)}$, the pK_a is calculated from Eq. 3.

As described in detail in Section 2.2, the directly calculated pK_a values have fairly poor accuracy and thus we derive a simple linear estimator to correct for shortcomings in the solvation model [24]. The linear model is based on our own training data set (described in the next section) and the resulting estimator $\mathscr{L}$ is applied to the pK_a from Eq. 3 to obtain improved predictions for the SAMPL6 data set, $\mathrm{p}K_a = \mathscr{L}[\mathrm{p}K_a^{\mathrm{calc}}]$.

2.1 Data sets

The QM1 subset of the training set contains 21 neutral acids belonging to several chemical families (Fig. 2): mono- (**1**), di- (**2**) and tri- (**3**) protic inorganic acids, aliphatic (**4**) and aromatic (**5**) sulfonic acids, diversely substituted carboxylic acids (**6**-**11**) and alcohols (**12**-**17**), phenols (**18** and **19**), phthalimide (**20**) and uracil (**21**). The experimental pK_a values of these compounds range from −3.00 to 17.10 (Table 1). The QM2 subset contains 17 compounds that are cationic acids (Fig. 3): hydrazine (**22**), guanidine (**23**), aliphatic mono- (**24**), di- (**25**) and tri- (**26**) substituted amines, diversely substituted aromatic amines (**27**-**31**) and pyridines (**32**-**38**). These compounds possess experimental pK_a values from 0.49 to 13.60 (Table 1).

The SAMPL6 data set consisted of 24 fragment-like small organic molecules (Fig. 1) with unknown pK_a values that were selected for their similarity to kinase inhibitors and for experimental tractability. It was provided by the SAMPL6 organizers through the public repository `https://github.com/MobleyLab/SAMPL6` as computer-generated microstates in SMILES format. The protonation state for each microstate was computed with an in-house script using the CACTVS Chemoinformatics Toolkit [27] (Xemistry GmbH, `https://www.xemistry.com/`), allowing the classification of microstates in two groups, neutral acids and cationic acids, for which different correction factors were applied in the approach using the piecewise linear model.

Three-dimensional coordinates for all compounds were generated in MOL2 format using CORINA version 3.60 (`http://www.molecular-networks.com`), then converted into the Gaussian input format using an in-house script. The PDF3D files, which can be visualized with Adobe Acrobat Reader (`https://get.adobe.com/fr/reader/`) were generated with CACTVS.

2.2 Quantum chemical microscopic pK_a calculations

Gas-phase geometry optimization and frequency calculation of the protonated and deprotonated forms were performed at the B3LYP/6-311+G(d,p) level using Gaussian 09 version D.01 [28] to obtain $\Delta G^\circ_{(g)}$. A single-point free energy evaluation at the same level using the Conductor-like Polarizable Continuum Solvation Model (CPCM) [29–32] and UAHF radii as implemented in Gaussian 09 version D.01 [28] yielded the solvation free energy $\Delta G^\circ_{\mathrm{solv}}$ so that $\Delta G^*_{(s)}$ (Eq. 7) and an estimate for

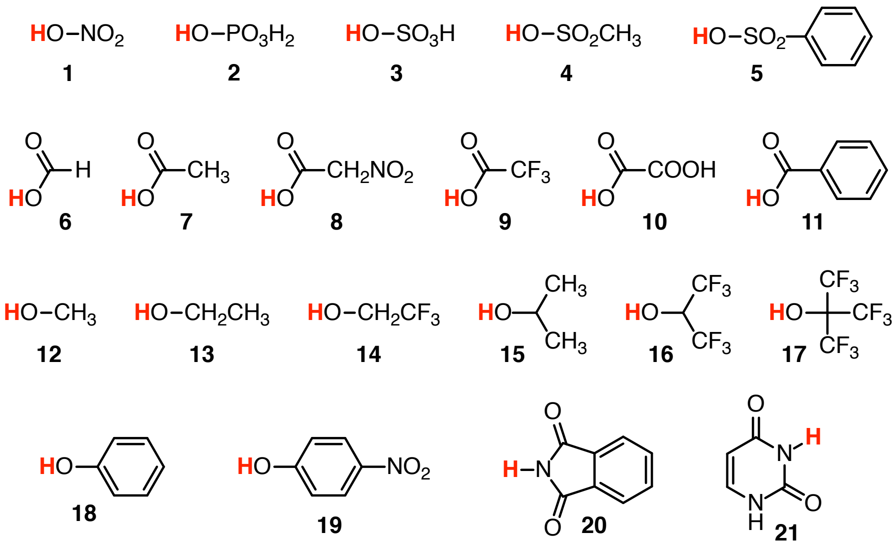


Fig. 2 Chemical structures of the QM1 training data set (neutral acids); see also Table 1.

the pK_a associated with this protonation/deprotonation event could be calculated via Eqs. 6 and 3.

In some cases, the geometry optimization did not converge with Gaussian 09 version D.01, but was successful with the version A.02 of Gaussian 09. Geometry optimization for microstates **SM04_micro016**, **SM07_micro016**, **SM17_micro008** and **SM17_micro009** did not converge in any conditions.

Muckerman et al [24] recognized systematic errors related to the solvation contribution ΔG^*_solv as responsible for poor accuracy, namely the solvation model undersolvates weak acids and over-solvates strong acids. They proposed a physically-motivated correction

$$\Delta G^*_\mathrm{corr}(\mathrm{HA}) := RT \ln 10 \cdot (\mathrm{p}K_\mathrm{a}^\mathrm{exp} - \mathrm{p}K_\mathrm{a}^\mathrm{calc}) \tag{8}$$

to $\Delta G^\circ_\mathrm{solv}$ with the linear model

$$\Delta G^*_\mathrm{corr} = a_0 + a_1 \cdot \mathrm{p}K_\mathrm{a}^\mathrm{exp}. \tag{9}$$

The parameters a_0 and a_1 are determined from a training set by linear regression. In order to apply the correction Eq. 9 to compounds with unknown pK_a, a linear estimator $\mathscr{L}$ can be derived by substituting $\mathrm{p}K_\mathrm{a}^\mathrm{exp} \approx \mathrm{p}K_\mathrm{a}^\mathrm{calc} + \Delta G^*_\mathrm{corr}/(RT \ln 10)$ in Eq. 9 and solving for ΔG^*_corr to yield

$$\Delta G^*_\mathrm{corr} = c_0 + c_1 \cdot \mathrm{p}K_\mathrm{a}^\mathrm{calc} \quad \text{with} \tag{10a}$$

$$c_0 = \frac{a_0}{1 - \lambda a_1} \tag{10b}$$

$$c_1 = \frac{a_1}{1 - \lambda a_1}, \quad \text{and} \quad \lambda := (RT \ln 10)^{-1} \tag{10c}$$

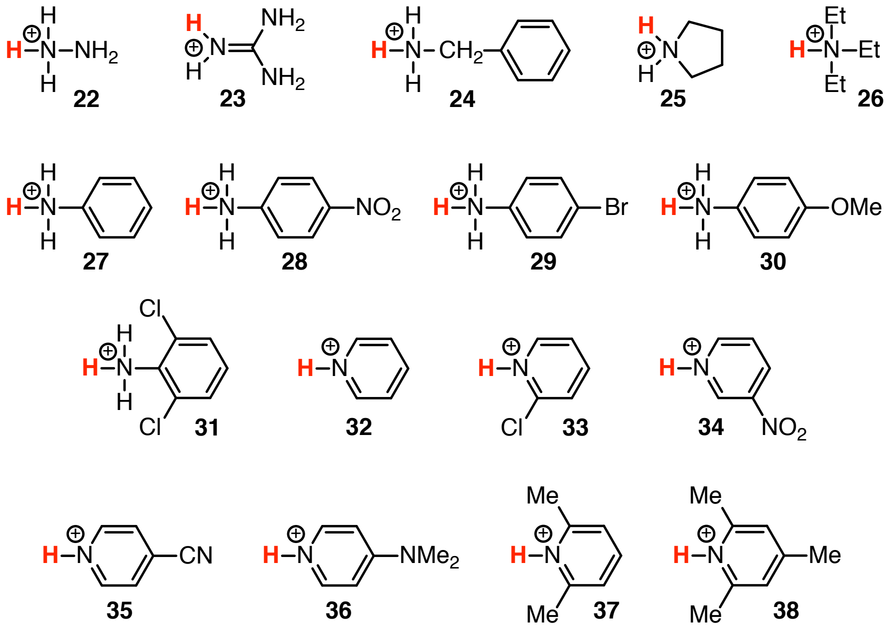


Fig. 3 Chemical structures of the QM2 training data set (cationic acids); see also Table 1.

The linear estimator $\mathscr{L}$ with parameters a_0 and a_1 for the microscopic pK_a is

$$\begin{aligned} \mathrm{p}K_\mathrm{a} = \mathscr{L}[\mathrm{p}K_\mathrm{a}^\mathrm{calc}] &= \mathrm{p}K_\mathrm{a}^\mathrm{calc} + \lambda \Delta G^*_\mathrm{corr} \\ &= \frac{\lambda a_0}{1-\lambda a_1} + \frac{1}{1-\lambda a_1} \cdot \mathrm{p}K_\mathrm{a}^\mathrm{calc}. \end{aligned} \tag{11}$$

2.3 Microstates *vs* Macrostates

We consider each tautomer of the acid HA and the base A$^-$ as a *microstate* with label i. The set of microstates with the same total number of protons $N_i = N$ is labeled the *macrostate* N. The macroscopic pK_a characterizes the transitions between any of the microstates with N protons to any microstate with $N-1$ protons.

In general, the free energy difference between two states (micro or macro states) that are separated by a single protonation process (i.e., the free energy to go from N to $N-1$ associated protons) is

$$\Delta G_{N,N-1} = -\Delta G_{N-1,N} = -\beta^{-1} \ln \left[\frac{P(N-1)}{P(N)} \right] \tag{12}$$

where $P(N-1)$ and $P(N)$ are the probabilities of observing the system with $N-1$ and N associated protons respectively and $\beta = (RT)^{-1}$. The Henderson-Hasselbalch equation

$$\mathrm{p}K_\mathrm{a} = \mathrm{pH} - \log_{10}\left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right) = \mathrm{pH} - \frac{1}{\ln 10} \ln \left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right) \tag{13}$$

can be rewritten in terms of the free energy of protonation $\Delta G_{N-1,N}$ (Eq. 12) to give

$$\mathrm{p}K_\mathrm{a} = \mathrm{pH} - \frac{\beta \Delta G_{N-1,N}}{\ln 10}, \tag{14a}$$

$$\Delta G_{N-1,N} = \beta^{-1} \ln 10 \cdot (\mathrm{pH} - \mathrm{p}K_\mathrm{a}). \tag{14b}$$

2.4 Calculation of macroscopic pK_as from microscopic pK_as

The microscopic pK_a values correspond to free energy differences $\Delta G_{ij}(\mathrm{pH}) = G_j(\mathrm{pH}) - G_i(\mathrm{pH})$ between microstates i and j (Eq. 14b); for notational convenience we drop the explicit pH dependence in the following for all free energies. Each state has a pH-dependent associated free energy G_i, which is not known. Constructing the G_i from the differences between them is not straightforward because these calculated free energy differences come with unknown errors that prevent, for example, that the sum along any closed thermodynamic cycle $i \to j \to k \to \cdots \to i$ is exactly zero as required by the fact that the G_i are thermodynamic state functions. We construct a set of M microstate free energies $\{G_i\}_{i=1}^M$ that is most consistent with the calculated ("measured") $\{\Delta G_{ij}\}$ using a maximum-likelihood estimator [33] based on the likelihood function

$$L(\{G_i\}|\{\Delta G_{ij}\}) = \prod_{ij} \exp\left(-\tfrac{1}{2}[(G_j - G_i) - \Delta G_{ij}]^2\right), \tag{15}$$

where we assumed normal distribution of errors with constant standard deviation. The product runs over all pairs (i, j) for which calculated ΔG_{ij} are available. L is proportional to the probability $P(\{\Delta G_{ij}\}|\{G_i\})$ that we could observe the measured data (all the calculated ΔG_{ij}) if we were given a specific set of the G_i (our model parameters). Maximizing the log-likelihood $\ln L$ (using functions in SciPy [34]) as a function of all the G_i provides the set $\{G_i\}_{i=1}^M$ that is most consistent with the given measurements $\{\Delta G_{ij}\}$. Further details and more general applications of this approach will be published elsewhere (I.M. Kenney *et al*, in preparation).

In order to calculate the macroscopic pK_as, we begin by calculating the free energy of protonation using principles of equilibrium statistical mechanics [25]. The probability of observing a macrostate with N associated protons is

$$P(N) = Z^{-1} \sum_i e^{-\beta G_i} \delta_{N_i,N} \tag{16}$$

where the sum is over all accessible microstates with free energy G_i, $\delta_{N_i,N}$ is unity when the microstate i has N protons and null otherwise, and Z is the partition function, defined by

$$Z = \sum_j e^{-\beta G_j}. \tag{17}$$

Eq. 16 combined with the general expression for the free energy of protonation (Eq. 12) yields the effective macroscopic protonation free energy as a function of the G_i,

$$\Delta G_{N-1,N} = \beta^{-1} \ln \left[\frac{\sum_i e^{-\beta G_i} \delta_{N_i,N-1}}{\sum_i e^{-\beta G_i} \delta_{N_i,N}} \right]. \tag{18}$$

$\Delta G_{N-1,N}$ is a function of the pH of the system and the microscopic pK_as relevant to the macrostate N. Together with Eq. 14a, Eq. 18 allows us to calculate the macroscopic pK_a value for removing the N^th proton from a molecule. With all microstate free energies $\{G_i\}_{i=1}^{M}$ known for a given pH value it is also straightforward to compute the pH-dependent microstate probabilities

$$p_i(\mathrm{pH}) = Z(\mathrm{pH})^{-1} e^{-\beta G_i(\mathrm{pH})} \tag{19}$$

where all terms depend on pH.

2.5 Error analysis

The difference between experimental and computed pK_a values ("signed error") for each compound, labeled with its identification code 'id', was calculated as

$$\Delta_\mathrm{id} = \mathrm{pK}_\mathrm{a,id} - \mathrm{pK}^\mathrm{exp}_\mathrm{a,id}. \tag{20}$$

The root-mean-square error (RMSE) was determined from the individual errors Δ as

$$\mathrm{RMSE} = \sqrt{\langle \Delta^2 \rangle} = \sqrt{N^{-1} \sum_\mathrm{id}^{N} \Delta_\mathrm{id}^2}, \tag{21}$$

the mean absolute error (MAE) as

$$\mathrm{MAE} = \langle |\Delta| \rangle = N^{-1} \sum_\mathrm{id}^{N} |\Delta_\mathrm{id}|, \tag{22}$$

and the signed mean error (ME, also called the "mean signed error", MSE) as

$$\mathrm{ME} = \langle \Delta \rangle = N^{-1} \sum_\mathrm{id}^{N} \Delta_\mathrm{id}. \tag{23}$$

We also report the Pearson correlation coefficient R^2 and the slope m of a linear regression to the data, as computed with the function `scipy.stats.linregress()` in the SciPy package [34].

The quantum chemical single point free energy calculations do not have a statistical error and we have not yet implemented the calculation of an error bound in the maximum likelihood estimator for the G_i. Therefore, all pK_a are provided without a statistical error. Judging from the performance of the training data set and the post-hoc analysis of the SAMPL6 compounds (see Results), the accuracy of the calculated pK_a values is 1–2 pK_a units.

Calculated pK_a were compared to experimental values with the script `typeIII_analysis.py` as provided by the SAMPL6 organizers in the public repository `https://github.com/MobleyLab/SAMPL6`. Calculated values were matched to experimental ones with the *Hungarian algorithm*, which finds the optimum pairing between two sets by minimizing the linear sum of squared errors.

3 Results and Discussion

3.1 Training data set

The first step in our protocol was the design of a training data set containing 38 structurally-diverse, simple organic and inorganic compounds with known pK_a values. This global data set could be classified by the charge of the acid and split into two subsets. The neutral acids (named *QM1*, Fig. 2) contained 21 compounds and the second set, the positively-charged acids (named *QM2*, Fig. 3), contained the remaining 17 compounds. The structures were chosen from different chemical families in order to obtain for the two subsets a relatively homogeneous distribution of data points over a wide range of values (see Table 1 for the experimental pK_as).

Predicted pK_a values were computed for all compounds from the training data set using the protocol described by Muckerman et al [24] (see the Methods section for details). The correlation of these computed values with the experimental pK_as is shown in (Fig. 4a), with a Pearson correlation coefficient $R^2 = 0.96$ (Table 1). The corresponding $\Delta G^{*}_{\mathrm{corr}}$ values were obtained using Eq. 8 and plotted against the experimental pK_a values. A global linear fit model, with a slope of $a_1 = -0.61$ and an intercept of $a_0 = 2.75$ (parameters in Eq. 9), was derived by using all compounds as a single data set (Fig. 4b). Alternatively, a piecewise linear fit model was derived by considering separately the two QM1 and QM2 subsets (Fig. 4c). In this latter case we obtained the parameters in Eq. 9 with a slope of $a_1^{\mathrm{QM1}} = a_1^{\mathrm{QM2}} = -0.62$ for both subsets and intercept values of $a_0^{\mathrm{QM1}} = 1.30$ and $a_0^{\mathrm{QM2}} = 4.65$ for the QM1 and QM2 subsets, respectively.

The linear estimators associated with these models (Eq. 10a) were calculated using Eq. 11. These corrections were applied to the whole training set, and to the QM1 and QM2 subsets, respectively, in order to evaluate to which extent the systematic errors related to the prediction method were removed compared with the pK_a values obtained directly from the *ab initio* calculations (Table 1). We can see that in all cases the corrected pK_a values are much closer to the experimental values, with the global model behaving slightly better than the piecewise model, as shown by, for instance, the smaller RMSE 1.66 *vs* 1.85 for the whole training set.

3.2 Macroscopic pK_a

The microscopic pK_a values for the SAMPL6 data set were computed using the same protocol as for the training data set (595 individual transformations). Again, the corrections from the global linear model were applied to the whole SAMPL6 data set and alternatively, those from the piecewise linear model to individual subsets of the SAMPL6 data set containing the neutral acids and the cationic acids, respectively.

Starting from these three sets of results (obtained directly from *ab initio* free energies or after correction with the two linear models, global and piecewise) we calculated pH-dependent microstate free energies and macroscopic pK_a values (Table 2). These results, formatted using the SAMPL6 submission template, were used as input for the `typeIII_analysis.py` script in order to compare to the experimental values

Table 1 Experimental and computed pK_a values for the compounds from the QM1 (Fig. 2) and QM2 (Fig. 3) training data sets. The difference Δ (Eq. 20) between computed and experimental pK_a values is shown for each compound. The experimental values were taken from Muckerman et al [24] and from Lundblad and Macdonald [35]. The root-mean-square error (RMSE), the mean absolute error (MAE), and the signed mean error (ME) were calculated according to Eqs. 21–23.

id	Exp.	QM computed		$\Delta G^*_{correction}$	Linear fit global		Linear fit piecewise	
	pK_a	pK_a	Δ		pK_a	Δ	pK_a	Δ
1	−1.40	−4.01	−2.61	3.56	0.43	1.83	−0.31	1.09
2	2.15	1.75	−0.40	0.54	2.88	0.73	2.14	−0.01
3	−3.00	−10.32	−7.32	9.98	−0.68	2.32	−1.41	1.59
4	−1.90	−2.92	−1.02	1.39	0.08	1.98	−0.65	1.25
5	−2.80	−3.44	−0.64	0.88	−0.54	2.26	−1.27	1.53
6	3.77	5.16	1.39	−1.89	4.00	0.23	3.26	−0.51
7	4.76	7.27	2.51	−3.42	4.69	−0.07	3.94	−0.82
8	1.68	4.73	3.05	−4.16	2.56	0.88	1.82	0.14
9	0.23	−1.82	−2.05	2.80	1.55	1.32	0.82	0.59
10	1.38	4.06	2.68	−3.65	2.35	0.97	1.61	0.23
11	4.21	6.77	2.56	−3.49	4.31	0.10	3.56	−0.65
12	15.54	23.37	7.83	−10.68	12.14	−3.40	11.37	−4.17
13	15.90	22.73	6.83	−9.31	12.39	−3.51	11.62	−4.28
14	12.43	14.22	1.79	−2.44	9.99	−2.44	9.22	−3.21
15	17.10	22.61	5.51	−7.51	13.22	−3.88	12.44	−4.66
16	9.30	8.57	−0.73	0.99	7.83	−1.47	7.07	−2.23
17	5.40	4.28	−1.12	1.53	5.13	−0.27	4.38	−1.02
18	9.95	14.99	5.04	−6.87	8.28	−1.67	7.51	−2.44
19	7.14	8.26	1.12	−1.53	6.33	−0.81	5.58	−1.56
20	8.30	13.49	5.19	−7.08	7.14	−1.16	6.38	−1.92
21	9.50	13.81	4.31	−5.87	7.97	−1.53	7.2	−2.30
RMSE (QM1)			3.86			1.90		2.19
MAE (QM1)			3.13			1.56		1.72
ME (QM1)			1.61			−0.36		−1.11
R^2 (QM1)		0.97						
m (QM1)		1.45						
22	8.12	9.05	0.93	−1.26	7.01	−1.11	7.93	−0.19
23	13.60	18.94	5.34	−7.28	10.8	−2.80	11.7	−1.90
24	9.30	8.48	−0.82	1.12	7.83	−1.47	8.75	−0.55
25	11.27	8.18	−3.09	4.22	9.19	−2.08	10.1	−1.17
26	10.72	12.64	1.92	−2.61	8.81	−1.91	9.72	−1.00
27	4.62	2.91	−1.71	2.33	4.59	−0.03	5.53	0.91
28	0.98	−4.10	−5.08	6.93	2.07	1.09	3.02	2.04
29	3.89	1.50	−2.39	3.26	4.09	0.20	5.02	1.13
30	5.36	5.35	−0.01	0.01	5.1	−0.26	6.04	0.68
31	1.53	−3.20	−4.73	6.44	2.45	0.92	3.4	1.87
32	5.24	4.67	−0.57	0.78	5.02	−0.22	5.95	0.71
33	0.49	−0.96	−1.45	1.98	1.73	1.24	2.69	2.20
34	0.81	−1.37	−2.18	2.98	1.96	1.15	2.91	2.10
35	1.86	0.14	−1.72	2.35	2.68	0.82	3.63	1.77
36	9.60	11.07	1.47	−2.00	8.04	−1.56	8.95	−0.65
37	6.70	7.89	1.19	−1.63	6.03	−0.67	6.96	0.26
38	7.33	8.22	0.89	−1.21	6.47	−0.86	7.39	0.06
RMSE (QM2)			2.60			1.30		1.33
MAE (QM2)			2.09			1.08		1.13
ME (QM2)			−0.71			−0.44		0.49
R^2 (QM2)		0.96						
m (QM2)		1.45						
RMSE (Global)			3.35			1.66		1.85
MAE (Global)			2.66			1.35		1.46
ME (Global)			0.58			−0.40		−0.40
R^2 (Global)		0.96						
m (Global)		1.44						

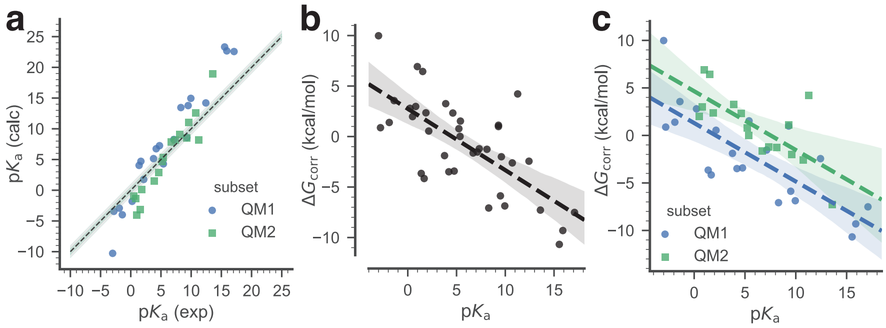


Fig. 4 Training data set. The pK_a of the training data set compounds are used to derive a simple linear model that relates the free energy correction ΔG^*_{corr} to the experimental pK_a. Two linear models were derived: a *global* linear model (black dashed line), utilizing all data, and a *piecewise* linear model that applies to either neutral acids (subset QM1, blue) or to positively charged acids (subset QM2, green). **a**: Correlation between experimental and calculated pK_a of the training data set. The dashed line indicates ideal correlation with the gray band indicating 1 pK_a unit deviation. **b**: Global linear fit of the calculated ΔG^*_{corr} to the experimental pK_a. **c**: Linear fits of the calculated ΔG^*_{corr} to the experimental pK_a, split between the QM1 and the QM2 subsets. In (b) and (c) the dashed lines are linear models to the data, with shaded bands indicating 95% confidence intervals from 1000 bootstrap samples.

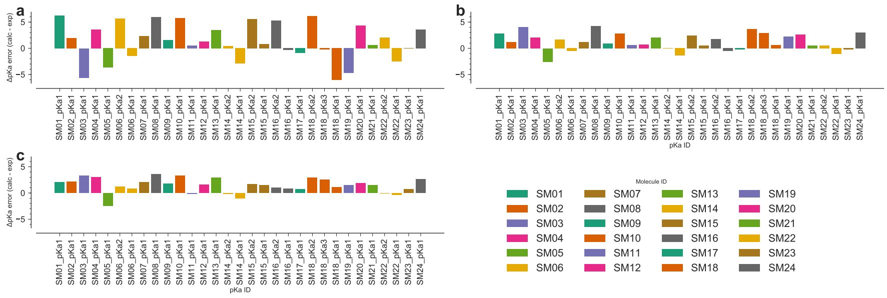


Fig. 5 Signed error Δ_{id} of individual predictions. The calculated pK_a was matched to the experimental pK_a for each compound (indicated by the SAMPL6 pKa ID) and the deviation from the experimental value represented as a bar. Observations for the same compound have the same color. **a**: pK_a were directly estimated from the quantum mechanical free energy differences. **b**: The quantum mechanical pK_a were corrected with the global linear model. **c**: compounds were corrected depending on their membership in subsets 1 or 2 with the piecewise linear model.

that were provided by the SAMPL6 organizers together with the analysis scripts. The input files with our results formatted as comma-separated value (CSV) files and the optimized structures for all microstates in MOL2 and PDF3D format are provided in the Electronic Supplementary Material. During the challenge we submitted macroscopic pK_a values only for three compounds (**SM15**, **SM20** and **SM22**). Here we describe the macroscopic pK_a predictions for the entire SAMPL6 data set.

Using this protocol we could predict the macroscopic pK_a values for the 24 SAMPL6 compounds with a RMSE of about 2 pK_a units when the corrections were applied and of 3.74 pK_a units when the *ab initio* free energies were used directly. The

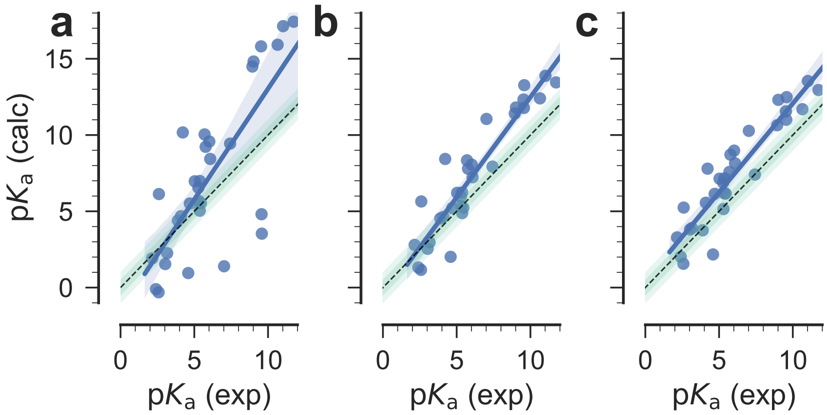


Fig. 6 Correlation between experimental and calculated pK_a values for the SAMPL6 compounds. **a**: pK_a were directly estimated from the quantum mechanical free energy differences. **b**: The quantum mechanical pK_a were corrected with the global linear model. **c**: compounds were corrected depending on their membership in subsets 1 or 2. The black dashed line indicates ideal correlation, the shaded green bars show 0.5 and 1 pK_a units deviation from ideal. Blue lines are linear regression fits to the data, with the blue shaded area indicating the 95% confidence interval from 1000 bootstrap samples.

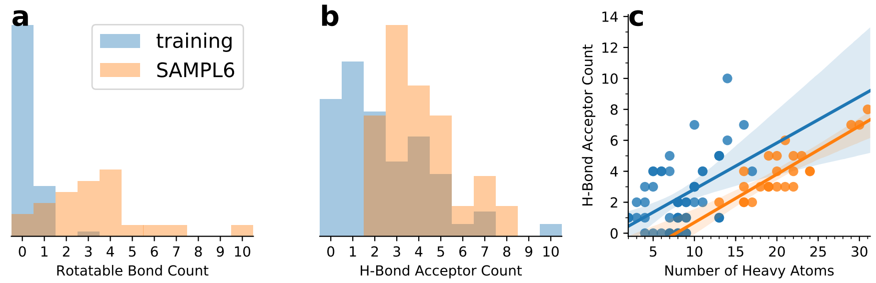


Fig. 7 Comparison of chemical properties of the training (light blue) and SAMPL6 (orange) data sets. **a**: normalized histograms of the number of rotatable bonds; **b**: normalized histograms of the number of hydrogen bond acceptors; **c**: correlation between the number of heavy atoms and the number of acceptors with linear regressions shown as solid lines and their 95% confidence interval from 1000 bootstraps indicated by shaded areas.

relative poor accuracy when directly using the quantum chemical free energies is in line with previous studies [15, 24].

The signed errors of individual predictions represented in Fig. 5 show that most of the prediction errors after correction are positive, with the notable exception of compound **SM05** for which these errors are consistently negative. High prediction errors ($3-4$ pK_a units) are obtained for compounds **SM03** and **SM08**, whereas compounds **SM01**, **SM04**, **SM10**, **SM13**, **SM18**, **SM20**, and **SM24** are predicted with errors of about $2-3$ pK_a units. The representation of the prediction errors in the order of increasing absolute experimental pKa values (Fig. S3, Electronic Supplementary

Material) shows that these are not related. Therefore, the source of remaining errors after correction should be sought elsewhere. As shown in Fig. 6, the results for the SAMPL6 data set are fairly insensitive to the fitting approach used (global or piecewise linear model), further indicating some level of robustness. Other statistical measures such as Pearson correlation coefficient $R^2 = 0.86$ and the slope of the linear regression $m = 1.17$ (for the piecewise linear model, see Table 2 for almost identical values for the global linear model) indicate encouraging correlations but the large mean error (1.42 for the piecewise linear model and 1.24 for the global linear model) hint at remaining systematic errors.

The fact that the linear fit did not remove these systematic errors implies that the training data set did not include properties that are important for the SAMPL6 data set and hence the linear or piecewise linear estimator cannot correct model errors related to these properties. In order to quantify similarities and differences between the two datasets we analyzed a number of chemical properties (see section *Properties of the training and SAMPL6 data sets* with Fig. S1 in the Electronic Supplementary Material file for details). Overall, the most obvious differences between our training and the SAMPL6 data set are the higher flexibility of the SAMPL6 molecules (with a median three and maximum ten rotatable bonds versus a median zero and maximum three, Fig. 7a) and the greater capability to accept hydrogen bonds (median four and maximum eight hydrogen bond acceptors versus median two and maximum ten; Fig. 7b), which correlates with a larger polar surface area (see Fig. S2 in the Electronic Supplementary Material file). However, Fig. 7c shows that the training compounds have *more* hydrogen bond acceptors for the same number of heavy atoms than the SAMPL6 compounds, i.e., for their larger size, the SAMPL6 compounds have fewer acceptors than one would expect from simple extrapolation of the training compounds. Similarly, the polar surface area of the SAMPL6 compounds would be overestimated from the training set (Fig. S2). These differences suggest that the interactions with water through hydrogen bonds are stronger in the training set than in the SAMPL6 set, which could lead to a systematic error in the estimator that was derived from the training set.

In the post-challenge analysis, we also tested the introduction of a conformational search step in our protocol and evaluated its influence on the quality of our predictions using two model compounds, **SM06** and **SM20**. The complete results are presented in the *Conformational search* section of the Electronic Supplementary Material file. In brief, for **SM06** the new microscopic pK_a value of **SM06_micro011** brought no changes in the predicted macroscopic pK_a values and for **SM20** we obtained macroscopic pK_a prediction errors 1.8-2.4 pK_a units higher compared with the values obtained without conformational search. It seems that, at least for these two compounds, the conformational search does not yield any substantial improvements in the prediction of macroscopic pK_a values.

3.3 Microstate probabilities

The SAMPL6 organizers recently made available experimental assignments of microstates with corresponding microstate pK_a for a number of compounds [36] (`https:`

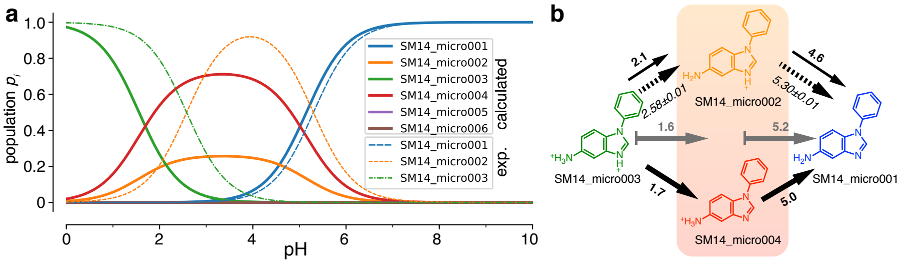


Fig. 8 Microstate probabilities p_i for **SM14**. **a**: Computed microstate probabilities (for the piecewise linear fit) are shown as heavy solid lines and experimentally derived probabilities as thin dashed lines. The experimental p_i were calculated in the same way as the calculated ones (Eq. 19) by directly using the experimental microstate pK_as. **b**: Microstate diagram with arrows indicating deprotonation. Bold numbers near solid arrows are the calculated microstate pK_a (from (a)) and italic numbers near dashed arrows are the experimental numbers, assigned to the experimentally identified microstate transitions. The gray solid arrows with gray bold numbers indicate the calculated macroscopic pK_a from $N = 3$ protons (microstate **SM14_micro003**) to $N = 2$ protons (mixture of **SM14_micro002** and **SM14_micro004**, indicated by the orange box) to $N = 1$ proton in **SM14_micro001** (and **SM14_micro005**, which is not shown because computation and experiment indicate that it is suppressed relative to **SM14_micro001**).

//github.com/MobleyLab/SAMPL6/blob/master/physical_properties/pKa/experimental_data/NMR_microstate_determination/). Here we focus on **SM14** as an example. Fig. 8 compares our computed microstate probabilities p_i (Eq. 19) to the ones derived from the experimental assignments of states **SM14_micro003**, **SM14_micro002**, and **SM14_micro001**. The important calculated microstates (from the linear piecewise model) were **SM14_micro003** ($N = 3$ protons), **SM14_micro004** and **SM14_micro002**, both with $N = 2$ protons, and **SM14_micro001** ($N = 1$). The calculated microscopic pK_a for the deprotonation of **SM14_micro003** to **SM14_micro002** was 2.1, similar to the experimental value 2.58 ± 0.01. The microscopic pK_a corresponding to the deprotonation of **SM14_micro002** to **SM14_micro001** was calculated as 4.6, also similar to the experimental one, 5.30 ± 0.01. A second microstate **SM14_micro005** exists with the same number of protons as **SM14_micro002** but both experiment and our computations indicated that this second state is suppressed and plays no role. Our calculations, however, assigned a higher population to **SM14_micro004** than to **SM14_micro002**, in contrast to the experimental findings, which, based on NMR nitrogen chemical shift measurements in the aprotic solvent acetonitrile-d$_3$ under pH titration, identified **SM14_micro002** as the dominant intermediate state. The partial agreement between these detailed experiments and our calculations is encouraging but a single comparison does not allow us to draw any broader conclusions except perhaps to highlight the ease with which our partition function-based formalism can be used to compute microscopic populations.

3.4 Computation time

The total computational cost required by this project was 641 CPU-days on a Linux cluster making use of Intel Xeon E5-4627 v3 CPUs running at 2.60 GHz. Given that

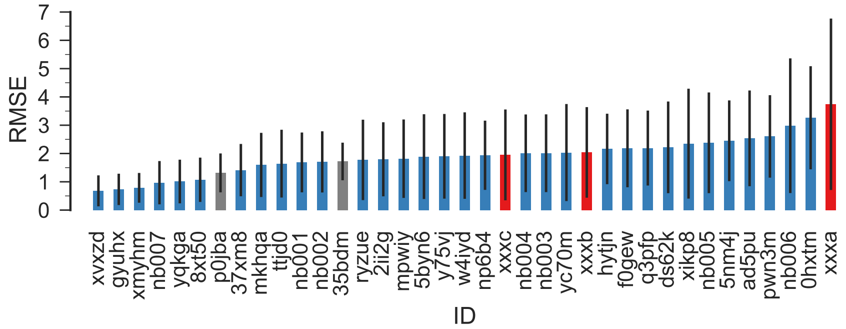


Fig. 9 RMSE of all SAMPL6 submissions (blue), including our new calculations for all SAMPL6 compounds (red) and for completeness our original submissions (gray), which only included predictions for **SM15**, **SM20**, and **SM22** and is only of limited statistical validity because of the large variance of the RMSE itself for only three samples [37]. The submission IDs *p0jba* and *xxxc* correspond to the piecewise linear model, *35bdm* and *xxxb* to the global linear model, and *xxxa* to directly using the quantum chemical free energies. Other IDs belong to other regular SAMPL6 submissions. The error bars indicate 95% confidence intervals from 1000 bootstrap samples.

344 microstates were computed, each microstate required 1.86 CPU-days on average. The calculations were carried out in parallel on 8 cores, so the average wall clock time for a microstate was 5.6 hours in these conditions. The most rigid compound, **SM22**, was the fastest with 1 CPU-hour for one microstate, whereas one of the biggest and most flexible compounds from the SAMPL6 data set, **SM18**, required about 3.2 CPU-days for one microstate.

4 Conclusions

Compared to other methods in the SAMPL6 challenge, our approach has below-average accuracy (Fig. 9 and Figs. S4–S7 in the Electronic Supplementary Material) and its computational cost is also higher than ML-based approaches (not considering the cost for compiling and validating the data and training the ML model). A key advantage of our approach is its generality as it does not depend on training on specific data sets although below we note that the quality of the training set for the correction step is a possible concern. With the linear model, which was derived from a very small and simple training set (38 compounds), we remove some of the errors related to the QM method used and its implementation in Gaussian (e.g., the implicit solvation model). The quality of the prediction is mostly independent of the structure, i.e., it can predict organic compounds from different families and even inorganic compounds with similar level of accuracy. In comparison, purely ML-based methods are trained on large experimental data sets (containing several thousands or tens of thousands compounds) and they can be vulnerable to chemical families that are not represented in the training set. Our approach appears reasonably robust because for our training set we obtain the same slope on the global data set and on the

Table 2 Experimental and computed pK_a values for the compounds from the SAMPL6 data set (Fig. 1). The difference Δ (Eq. 20) between computed and experimental pK_a values is shown for each compound. The experimental values were provided by the SAMPL6 organizers. The root-mean-square error (RMSE), the mean absolute error (MAE), and the signed mean error (ME) were calculated according to Eqs. 21–23. The Pearson correlation coefficient R^2 and the slope m were calculated from a linear regression.

Compound	pK_a	Exp.	QM computed		Linear fit global		Linear fit piecewise	
ID	ID	pK_a	pK_a	Δ	pK_a	Δ	pK_a	Δ
SM01	pKa1	9.53(1)	15.81	6.28	12.33	2.80	11.55	2.02
SM02	pKa1	5.03(1)	6.97	1.94	6.21	1.18	7.14	2.11
SM03	pKa1	7.02(1)	1.40	−5.62	11.06	4.04	10.27	3.25
SM04	pKa1	6.02(1)	9.58	3.56	8.06	2.04	8.98	2.96
SM05	pKa1	4.59(1)	0.95	−3.64	2.02	−2.57	2.17	−2.42
SM06	pKa1	3.03(4)	1.54	−1.49	2.54	−0.49	3.81	0.78
SM06	pKa2	11.74(1)	17.43	5.69	13.45	1.71	12.95	1.21
SM07	pKa1	6.08(1)	8.44	2.36	7.23	1.15	8.15	2.07
SM08	pKa1	4.22(1)	10.17	5.95	8.43	4.21	7.80	3.58
SM09	pKa1	5.37(1)	6.99	1.62	6.23	0.86	7.16	1.79
SM10	pKa1	9.02(1)	14.82	5.80	11.81	2.79	12.31	3.29
SM11	pKa1	3.89(1)	4.39	0.50	4.53	0.64	3.75	−0.14
SM12	pKa1	5.28(1)	6.55	1.27	5.96	0.68	6.89	1.61
SM13	pKa1	5.77(1)	9.23	3.46	7.79	2.02	8.72	2.95
SM14	pKa1	2.58(1)	−0.31	−2.89	1.16	−1.42	1.56	−1.02
SM14	pKa2	5.30(1)	5.68	0.38	5.34	0.04	5.15	−0.15
SM15	pKa1	4.70(1)	5.51	0.81	5.21[a]	0.51	6.14[b]	1.44
SM15	pKa2	8.94(1)	14.49	5.55	11.41[a]	2.47	10.64[b]	1.70
SM16	pKa1	5.37(1)	5.04	−0.33	4.88	−0.49	6.17	0.80
SM16	pKa2	10.65(1)	15.92	5.27	12.40	1.75	11.69	1.04
SM17	pKa1	3.16(1)	2.26	−0.90	2.96	−0.20	3.90	0.74
SM18	pKa1	2.15(2)	1.91	−0.24	2.80	0.65	3.29	1.14
SM18	pKa2	9.58(3)	3.54	−6.04	13.27	3.69	12.49	2.91
SM18	pKa3	11.02(4)	17.14	6.12	13.88	2.86	13.54	2.52
SM19	pKa1	9.56(2)	4.81	−4.75	11.78	2.22	11.00	1.44
SM20	pKa1	5.70(3)	10.04	4.34	8.34[a]	2.64	7.58[b]	1.88
SM21	pKa1	4.10(1)	4.68	0.58	4.63	0.53	5.56	1.46
SM22	pKa1	2.40(2)	−0.10	−2.50	1.32[a]	−1.08	2.02[b]	−0.38
SM22	pKa2	7.43(1)	9.44	2.01	7.93[a]	0.50	7.41[b]	−0.02
SM23	pKa1	5.45(1)	5.53	0.08	5.23	−0.22	6.16	0.71
SM24	pKa1	2.60(1)	6.13	3.53	5.65	3.05	5.25	2.65
RMSE				3.74		2.04		1.95
MAE				3.08		1.66		1.68
ME				1.25		1.24		1.42
R^2			0.58		0.87		0.86	
m			1.45		1.31		1.17	

[a] These results represent our submission *35bdm* to SAMPL6.

[b] These results represent our submission *p0jba* to SAMPL6.

individual subsets, which are chemically quite different. The results for the SAMPL6 data set are also fairly insensitive to the fitting approach used (global or piecewise linear model), further demonstrating robustness. The correlations with experimental data are generally good but suffer from systematic errors, possibly from differences between the training set and the SAMPL6 set that bias the estimator that is needed to correct the raw QM pK_a values. The statistical measures indicate clear room from improvement. It appears that a better correction scheme, using a larger data set that better matches the test data set with respect to its hydrogen bonding properties and is generally more representative of drug-like molecules could improve the predictions, perhaps in conjunction with more sophisticated classifiers and estimators than simple separation by charge and linear regression. We expect that improvements in the model physics, namely in the treatment of solvation, could also lead to further increases in accuracy.

We currently consider the method described here (and originally developed by Muckerman et al [24]) as an acceptable compromise between speed, accuracy and generality across the chemical space. It seems especially useful when one encounters novel compounds and wants to assess them based on their absolute pK_a values. The calculations are tractable with typical computational resources, absolute pK_as are accurate to about 2 units (within the “chemical accuracy” range [15]) and do not seem to be biased with respect to specific chemical groups, and thus the relative ordering of compounds is also meaningful.

References

1. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, Pande VS (2008) Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. J Med Chem 51(4):769–779, DOI 10.1021/jm070549+
2. Guthrie JP (2009) A blind challenge for computational solvation free energies: Introduction and overview. J Phys Chem B 113(14):4501–4507, DOI 10.1021/jp806724u
3. Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ (2010) The SAMPL2 blind prediction challenge: Introduction and overview. J Comput Aided Mol Des 24(4):259–279, DOI 10.1007/s10822-010-9350-8
4. Geballe MT, Guthrie JP (2012) The SAMPL3 blind prediction challenge: transfer energy overview. J Comput Aided Mol Des 26(5):489–96, DOI 10.1007/s10822-012-9568-8
5. Mobley DL, Wymer KL, Lim NM, Guthrie JP (2014) Blind prediction of solvation free energies from the SAMPL4 challenge. J Comput Aided Mol Des 28(3):135–50, DOI 10.1007/s10822-014-9718-2
6. Bannan CC, Calabró G, Kyu DY, Mobley DL (2016) Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. J Chem Theory Comput 12(8):4015–24, DOI 10.1021/acs.jctc.6b00449
7. Beckstein O, Iorga BI (2012) Prediction of hydration free energies for aliphatic and aromatic chloro derivatives using molecular dynamics simulations with

the OPLS-AA force field. J Comput Aided Mol Des 26(5):635–645, DOI 10.1007/s10822-011-9527-9

8. Beckstein O, Fourrier A, Iorga BI (2014) Prediction of hydration free energies for the SAMPL4 diverse set of compounds using molecular dynamics simulations with the OPLS-AA force field. J Comput Aided Mol Des 28(3):265–276, DOI 10.1007/s10822-014-9727-1
9. Kenney IM, Beckstein O, Iorga BI (2016) Prediction of cyclohexane-water distribution coefficients for the SAMPL5 data set using molecular dynamics simulations with the OPLS-AA force field. J Comput Aided Mol Des 30(11):1045–1058, DOI 10.1007/s10822-016-9949-5
10. Babić S, Horvat AJ, Mutavdžić Pavlović D, Kaštelan-Macan M (2007) Determination of pK$_a$ values of active pharmaceutical ingredients. TrAC Trends in Analytical Chemistry 26(11):1043–1061, DOI https://doi.org/10.1016/j.trac.2007.09.004
11. Lee AC, Crippen GM (2009) Predicting pK_a. Journal of Chemical Information and Modeling 49(9):2013–2033, DOI 10.1021/ci900209w
12. Alexov E, Mehler EL, Baker N, Baptista AM, Huang Y, Milletti F, Nielsen JE, Farrell D, Carstensen T, Olsson MHM, Shen JK, Warwicker J, Williams S, Word JM (2011) Progress in the prediction of pK_a values in proteins. Proteins 79(12):3260–75, DOI 10.1002/prot.23189
13. Rupp M, Korner R, Tetko IV (2011) Predicting the pK_a of small molecules. Combinatorial Chemistry & High Throughput Screening 14(5):307–327, DOI 10.2174/138620711795508403
14. Reijenga J, van Hoof A, van Loon A, Teunissen B (2013) Development of methods for the determination of pK_a values. Anal Chem Insights 8:53–71, DOI 10.4137/ACI.S12304
15. Ho J, Coote ML (2009) A universal approach for continuum solvent pK_a calculations: are we there yet? Theoretical Chemistry Accounts 125(1-2):3–21, DOI 10.1007/s00214-009-0667-0
16. Mongan J, Case DA, McCammon JA (2004) Constant pH molecular dynamics in generalized born implicit solvent. J Comput Chem 25(16):2038–48, DOI 10.1002/jcc.20139
17. Chen W, Morrow BH, Shi C, Shen JK (2014) Recent development and application of constant pH molecular dynamics. Molecular Simulation 40(10-11):830–838, DOI 10.1080/08927022.2014.907492
18. Swails JM, York DM, Roitberg AE (2014) Constant pH replica exchange molecular dynamics in explicit solvent using discrete protonation states: Implementation, testing, and validation. J Chem Theory Comput 10(3):1341–1352, DOI 10.1021/ct401042b
19. Radak BK, Chipot C, Suh D, Jo S, Jiang W, Phillips JC, Schulten K, Roux B (2017) Constant-pH molecular dynamics simulations for large biomolecular systems. Journal of Chemical Theory and Computation 13(12):5933–5944, DOI 10.1021/acs.jctc.7b00875
20. Di Russo NV, Estrin DA, Martí MA, Roitberg AE (2012) pH-dependent conformational changes in proteins and their effect on experimental pK(a)s: the case of nitrophorin 4. PLoS Comput Biol 8(11):e1002,761, DOI 10.1371/jour-

nal.pcbi.1002761

21. Morrow BH, Koenig PH, Shen JK (2013) Self-assembly and bilayer-micelle transition of fatty acids studied by replica-exchange constant pH molecular dynamics. Langmuir 29(48):14,823–30, DOI 10.1021/la403398n
22. Huang Y, Chen W, Dotson DL, Beckstein O, Shen J (2016) Mechanism of pH-dependent activation of the sodium-proton antiporter NhaA. Nature Communications 7:12,940, DOI 10.1038/ncomms12940
23. Alongi KS, Shields GC (2010) Theoretical calculations of acid dissociation constants: A review article. In: Ann Rep Comput Chem, vol 6, Elsevier Science B.V., chap 8, pp 113–138, DOI 10.1016/S1574-1400(10)06008-1
24. Muckerman JT, Skone JH, Ning M, Wasada-Tsutsui Y (2013) Toward the accurate calculation of pK_a values in water and acetonitrile. Biochimica et Biophysica Acta 1827:882–891, DOI 10.1016/j.bbabio.2013.03.011
25. McQuarrie DA (1976) Statistical Mechanics. HarperCollins, New York
26. Zhang H, Jiang Y, Yan H, Yin C, Tan T, van der Spoel D (2017) Free-energy calculations of ionic hydration consistent with the experimental hydration free energy of the proton. The Journal of Physical Chemistry Letters 8(12):2705–2712, DOI 10.1021/acs.jpclett.7b01125, pMID: 28561580
27. Ihlenfeldt W, Takahashi Y, Abe H, Sasaki S (1994) Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility. J Chem Inf Comput Sci 34(1):109–116 (http://www.xemistry.com/)
28. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA Jr, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas O, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT
29. Klamt A, Schüürmann G (1993) COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J Chem Soc, Perkin Trans 2 pp 799–805, DOI 10.1039/P29930000799
30. Andzelm J, Külmel C, Klamt A (1995) Incorporation of solvent effects into density functional calculations of molecular energies and geometries. The Journal of Chemical Physics 103(21):9312–9320, DOI 10.1063/1.469990
31. Barone V, Cossi M (1998) Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. The Journal of Physical Chemistry A 102(11):1995–2001, DOI 10.1021/jp9716997
32. Cossi M, Rega N, Scalmani G, Barone V (2003) Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. Jour-

nal of Computational Chemistry 24(6):669–681, DOI 10.1002/jcc.10189

33. Bishop CM (2006) Pattern Recognition and Machine Learning. Information Science and Statistics, Springer
34. Jones E, Oliphant T, Peterson P, et al (2001–) SciPy: Open source scientific tools for Python. URL `http://www.scipy.org/`, [Online; accessed 2018-05-31]
35. Lundblad R, Macdonald F (2010) Handbook of Biochemistry and Molecular Biology, Fourth Edition. Taylor & Francis
36. Ndukwe IE, Wang X, Reibarkh M, Isik M, Martin GE (2018) NMR characterization of microstates of SM14. Tech. rep., Merck NMR Structure Elucidation Group
37. Faber NKM (1999) Estimating the uncertainty in estimates of root mean square error of prediction: application to determining the size of an adequate test set in multivariate calibration. Chemometrics and Intelligent Laboratory Systems 49(1):79 – 89, DOI 10.1016/S0169-7439(99)00027-1

Manuscript with changes highlighted

Click here to view linked References

Click here to download Manuscript manuscript-sampl6-revision-CHANGES.pdf

Journal of Computer-Aided Molecular Design manuscript No.
(will be inserted by the editor)

SAMPL6: Calculation of macroscopic pK_a values from *ab initio* quantum mechanical free energies

Edithe Selwa · Ian M. Kenney · Oliver Beckstein · Bogdan I. Iorga

Received: date / Accepted: date

Abstract Macroscopic pK_a values were calculated for all compounds in the SAMPL6 blind prediction challenge, based on quantum chemical calculations with a continuum solvation model and a linear correction derived from a small training set. Microscopic pK_a values were derived from the gas-phase free energy difference between protonated and deprotonated forms together with the Conductor-like Polarizable Continuum Solvation Model and the experimental solvation free energy of the proton. pH-dependent microstate free energies were obtained from the microscopic pK_as with a maximum likelihood estimator and appropriately summed to yield macroscopic pK_a values or microstate populations as function of pH. We assessed the accuracy of three approaches to calculate the microscopic pK_as: direct use of the quantum mechanical

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01GM118772 (to OB). BII was supported in part by grants ANR-10-LABX-33 (LabEx LERMIT) and ANR-14-JAMR-0002-03 (JPIAMR) from the French National Research Agency (ANR), and by a grant DIM MAL-INF from the Région Ile-de-France.

E. Selwa
Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Université Paris-Saclay, Labex LERMIT, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette, France

I. M. Kenney
Department of Physics, Arizona State University, P.O. Box 871504, Tempe, AZ 85287-1504, USA

O. Beckstein
Department of Physics and Center for Biological Physics, Arizona State University, P.O. Box 871504, Tempe, AZ 85287-1504, USA
Tel.: +1 480 727 9765
Fax: +1 480 965-4669
E-mail: oliver.beckstein@asu.edu

B. I. Iorga
Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Université Paris-Saclay, Labex LERMIT, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette, France
Tel.: +33 1 69 82 30 94
Fax: +33 1 69 07 72 47
E-mail: bogdan.iorga@cnrs.fr

E. Selwa and I. M. Kenney contributed equally to this work.

free energy differences and correction of the direct values for short-comings in the QM solvation model with two different linear models that we independently derived from a small training set of 38 compounds with known pK_a. The predictions that were corrected with the linear models had much better accuracy [root-mean-square error (RMSE) 2.04 and 1.95 pK_a units] than the direct calculation (RMSE 3.74). Statistical measures indicate that some systematic errors remain, likely due to differences in the SAMPL6 data set and the small training set with respect to their interactions with water. Overall, the current approach provides a viable physics-based route to estimate macroscopic pK_a values for novel compounds with reasonable accuracy.

Keywords pK_a · pH · quantum chemistry · SAMPL challenge

1 Introduction

The SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) challenges allow the molecular modeling community to assess, in "blind" conditions, the accuracy and efficiency of current computational chemistry methods and tools, leading to continuous improvements of the available computational methods. The previous SAMPL challenges [1–5] involved hydration free energy calculations, with the exception of the last edition, SAMPL5, which was dedicated to the prediction of distribution coefficients [6]. Our past participations in SAMPL challenges [7–9] represented unique opportunities for us to test our approaches and to develop and improve new computational tools. In 2018, the SAMPL6 challenge focused on the prediction of microscopic and macroscopic pK_a values for fragment-like organic compounds.

The equilibrium acid dissociation reaction in aqueous solution

$$\mathrm{HA}_{(s)} \rightleftharpoons \mathrm{H}^+_{(s)} + \mathrm{A}^-_{(s)} \tag{1}$$

with acid dissociation constant $K_a = [\mathrm{A}^-][\mathrm{H}^+]/[\mathrm{HA}]$ is of broad importance in biological systems, in synthetic chemistry, and pharmacology [10–14]. The pK_a, defined as

$$\mathrm{p}K_a = -\log_{10}\frac{K_a}{c_0} \tag{2}$$

for the standard state concentration $c_0 = 1\,\mathrm{mol/l}$, measures thermodynamic acidity. The theoretical prediction of pK_a values is an ongoing challenge [15]. In the narrow realm of protein biochemistry, good progress has been made in calculating the physiologically important *changes* in pK_as of standard amino acid residues in different environments with accuracies better than 1 pK_a unit [12], especially with constant pH molecular dynamics simulations [16–19], which have been applied to study a wide range of phenomena [20–22]. *Absolute* pK_a calculations of arbitrary molecules using physics-based quantum chemistry approaches (as opposed to machine learning (ML) ones) have been more challenging and accuracy of 1 pK_a unit has been difficult to achieve consistently [15, 23] whereas a range of methods can achieve "chemical accuracy" (defined as 2.5 pK_a units by Ho and Coote [15]). The clear advantage of *ab initio* approaches is that they can be applied to any novel compound. Here we report on pK_a calculations of the 24 compounds in the SAMPL6 challenge (Fig. 1) with a quantum-chemical approach originally developed by Muckerman et al [24]. The SAMPL6 compounds are, however, chemically more complex and 23 contain multiple titratable protons and, in some cases, tautomers so that *macroscopic* pK_a have to be calculated.

The calculation of microscopic pK_as, i.e., the free energy difference for the deprotonation reaction Eq. 1 at the standard state (concentration 1 mol/l and temperature $T = 298.15$ K, indicated by the superscript "*")

$$\mathrm{p}K_\mathrm{a} = \frac{\Delta G^*_{(\mathrm{s})}}{RT \ln 10}, \tag{3}$$

is straightforward using quantum chemical gas-phase calculations. However, it is well-known [15, 23] that direct calculations lead to large errors in the calculated pK_as, mainly due to the poor continuum solvation models that have to be employed in order to obtain free energies in solution. One approach to correct for these systematic errors is to generate a model to correct the raw quantum chemical free energies [24]. We generated linear models from a training set with 38 simple compounds with experimentally known pK_a (Fig. 2 and 3). We fit a global model to all the data (the global linear model) and we split the training set with a simple classifier, namely the charge of the acid, yielding a piecewise linear model with separate linear functions for neutral and cationic acids. We calculated the macroscopic pK_as for all 24 SAMPL6 compounds and compared the accuracy of the three approaches [QM computed (raw), linear fit global, and linear fit piecewise].

2 Methods

Following Muckerman et al [24], our strategy was to compute gas-phase free energy differences

$$\Delta G^\circ_{(\mathrm{g})} = G^\circ(\mathrm{A}^-_{(\mathrm{g})}) + G^\circ(\mathrm{H}^+_{(\mathrm{g})}) - G^\circ(\mathrm{HA}_{(\mathrm{g})}) \tag{4}$$

(denoted as standard state free energies at 1 atm pressure and 298.15 K) for the deprotonation reaction for all titratable protons,

$$\mathrm{HA}_{(\mathrm{g})} \rightleftharpoons \mathrm{H}^+_{(\mathrm{g})} + \mathrm{A}^-_{(\mathrm{g})}. \tag{5}$$

To obtain solution free energy differences corresponding to Eq. 1,

$$\Delta G^*_{(\mathrm{s})} = G^*(\mathrm{A}^-_{(\mathrm{s})}) + G^*(\mathrm{H}^+_{(\mathrm{s})}) - G^*(\mathrm{HA}_{(\mathrm{s})}) \tag{6}$$

(where the standard state refers to 1 mol/l), a solvation free energy contribution $\Delta G^\circ_\mathrm{solv}$ is added to the gas-phase free energies of the acid HA and the base A^- from Eq. 4,

$$G^*_{(\mathrm{s})} = G^\circ_{(\mathrm{g})} + \Delta G^{\circ\to *} + \Delta G^\circ_\mathrm{solv} = G^\circ_{(\mathrm{g})} + \Delta G^*_\mathrm{solv} \tag{7}$$

with $\Delta G^{\circ\to *} = 1.894\ \mathrm{kcal/mol}$ accounting for the change in standard state in the gas phase. The free energy of the proton in the gas phase is calculated analytically in the ideal gas limit (the Sackur-Tetrode equation [25]), $G^\circ(\mathrm{H}^+_{(\mathrm{g})}) = -6.28$ kcal/mol, and for the solvation free energy of the proton we chose the same value as Muckerman

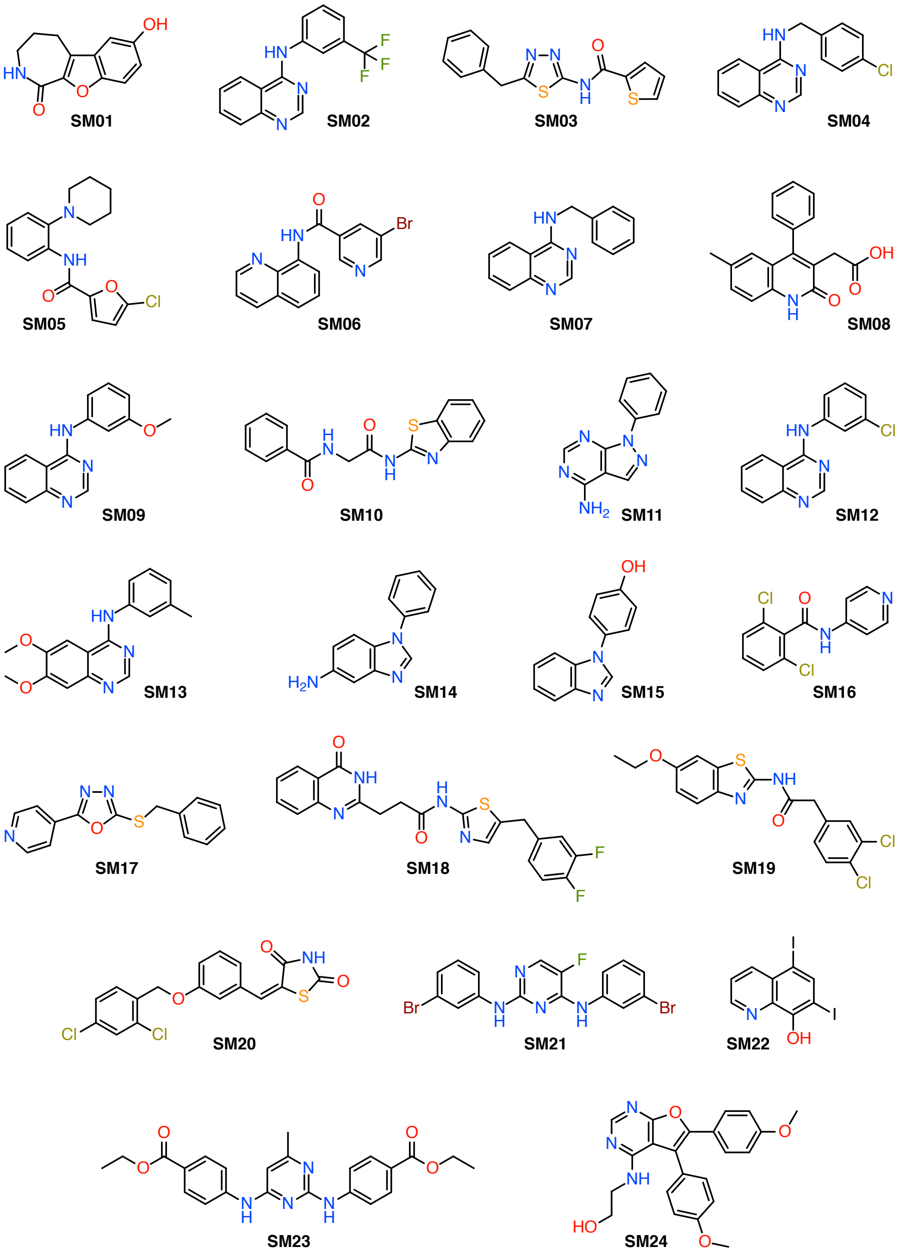


Fig. 1 Chemical structures of the SAMPL6 data set. **SM20** is the only compound that contains a single titratable proton; all other compounds contain multiple titratable protons and, in some cases, tautomers.

et al [24], $G^*(\mathrm{HA}_{(s)}) = -272.2$ kcal/mol although other values are also discussed in the literature [15, 26]. With $\Delta G^*_{(s)}$, the pK_a is calculated from Eq. 3.

As described in detail in Section 2.2, the directly calculated pK_a values have fairly poor accuracy and thus we derive a simple linear estimator to correct for shortcomings in the solvation model [24]. The linear model is based on our own training data set (described in the next section) and the resulting estimator $\mathscr{L}$ is applied to the pK_a from Eq. 3 to obtain improved predictions for the SAMPL6 data set, $\mathrm{p}K_a = \mathscr{L}[\mathrm{p}K_a^{\mathrm{calc}}]$.

2.1 Data sets

The QM1 subset of the training set contains 21 neutral acids belonging to several chemical families (Fig. 2): mono- (**1**), di- (**2**) and tri- (**3**) protic inorganic acids, aliphatic (**4**) and aromatic (**5**) sulfonic acids, diversely substituted carboxylic acids (**6**-**11**) and alcohols (**12**-**17**), phenols (**18** and **19**), phthalimide (**20**) and uracil (**21**). The experimental pK_a values of these compounds range from −3.00 to 17.10 (Table 1). The QM2 subset contains 17 compounds that are cationic acids (Fig. 3): hydrazine (**22**), guanidine (**23**), aliphatic mono- (**24**), di- (**25**) and tri- (**26**) substituted amines, diversely substituted aromatic amines (**27**-**31**) and pyridines (**32**-**38**). These compounds possess experimental pK_a values from 0.49 to 13.60 (Table 1).

The SAMPL6 data set consisted of 24 fragment-like small organic molecules (Fig. 1) with unknown pK_a values that were selected for their similarity to kinase inhibitors and for experimental tractability. It was provided by the SAMPL6 organizers through the public repository `https://github.com/MobleyLab/SAMPL6` as computer-generated microstates in SMILES format. The protonation state for each microstate was computed with an in-house script using the CACTVS Chemoinformatics Toolkit [27] (Xemistry GmbH, `https://www.xemistry.com/`), allowing the classification of microstates in two groups, neutral acids and cationic acids, for which different correction factors were applied in the approach using the piecewise linear model.

Three-dimensional coordinates for all compounds were generated in MOL2 format using CORINA version 3.60 (`http://www.molecular-networks.com`), then converted into the Gaussian input format using an in-house script. The PDF3D files, which can be visualized with Adobe Acrobat Reader (`https://get.adobe.com/fr/reader/`) were generated with CACTVS.

2.2 Quantum chemical microscopic pK_a calculations

Gas-phase geometry optimization and frequency calculation of the protonated and deprotonated forms were performed at the B3LYP/6-311+G(d,p) level using Gaussian 09 version D.01 [28] to obtain $\Delta G^\circ_{(g)}$. A single-point free energy evaluation at the same level using the Conductor-like Polarizable Continuum Solvation Model (CPCM) [29–32] and UAHF radii as implemented in Gaussian 09 version D.01 [28] yielded the solvation free energy $\Delta G^\circ_{\mathrm{solv}}$ so that $\Delta G^*_{(s)}$ (Eq. 7) and an estimate for

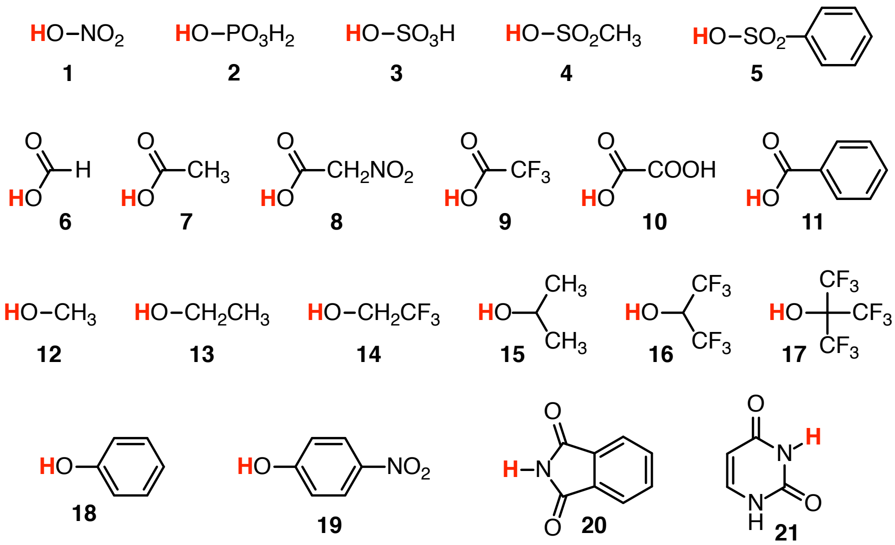


Fig. 2 Chemical structures of the QM1 training data set (neutral acids); see also Table 1.

the pK_a associated with this protonation/deprotonation event could be calculated via Eqs. 6 and 3.

In some cases, the geometry optimization did not converge with Gaussian 09 version D.01, but was successful with the version A.02 of Gaussian 09. Geometry optimization for microstates **SM04_micro016**, **SM07_micro016**, **SM17_micro008** and **SM17_micro009** did not converge in any conditions.

Muckerman et al [24] recognized systematic errors related to the solvation contribution $\Delta G^{*}_{\mathrm{solv}}$ as responsible for poor accuracy, namely the solvation model undersolvates weak acids and over-solvates strong acids. They proposed a physically-motivated correction

$$\Delta G^{*}_{\mathrm{corr}}(\mathrm{HA}) := RT \ln 10 \cdot (\mathrm{p}K_\mathrm{a}^{\mathrm{exp}} - \mathrm{p}K_\mathrm{a}^{\mathrm{calc}}) \tag{8}$$

to $\Delta G^{\circ}_{\mathrm{solv}}$ with the linear model

$$\Delta G^{*}_{\mathrm{corr}} = a_0 + a_1 \cdot \mathrm{p}K_\mathrm{a}^{\mathrm{exp}}. \tag{9}$$

The parameters a_0 and a_1 are determined from a training set by linear regression. In order to apply the correction Eq. 9 to compounds with unknown pK_a, a linear estimator $\mathscr{L}$ can be derived by substituting $\mathrm{p}K_\mathrm{a}^{\mathrm{exp}} \approx \mathrm{p}K_\mathrm{a}^{\mathrm{calc}} + \Delta G^{*}_{\mathrm{corr}}/(RT \ln 10)$ in Eq. 9 and solving for $\Delta G^{*}_{\mathrm{corr}}$ to yield

$$\Delta G^{*}_{\mathrm{corr}} = c_0 + c_1 \cdot \mathrm{p}K_\mathrm{a}^{\mathrm{calc}} \quad \text{with} \tag{10a}$$

$$c_0 = \frac{a_0}{1 - \lambda a_1} \tag{10b}$$

$$c_1 = \frac{a_1}{1 - \lambda a_1}, \quad \text{and} \quad \lambda := (RT \ln 10)^{-1} \tag{10c}$$

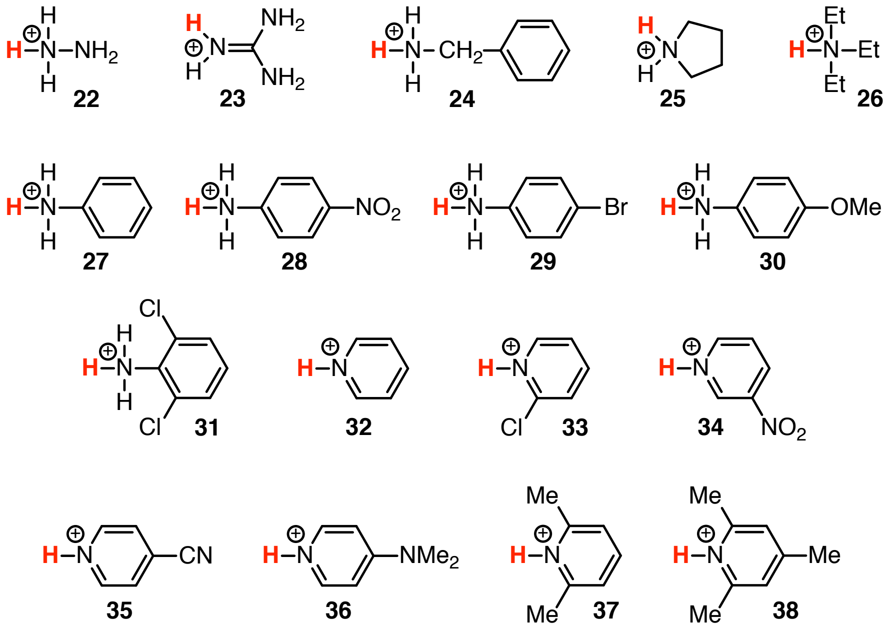


Fig. 3 Chemical structures of the QM2 training data set (cationic acids); see also Table 1.

The linear estimator $\mathscr{L}$ with parameters a_0 and a_1 for the microscopic pK_a is

$$\begin{aligned}\mathrm{p}K_\mathrm{a} = \mathscr{L}[\mathrm{p}K_\mathrm{a}^\mathrm{calc}] &= \mathrm{p}K_\mathrm{a}^\mathrm{calc} + \lambda \Delta G^*_\mathrm{corr} \\ &= \frac{\lambda a_0}{1-\lambda a_1} + \frac{1}{1-\lambda a_1} \cdot \mathrm{p}K_\mathrm{a}^\mathrm{calc}. \end{aligned} \tag{11}$$

2.3 Microstates *vs* Macrostates

We consider each tautomer of the acid HA and the base A^- as a *microstate* with label i. The set of microstates with the same total number of protons $N_i = N$ is labeled the *macrostate* N. The macroscopic pK_a characterizes the transitions between any of the microstates with N protons to any microstate with $N-1$ protons.

In general, the free energy difference between two states (micro or macro states) that are separated by a single protonation process (i.e., the free energy to go from N to $N-1$ associated protons) is

$$\Delta G_{N,N-1} = -\Delta G_{N-1,N} = -\beta^{-1} \ln \left[\frac{P(N-1)}{P(N)}\right] \tag{12}$$

where $P(N-1)$ and $P(N)$ are the probabilities of observing the system with $N-1$ and N associated protons respectively and $\beta = (RT)^{-1}$. The Henderson-Hasselbalch equation

$$\mathrm{p}K_\mathrm{a} = \mathrm{pH} - \log_{10}\left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right) = \mathrm{pH} - \frac{1}{\ln 10} \ln \left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right) \tag{13}$$

can be rewritten in terms of the free energy of protonation $\Delta G_{N-1,N}$ (Eq. 12) to give

$$\mathrm{p}K_\mathrm{a} = \mathrm{pH} - \frac{\beta \Delta G_{N-1,N}}{\ln 10}, \tag{14a}$$

$$\Delta G_{N-1,N} = \beta^{-1} \ln 10 \cdot (\mathrm{pH} - \mathrm{p}K_\mathrm{a}). \tag{14b}$$

2.4 Calculation of macroscopic pK_as from microscopic pK_as

The microscopic pK_a values correspond to free energy differences $\Delta G_{ij}(\mathrm{pH}) = G_j(\mathrm{pH}) - G_i(\mathrm{pH})$ between microstates i and j (Eq. 14b); for notational convenience we drop the explicit pH dependence in the following for all free energies. Each state has a pH-dependent associated free energy G_i, which is not known. Constructing the G_i from the differences between them is not straightforward because these calculated free energy differences come with unknown errors that prevent, for example, that the sum along any closed thermodynamic cycle $i \to j \to k \to \cdots \to i$ is exactly zero as required by the fact that the G_i are thermodynamic state functions. We construct a set of M microstate free energies $\{G_i\}_{i=1}^{M}$ that is most consistent with the calculated ("measured") $\{\Delta G_{ij}\}$ using a maximum-likelihood estimator [33] based on the likelihood function

$$L(\{G_i\}|\{\Delta G_{ij}\}) = \prod_{ij} \exp\left(-\tfrac{1}{2}[(G_j - G_i) - \Delta G_{ij}]^2\right), \tag{15}$$

where we assumed normal distribution of errors with constant standard deviation. The product runs over all pairs (i, j) for which calculated ΔG_{ij} are available. L is proportional to the probability $P(\{\Delta G_{ij}\}|\{G_i\})$ that we could observe the measured data (all the calculated ΔG_{ij}) if we were given a specific set of the G_i (our model parameters). Maximizing the log-likelihood $\ln L$ (using functions in SciPy [34]) as a function of all the G_i provides the set $\{G_i\}_{i=1}^{M}$ that is most consistent with the given measurements $\{\Delta G_{ij}\}$. Further details and more general applications of this approach will be published elsewhere (I.M. Kenney *et al*, in preparation).

In order to calculate the macroscopic pK_as, we begin by calculating the free energy of protonation using principles of equilibrium statistical mechanics [25]. The probability of observing a macrostate with N associated protons is

$$P(N) = Z^{-1} \sum_i e^{-\beta G_i} \delta_{N_i,N} \tag{16}$$

where the sum is over all accessible microstates with free energy G_i, $\delta_{N_i,N}$ is unity when the microstate i has N protons and null otherwise, and Z is the partition function, defined by

$$Z = \sum_j e^{-\beta G_j}. \tag{17}$$

Eq. 16 combined with the general expression for the free energy of protonation (Eq. 12) yields the effective macroscopic protonation free energy as a function of the G_i,

$$\Delta G_{N-1,N} = \beta^{-1} \ln \left[\frac{\sum_i e^{-\beta G_i} \delta_{N_i,N-1}}{\sum_i e^{-\beta G_i} \delta_{N_i,N}} \right]. \tag{18}$$

$\Delta G_{N-1,N}$ is a function of the pH of the system and the microscopic pK_as relevant to the macrostate N. Together with Eq. 14a, Eq. 18 allows us to calculate the macroscopic pK_a value for removing the N^th proton from a molecule. With all microstate free energies $\{G_i\}_{i=1}^{M}$ known for a given pH value it is also straightforward to compute the pH-dependent microstate probabilities

$$p_i(\mathrm{pH}) = Z(\mathrm{pH})^{-1} e^{-\beta G_i(\mathrm{pH})} \tag{19}$$

where all terms depend on pH.

2.5 Error analysis

The difference between experimental and computed pK_a values ("signed error") for each compound, labeled with its identification code 'id', was calculated as

$$\Delta_\mathrm{id} = \mathrm{pK}_\mathrm{a,id} - \mathrm{pK}^\mathrm{exp}_\mathrm{a,id}. \tag{20}$$

The root-mean-square error (RMSE) was determined from the individual errors Δ as

$$\mathrm{RMSE} = \sqrt{\langle \Delta^2 \rangle} = \sqrt{N^{-1} \sum_\mathrm{id}^{N} \Delta_\mathrm{id}^2}, \tag{21}$$

the mean absolute error (MAE) as

$$\mathrm{MAE} = \langle |\Delta| \rangle = N^{-1} \sum_\mathrm{id}^{N} |\Delta_\mathrm{id}|, \tag{22}$$

and the signed mean error (ME, also called the "mean signed error", MSE) as

$$\mathrm{ME} = \langle \Delta \rangle = N^{-1} \sum_\mathrm{id}^{N} \Delta_\mathrm{id}. \tag{23}$$

We also report the Pearson correlation coefficient R^2 and the slope m of a linear regression to the data, as computed with the function `scipy.stats.linregress()` in the SciPy package [34].

The quantum chemical single point free energy calculations do not have a statistical error and we have not yet implemented the calculation of an error bound in the maximum likelihood estimator for the G_i. Therefore, all pK_a are provided without a statistical error. Judging from the performance of the training data set and the post-hoc analysis of the SAMPL6 compounds (see Results), the accuracy of the calculated pK_a values is 1–2 pK_a units.

Calculated pK_a were compared to experimental values with the script `typeIII_analysis.py` as provided by the SAMPL6 organizers in the public repository `https://github.com/MobleyLab/SAMPL6`. Calculated values were matched to experimental ones with the *Hungarian algorithm*, which finds the optimum pairing between two sets by minimizing the linear sum of squared errors.

3 Results and Discussion

3.1 Training data set

The first step in our protocol was the design of a training data set containing 38 structurally-diverse, simple organic and inorganic compounds with known pK_a values. This global data set could be classified by the charge of the acid and split into two subsets. The neutral acids (named *QM1*, Fig. 2) contained 21 compounds and the second set, the positively-charged acids (named *QM2*, Fig. 3), contained the remaining 17 compounds. The structures were chosen from different chemical families in order to obtain for the two subsets a relatively homogeneous distribution of data points over a wide range of values (see Table 1 for the experimental pK_as).

Predicted pK_a values were computed for all compounds from the training data set using the protocol described by Muckerman et al [24] (see the Methods section for details). The correlation of these computed values with the experimental pK_as is shown in (Fig. 4a), with a Pearson correlation coefficient $R^2 = 0.96$ (Table 1). The corresponding $\Delta G^*_{\mathrm{corr}}$ values were obtained using Eq. 8 and plotted against the experimental pK_a values. A global linear fit model, with a slope of $a_1 = -0.61$ and an intercept of $a_0 = 2.75$ (parameters in Eq. 9), was derived by using all compounds as a single data set (Fig. 4b). Alternatively, a piecewise linear fit model was derived by considering separately the two QM1 and QM2 subsets (Fig. 4c). In this latter case we obtained the parameters in Eq. 9 with a slope of $a_1^{\mathrm{QM1}} = a_1^{\mathrm{QM2}} = -0.62$ for both subsets and intercept values of $a_0^{\mathrm{QM1}} = 1.30$ and $a_0^{\mathrm{QM2}} = 4.65$ for the QM1 and QM2 subsets, respectively.

The linear estimators associated with these models (Eq. 10a) were calculated using Eq. 11. These corrections were applied to the whole training set, and to the QM1 and QM2 subsets, respectively, in order to evaluate to which extent the systematic errors related to the prediction method were removed compared with the pK_a values obtained directly from the *ab initio* calculations (Table 1). We can see that in all cases the corrected pK_a values are much closer to the experimental values, with the global model behaving slightly better than the piecewise model, as shown by, for instance, the smaller RMSE 1.66 *vs* 1.85 for the whole training set.

3.2 Macroscopic pK_a

The microscopic pK_a values for the SAMPL6 data set were computed using the same protocol as for the training data set (595 individual transformations). Again, the corrections from the global linear model were applied to the whole SAMPL6 data set and alternatively, those from the piecewise linear model to individual subsets of the SAMPL6 data set containing the neutral acids and the cationic acids, respectively.

Starting from these three sets of results (obtained directly from *ab initio* free energies or after correction with the two linear models, global and piecewise) we calculated pH-dependent microstate free energies and macroscopic pK_a values (Table 2). These results, formatted using the SAMPL6 submission template, were used as input for the `typeIII_analysis.py` script in order to compare to the experimental values

Table 1 Experimental and computed pK_a values for the compounds from the QM1 (Fig. 2) and QM2 (Fig. 3) training data sets. The difference Δ (Eq. 20) between computed and experimental pK_a values is shown for each compound. The experimental values were taken from Muckerman et al [24] and from Lundblad and Macdonald [35]. The root-mean-square error (RMSE), the mean absolute error (MAE), and the signed mean error (ME) were calculated according to Eqs. 21–23.

id	Exp.	QM computed		$\Delta G^*_{correction}$	Linear fit global		Linear fit piecewise	
	pK_a	pK_a	Δ		pK_a	Δ	pK_a	Δ
1	−1.40	−4.01	−2.61	3.56	0.43	1.83	−0.31	1.09
2	2.15	1.75	−0.40	0.54	2.88	0.73	2.14	−0.01
3	−3.00	−10.32	−7.32	9.98	−0.68	2.32	−1.41	1.59
4	−1.90	−2.92	−1.02	1.39	0.08	1.98	−0.65	1.25
5	−2.80	−3.44	−0.64	0.88	−0.54	2.26	−1.27	1.53
6	3.77	5.16	1.39	−1.89	4.00	0.23	3.26	−0.51
7	4.76	7.27	2.51	−3.42	4.69	−0.07	3.94	−0.82
8	1.68	4.73	3.05	−4.16	2.56	0.88	1.82	0.14
9	0.23	−1.82	−2.05	2.80	1.55	1.32	0.82	0.59
10	1.38	4.06	2.68	−3.65	2.35	0.97	1.61	0.23
11	4.21	6.77	2.56	−3.49	4.31	0.10	3.56	−0.65
12	15.54	23.37	7.83	−10.68	12.14	−3.40	11.37	−4.17
13	15.90	22.73	6.83	−9.31	12.39	−3.51	11.62	−4.28
14	12.43	14.22	1.79	−2.44	9.99	−2.44	9.22	−3.21
15	17.10	22.61	5.51	−7.51	13.22	−3.88	12.44	−4.66
16	9.30	8.57	−0.73	0.99	7.83	−1.47	7.07	−2.23
17	5.40	4.28	−1.12	1.53	5.13	−0.27	4.38	−1.02
18	9.95	14.99	5.04	−6.87	8.28	−1.67	7.51	−2.44
19	7.14	8.26	1.12	−1.53	6.33	−0.81	5.58	−1.56
20	8.30	13.49	5.19	−7.08	7.14	−1.16	6.38	−1.92
21	9.50	13.81	4.31	−5.87	7.97	−1.53	7.2	−2.30
RMSE (QM1)			3.86			1.90		2.19
MAE (QM1)			3.13			1.56		1.72
ME (QM1)			1.61			−0.36		−1.11
R^2 (QM1)		0.97						
m (QM1)		1.45						
22	8.12	9.05	0.93	−1.26	7.01	−1.11	7.93	−0.19
23	13.60	18.94	5.34	−7.28	10.8	−2.80	11.7	−1.90
24	9.30	8.48	−0.82	1.12	7.83	−1.47	8.75	−0.55
25	11.27	8.18	−3.09	4.22	9.19	−2.08	10.1	−1.17
26	10.72	12.64	1.92	−2.61	8.81	−1.91	9.72	−1.00
27	4.62	2.91	−1.71	2.33	4.59	−0.03	5.53	0.91
28	0.98	−4.10	−5.08	6.93	2.07	1.09	3.02	2.04
29	3.89	1.50	−2.39	3.26	4.09	0.20	5.02	1.13
30	5.36	5.35	−0.01	0.01	5.1	−0.26	6.04	0.68
31	1.53	−3.20	−4.73	6.44	2.45	0.92	3.4	1.87
32	5.24	4.67	−0.57	0.78	5.02	−0.22	5.95	0.71
33	0.49	−0.96	−1.45	1.98	1.73	1.24	2.69	2.20
34	0.81	−1.37	−2.18	2.98	1.96	1.15	2.91	2.10
35	1.86	0.14	−1.72	2.35	2.68	0.82	3.63	1.77
36	9.60	11.07	1.47	−2.00	8.04	−1.56	8.95	−0.65
37	6.70	7.89	1.19	−1.63	6.03	−0.67	6.96	0.26
38	7.33	8.22	0.89	−1.21	6.47	−0.86	7.39	0.06
RMSE (QM2)			2.60			1.30		1.33
MAE (QM2)			2.09			1.08		1.13
ME (QM2)			−0.71			−0.44		0.49
R^2 (QM2)		0.96						
m (QM2)		1.45						
RMSE (Global)			3.35			1.66		1.85
MAE (Global)			2.66			1.35		1.46
ME (Global)			0.58			−0.40		−0.40
R^2 (Global)		0.96						
m (Global)		1.44						

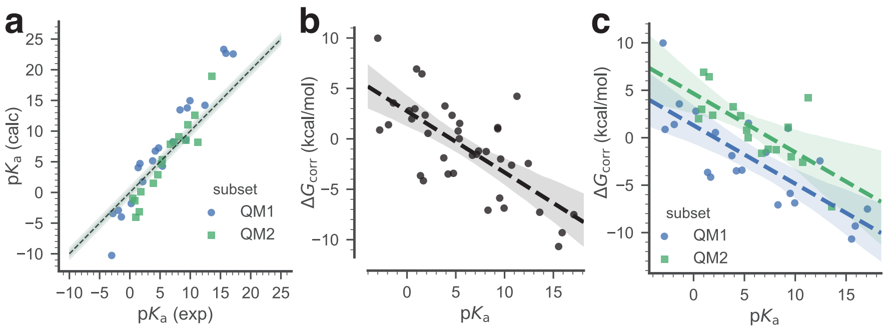


Fig. 4 Training data set. The pK_a of the training data set compounds are used to derive a simple linear model that relates the free energy correction ΔG^*_{corr} to the experimental pK_a. Two linear models were derived: a *global* linear model (black dashed line), utilizing all data, and a *piecewise* linear model that applies to either neutral acids (subset QM1, blue) or to positively charged acids (subset QM2, green). **a**: Correlation between experimental and calculated pK_a of the training data set. The dashed line indicates ideal correlation with the gray band indicating 1 pK_a unit deviation. **b**: Global linear fit of the calculated ΔG^*_{corr} to the experimental pK_a. **c**: Linear fits of the calculated ΔG^*_{corr} to the experimental pK_a, split between the QM1 and the QM2 subsets. In (b) and (c) the dashed lines are linear models to the data, with shaded bands indicating 95% confidence intervals from 1000 bootstrap samples.

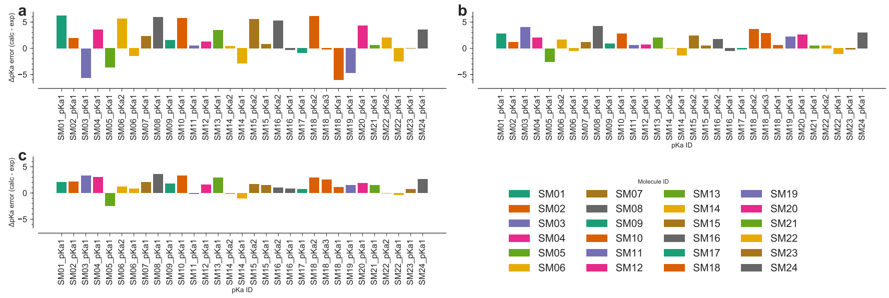


Fig. 5 Signed error Δ_{id} of individual predictions. The calculated pK_a was matched to the experimental pK_a for each compound (indicated by the SAMPL6 pKa ID) and the deviation from the experimental value represented as a bar. Observations for the same compound have the same color. **a**: pK_a were directly estimated from the quantum mechanical free energy differences. **b**: The quantum mechanical pK_a were corrected with the global linear model. **c**: compounds were corrected depending on their membership in subsets 1 or 2 with the piecewise linear model.

that were provided by the SAMPL6 organizers together with the analysis scripts. The input files with our results formatted as comma-separated value (CSV) files and the optimized structures for all microstates in MOL2 and PDF3D format are provided in the Electronic Supplementary Material. During the challenge we submitted macroscopic pK_a values only for three compounds (**SM15**, **SM20** and **SM22**). Here we describe the macroscopic pK_a predictions for the entire SAMPL6 data set.

Using this protocol we could predict the macroscopic pK_a values for the 24 SAMPL6 compounds with a RMSE of about 2 pK_a units when the corrections were applied and of 3.74 pK_a units when the *ab initio* free energies were used directly. The

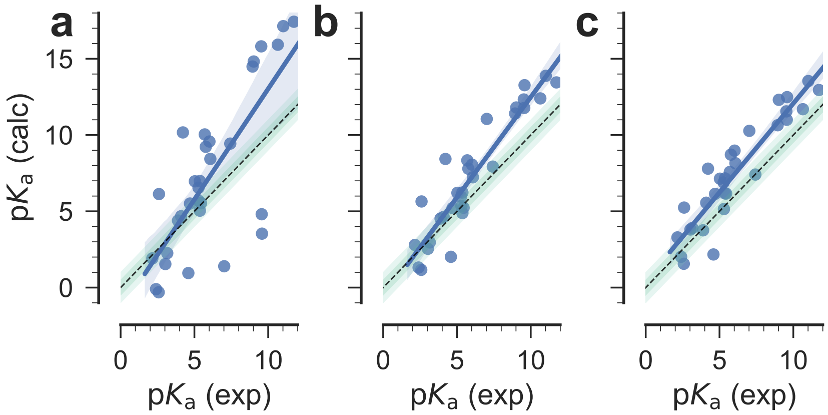


Fig. 6 Correlation between experimental and calculated pK_a values for the SAMPL6 compounds. **a**: pK_a were directly estimated from the quantum mechanical free energy differences. **b**: The quantum mechanical pK_a were corrected with the global linear model. **c**: compounds were corrected depending on their membership in subsets 1 or 2. The black dashed line indicates ideal correlation, the shaded green bars show 0.5 and 1 pK_a units deviation from ideal. Blue lines are linear regression fits to the data, with the blue shaded area indicating the 95% confidence interval from 1000 bootstrap samples.

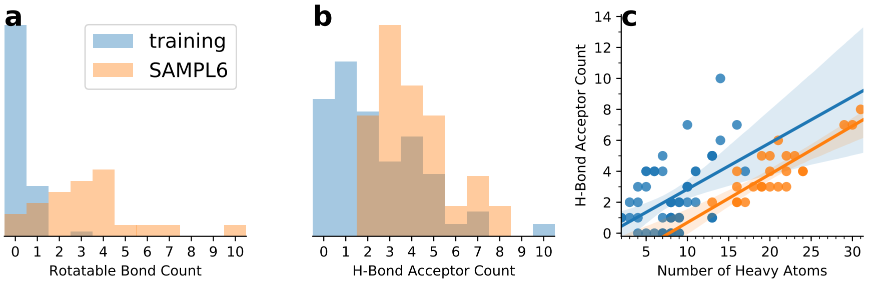


Fig. 7 Comparison of chemical properties of the training (light blue) and SAMPL6 (orange) data sets. **a**: normalized histograms of the number of rotatable bonds; **b**: normalized histograms of the number of hydrogen bond acceptors; **c**: correlation between the number of heavy atoms and the number of acceptors with linear regressions shown as solid lines and their 95% confidence interval from 1000 bootstraps indicated by shaded areas.

relative poor accuracy when directly using the quantum chemical free energies is in line with previous studies [15, 24].

The signed errors of individual predictions represented in Fig. 5 show that most of the prediction errors after correction are positive, with the notable exception of compound **SM05** for which these errors are consistently negative. High prediction errors ($3-4$ pK_a units) are obtained for compounds **SM03** and **SM08**, whereas compounds **SM01**, **SM04**, **SM10**, **SM13**, **SM18**, **SM20**, and **SM24** are predicted with errors of about $2-3$ pK_a units. The representation of the prediction errors in the order of increasing absolute experimental pKa values (Fig. S3, Electronic Supplementary

Material) shows that these are not related. Therefore, the source of remaining errors after correction should be sought elsewhere. As shown in Fig. 6, the results for the SAMPL6 data set are fairly insensitive to the fitting approach used (global or piecewise linear model), further indicating some level of robustness. Other statistical measures such as Pearson correlation coefficient $R^2 = 0.86$ and the slope of the linear regression $m = 1.17$ (for the piecewise linear model, see Table 2 for almost identical values for the global linear model) indicate encouraging correlations but the large mean error (1.42 for the piecewise linear model and 1.24 for the global linear model) hint at remaining systematic errors.

The fact that the linear fit did not remove these systematic errors implies that the training data set did not include properties that are important for the SAMPL6 data set and hence the linear or piecewise linear estimator cannot correct model errors related to these properties. In order to quantify similarities and differences between the two datasets we analyzed a number of chemical properties (see section *Properties of the training and SAMPL6 data sets* with Fig. S1 in the Electronic Supplementary Material file for details). Overall, the most obvious differences between our training and the SAMPL6 data set are the higher flexibility of the SAMPL6 molecules (with a median three and maximum ten rotatable bonds versus a median zero and maximum three, Fig. 7a) and the greater capability to accept hydrogen bonds (median four and maximum eight hydrogen bond acceptors versus median two and maximum ten; Fig. 7b), which correlates with a larger polar surface area (see Fig. S2 in the Electronic Supplementary Material file). However, Fig. 7c shows that the training compounds have *more* hydrogen bond acceptors for the same number of heavy atoms than the SAMPL6 compounds, i.e., for their larger size, the SAMPL6 compounds have fewer acceptors than one would expect from simple extrapolation of the training compounds. Similarly, the polar surface area of the SAMPL6 compounds would be overestimated from the training set (Fig. S2). These differences suggest that the interactions with water through hydrogen bonds are stronger in the training set than in the SAMPL6 set, which could lead to a systematic error in the estimator that was derived from the training set.

In the post-challenge analysis, we also tested the introduction of a conformational search step in our protocol and evaluated its influence on the quality of our predictions using two model compounds, **SM06** and **SM20**. The complete results are presented in the *Conformational search* section of the Electronic Supplementary Material file. In brief, for **SM06** the new microscopic pK_a value of **SM06_micro011** brought no changes in the predicted macroscopic pK_a values and for **SM20** we obtained macroscopic pK_a prediction errors 1.8-2.4 pK_a units higher compared with the values obtained without conformational search. It seems that, at least for these two compounds, the conformational search does not yield any substantial improvements in the prediction of macroscopic pK_a values.

3.3 Microstate probabilities

The SAMPL6 organizers recently made available experimental assignments of microstates with corresponding microstate pK_a for a number of compounds [36] (`https:`

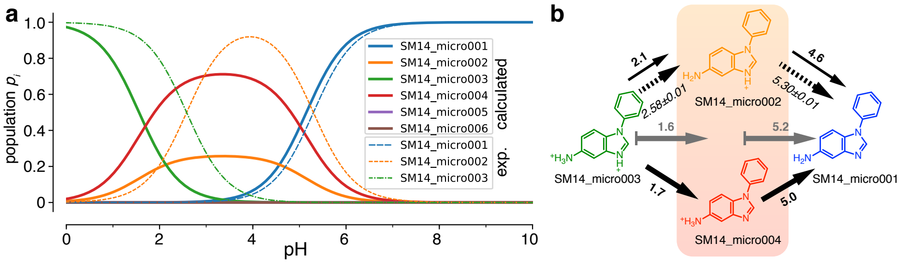


Fig. 8 Microstate probabilities p_i for **SM14**. **a**: Computed microstate probabilities (for the piecewise linear fit) are shown as heavy solid lines and experimentally derived probabilities as thin dashed lines. The experimental p_i were calculated in the same way as the calculated ones (Eq. 19) by directly using the experimental microstate pK_as. **b**: Microstate diagram with arrows indicating deprotonation. Bold numbers near solid arrows are the calculated microstate pK_a (from (a)) and italic numbers near dashed arrows are the experimental numbers, assigned to the experimentally identified microstate transitions. The gray solid arrows with gray bold numbers indicate the calculated macroscopic pK_a from $N = 3$ protons (microstate **SM14_micro003**) to $N = 2$ protons (mixture of **SM14_micro002** and **SM14_micro004**, indicated by the orange box) to $N = 1$ proton in **SM14_micro001** (and **SM14_micro005**, which is not shown because computation and experiment indicate that it is suppressed relative to **SM14_micro001**).

`//github.com/MobleyLab/SAMPL6/blob/master/physical_properties/pKa/experimental_data/NMR_microstate_determination/`). Here we focus on **SM14** as an example. Fig. 8 compares our computed microstate probabilities p_i (Eq. 19) to the ones derived from the experimental assignments of states **SM14_micro003**, **SM14_micro002**, and **SM14_micro001**. The important calculated microstates (from the linear piecewise model) were **SM14_micro003** ($N = 3$ protons), **SM14_micro004** and **SM14_micro002**, both with $N = 2$ protons, and **SM14_micro001** ($N = 1$). The calculated microscopic pK_a for the deprotonation of **SM14_micro003** to **SM14_micro002** was 2.1, similar to the experimental value 2.58 ± 0.01. The microscopic pK_a corresponding to the deprotonation of **SM14_micro002** to **SM14_micro001** was calculated as 4.6, also similar to the experimental one, 5.30 ± 0.01. A second microstate **SM14_micro005** exists with the same number of protons as **SM14_micro002** but both experiment and our computations indicated that this second state is suppressed and plays no role. Our calculations, however, assigned a higher population to **SM14_micro004** than to **SM14_micro002**, in contrast to the experimental findings, which, based on NMR nitrogen chemical shift measurements in the aprotic solvent acetonitrile-d_3 under pH titration, identified **SM14_micro002** as the dominant intermediate state. The partial agreement between these detailed experiments and our calculations is encouraging but a single comparison does not allow us to draw any broader conclusions except perhaps to highlight the ease with which our partition function-based formalism can be used to compute microscopic populations.

3.4 Computation time

The total computational cost required by this project was 641 CPU-days on a Linux cluster making use of Intel Xeon E5-4627 v3 CPUs running at 2.60 GHz. Given that

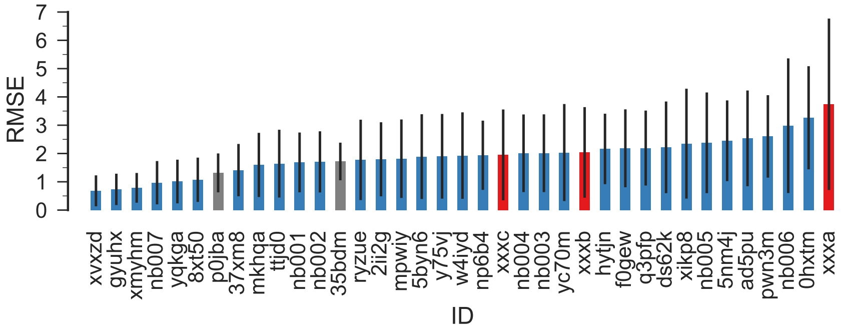


Fig. 9 RMSE of all SAMPL6 submissions (blue), including our new calculations for all SAMPL6 compounds (red) and for completeness our original submissions (gray), which only included predictions for **SM15**, **SM20**, and **SM22** and is only of limited statistical validity because of the large variance of the RMSE itself for only three samples [37]. The submission IDs *p0jba* and *xxxc* correspond to the piecewise linear model, *35bdm* and *xxxb* to the global linear model, and *xxxa* to directly using the quantum chemical free energies. Other IDs belong to other regular SAMPL6 submissions. The error bars indicate 95% confidence intervals from 1000 bootstrap samples.

344 microstates were computed, each microstate required 1.86 CPU-days on average. The calculations were carried out in parallel on 8 cores, so the average wall clock time for a microstate was 5.6 hours in these conditions. The most rigid compound, **SM22**, was the fastest with 1 CPU-hour for one microstate, whereas one of the biggest and most flexible compounds from the SAMPL6 data set, **SM18**, required about 3.2 CPU-days for one microstate.

4 Conclusions

Compared to other methods in the SAMPL6 challenge, our approach has below-average accuracy (Fig. 9 and Figs. S4–S7 in the Electronic Supplementary Material) and its computational cost is also higher than ML-based approaches (not considering the cost for compiling and validating the data and training the ML model). A key advantage of our approach is its generality as it does not depend on training on specific data sets although below we note that the quality of the training set for the correction step is a possible concern. With the linear model, which was derived from a very small and simple training set (38 compounds), we remove some of the errors related to the QM method used and its implementation in Gaussian (e.g., the implicit solvation model). The quality of the prediction is mostly independent of the structure, i.e., it can predict organic compounds from different families and even inorganic compounds with similar level of accuracy. In comparison, purely ML-based methods are trained on large experimental data sets (containing several thousands or tens of thousands compounds) and they can be vulnerable to chemical families that are not represented in the training set. Our approach appears reasonably robust because for our training set we obtain the same slope on the global data set and on the

Table 2 Experimental and computed pK_a values for the compounds from the SAMPL6 data set (Fig. 1). The difference Δ (Eq. 20) between computed and experimental pK_a values is shown for each compound. The experimental values were provided by the SAMPL6 organizers. The root-mean-square error (RMSE), the mean absolute error (MAE), and the signed mean error (ME) were calculated according to Eqs. 21–23. The Pearson correlation coefficient R^2 and the slope m were calculated from a linear regression.

Compound	pK_a	Exp.	QM computed		Linear fit global		Linear fit piecewise	
ID	ID	pK_a	pK_a	Δ	pK_a	Δ	pK_a	Δ
SM01	pKa1	9.53(1)	15.81	6.28	12.33	2.80	11.55	2.02
SM02	pKa1	5.03(1)	6.97	1.94	6.21	1.18	7.14	2.11
SM03	pKa1	7.02(1)	1.40	−5.62	11.06	4.04	10.27	3.25
SM04	pKa1	6.02(1)	9.58	3.56	8.06	2.04	8.98	2.96
SM05	pKa1	4.59(1)	0.95	−3.64	2.02	−2.57	2.17	−2.42
SM06	pKa1	3.03(4)	1.54	−1.49	2.54	−0.49	3.81	0.78
SM06	pKa2	11.74(1)	17.43	5.69	13.45	1.71	12.95	1.21
SM07	pKa1	6.08(1)	8.44	2.36	7.23	1.15	8.15	2.07
SM08	pKa1	4.22(1)	10.17	5.95	8.43	4.21	7.80	3.58
SM09	pKa1	5.37(1)	6.99	1.62	6.23	0.86	7.16	1.79
SM10	pKa1	9.02(1)	14.82	5.80	11.81	2.79	12.31	3.29
SM11	pKa1	3.89(1)	4.39	0.50	4.53	0.64	3.75	−0.14
SM12	pKa1	5.28(1)	6.55	1.27	5.96	0.68	6.89	1.61
SM13	pKa1	5.77(1)	9.23	3.46	7.79	2.02	8.72	2.95
SM14	pKa1	2.58(1)	−0.31	−2.89	1.16	−1.42	1.56	−1.02
SM14	pKa2	5.30(1)	5.68	0.38	5.34	0.04	5.15	−0.15
SM15	pKa1	4.70(1)	5.51	0.81	5.21[a]	0.51	6.14[b]	1.44
SM15	pKa2	8.94(1)	14.49	5.55	11.41[a]	2.47	10.64[b]	1.70
SM16	pKa1	5.37(1)	5.04	−0.33	4.88	−0.49	6.17	0.80
SM16	pKa2	10.65(1)	15.92	5.27	12.40	1.75	11.69	1.04
SM17	pKa1	3.16(1)	2.26	−0.90	2.96	−0.20	3.90	0.74
SM18	pKa1	2.15(2)	1.91	−0.24	2.80	0.65	3.29	1.14
SM18	pKa2	9.58(3)	3.54	−6.04	13.27	3.69	12.49	2.91
SM18	pKa3	11.02(4)	17.14	6.12	13.88	2.86	13.54	2.52
SM19	pKa1	9.56(2)	4.81	−4.75	11.78	2.22	11.00	1.44
SM20	pKa1	5.70(3)	10.04	4.34	8.34[a]	2.64	7.58[b]	1.88
SM21	pKa1	4.10(1)	4.68	0.58	4.63	0.53	5.56	1.46
SM22	pKa1	2.40(2)	−0.10	−2.50	1.32[a]	−1.08	2.02[b]	−0.38
SM22	pKa2	7.43(1)	9.44	2.01	7.93[a]	0.50	7.41[b]	−0.02
SM23	pKa1	5.45(1)	5.53	0.08	5.23	−0.22	6.16	0.71
SM24	pKa1	2.60(1)	6.13	3.53	5.65	3.05	5.25	2.65
RMSE				3.74		2.04		1.95
MAE				3.08		1.66		1.68
ME				1.25		1.24		1.42
R^2			0.58		0.87		0.86	
m			1.45		1.31		1.17	

[a] These results represent our submission *35bdm* to SAMPL6.

[b] These results represent our submission *p0jba* to SAMPL6.

individual subsets, which are chemically quite different. The results for the SAMPL6 data set are also fairly insensitive to the fitting approach used (global or piecewise linear model), further demonstrating robustness. The correlations with experimental data are generally good but suffer from systematic errors, possibly from differences between the training set and the SAMPL6 set that bias the estimator that is needed to correct the raw QM pK_a values. The statistical measures indicate clear room from improvement. It appears that a better correction scheme, using a larger data set that better matches the test data set with respect to its hydrogen bonding properties and is generally more representative of drug-like molecules could improve the predictions, perhaps in conjunction with more sophisticated classifiers and estimators than simple separation by charge and linear regression. We expect that improvements in the model physics, namely in the treatment of solvation, could also lead to further increases in accuracy.

We currently consider the method described here (and originally developed by Muckerman et al [24]) as an acceptable compromise between speed, accuracy and generality across the chemical space. It seems especially useful when one encounters novel compounds and wants to assess them based on their absolute pK_a values. The calculations are tractable with typical computational resources, absolute pK_as are accurate to about 2 units (within the "chemical accuracy" range [15]) and do not seem to be biased with respect to specific chemical groups, and thus the relative ordering of compounds is also meaningful.

References

1. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, Pande VS (2008) Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. J Med Chem 51(4):769–779, DOI 10.1021/jm070549+
2. Guthrie JP (2009) A blind challenge for computational solvation free energies: Introduction and overview. J Phys Chem B 113(14):4501–4507, DOI 10.1021/jp806724u
3. Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ (2010) The SAMPL2 blind prediction challenge: Introduction and overview. J Comput Aided Mol Des 24(4):259–279, DOI 10.1007/s10822-010-9350-8
4. Geballe MT, Guthrie JP (2012) The SAMPL3 blind prediction challenge: transfer energy overview. J Comput Aided Mol Des 26(5):489–96, DOI 10.1007/s10822-012-9568-8
5. Mobley DL, Wymer KL, Lim NM, Guthrie JP (2014) Blind prediction of solvation free energies from the SAMPL4 challenge. J Comput Aided Mol Des 28(3):135–50, DOI 10.1007/s10822-014-9718-2
6. Bannan CC, Calabró G, Kyu DY, Mobley DL (2016) Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. J Chem Theory Comput 12(8):4015–24, DOI 10.1021/acs.jctc.6b00449
7. Beckstein O, Iorga BI (2012) Prediction of hydration free energies for aliphatic and aromatic chloro derivatives using molecular dynamics simulations with

the OPLS-AA force field. J Comput Aided Mol Des 26(5):635–645, DOI 10.1007/s10822-011-9527-9

8. Beckstein O, Fourrier A, Iorga BI (2014) Prediction of hydration free energies for the SAMPL4 diverse set of compounds using molecular dynamics simulations with the OPLS-AA force field. J Comput Aided Mol Des 28(3):265–276, DOI 10.1007/s10822-014-9727-1
9. Kenney IM, Beckstein O, Iorga BI (2016) Prediction of cyclohexane-water distribution coefficients for the SAMPL5 data set using molecular dynamics simulations with the OPLS-AA force field. J Comput Aided Mol Des 30(11):1045–1058, DOI 10.1007/s10822-016-9949-5
10. Babić S, Horvat AJ, Mutavdžić Pavlović D, Kaštelan-Macan M (2007) Determination of pK$_a$ values of active pharmaceutical ingredients. TrAC Trends in Analytical Chemistry 26(11):1043–1061, DOI https://doi.org/10.1016/j.trac.2007.09.004
11. Lee AC, Crippen GM (2009) Predicting pK_a. Journal of Chemical Information and Modeling 49(9):2013–2033, DOI 10.1021/ci900209w
12. Alexov E, Mehler EL, Baker N, Baptista AM, Huang Y, Milletti F, Nielsen JE, Farrell D, Carstensen T, Olsson MHM, Shen JK, Warwicker J, Williams S, Word JM (2011) Progress in the prediction of pK_a values in proteins. Proteins 79(12):3260–75, DOI 10.1002/prot.23189
13. Rupp M, Korner R, Tetko IV (2011) Predicting the pK_a of small molecules. Combinatorial Chemistry & High Throughput Screening 14(5):307–327, DOI 10.2174/138620711795508403
14. Reijenga J, van Hoof A, van Loon A, Teunissen B (2013) Development of methods for the determination of pK_a values. Anal Chem Insights 8:53–71, DOI 10.4137/ACI.S12304
15. Ho J, Coote ML (2009) A universal approach for continuum solvent pK_a calculations: are we there yet? Theoretical Chemistry Accounts 125(1-2):3–21, DOI 10.1007/s00214-009-0667-0
16. Mongan J, Case DA, McCammon JA (2004) Constant pH molecular dynamics in generalized born implicit solvent. J Comput Chem 25(16):2038–48, DOI 10.1002/jcc.20139
17. Chen W, Morrow BH, Shi C, Shen JK (2014) Recent development and application of constant pH molecular dynamics. Molecular Simulation 40(10-11):830–838, DOI 10.1080/08927022.2014.907492
18. Swails JM, York DM, Roitberg AE (2014) Constant pH replica exchange molecular dynamics in explicit solvent using discrete protonation states: Implementation, testing, and validation. J Chem Theory Comput 10(3):1341–1352, DOI 10.1021/ct401042b
19. Radak BK, Chipot C, Suh D, Jo S, Jiang W, Phillips JC, Schulten K, Roux B (2017) Constant-pH molecular dynamics simulations for large biomolecular systems. Journal of Chemical Theory and Computation 13(12):5933–5944, DOI 10.1021/acs.jctc.7b00875
20. Di Russo NV, Estrin DA, Martí MA, Roitberg AE (2012) pH-dependent conformational changes in proteins and their effect on experimental pK(a)s: the case of nitrophorin 4. PLoS Comput Biol 8(11):e1002,761, DOI 10.1371/jour-

nal.pcbi.1002761

21. Morrow BH, Koenig PH, Shen JK (2013) Self-assembly and bilayer-micelle transition of fatty acids studied by replica-exchange constant pH molecular dynamics. Langmuir 29(48):14,823–30, DOI 10.1021/la403398n
22. Huang Y, Chen W, Dotson DL, Beckstein O, Shen J (2016) Mechanism of pH-dependent activation of the sodium-proton antiporter NhaA. Nature Communications 7:12,940, DOI 10.1038/ncomms12940
23. Alongi KS, Shields GC (2010) Theoretical calculations of acid dissociation constants: A review article. In: Ann Rep Comput Chem, vol 6, Elsevier Science B.V., chap 8, pp 113–138, DOI 10.1016/S1574-1400(10)06008-1
24. Muckerman JT, Skone JH, Ning M, Wasada-Tsutsui Y (2013) Toward the accurate calculation of pK_a values in water and acetonitrile. Biochimica et Biophysica Acta 1827:882–891, DOI 10.1016/j.bbabio.2013.03.011
25. McQuarrie DA (1976) Statistical Mechanics. HarperCollins, New York
26. Zhang H, Jiang Y, Yan H, Yin C, Tan T, van der Spoel D (2017) Free-energy calculations of ionic hydration consistent with the experimental hydration free energy of the proton. The Journal of Physical Chemistry Letters 8(12):2705–2712, DOI 10.1021/acs.jpclett.7b01125, pMID: 28561580
27. Ihlenfeldt W, Takahashi Y, Abe H, Sasaki S (1994) Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility. J Chem Inf Comput Sci 34(1):109–116 (http://www.xemistry.com/)
28. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA Jr, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas O, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT
29. Klamt A, Schüürmann G (1993) COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J Chem Soc, Perkin Trans 2 pp 799–805, DOI 10.1039/P29930000799
30. Andzelm J, Külmel C, Klamt A (1995) Incorporation of solvent effects into density functional calculations of molecular energies and geometries. The Journal of Chemical Physics 103(21):9312–9320, DOI 10.1063/1.469990
31. Barone V, Cossi M (1998) Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. The Journal of Physical Chemistry A 102(11):1995–2001, DOI 10.1021/jp9716997
32. Cossi M, Rega N, Scalmani G, Barone V (2003) Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. Jour-

nal of Computational Chemistry 24(6):669–681, DOI 10.1002/jcc.10189

33. Bishop CM (2006) Pattern Recognition and Machine Learning. Information Science and Statistics, Springer
34. Jones E, Oliphant T, Peterson P, et al (2001–) SciPy: Open source scientific tools for Python. URL `http://www.scipy.org/`, [Online; accessed 2018-05-31]
35. Lundblad R, Macdonald F (2010) Handbook of Biochemistry and Molecular Biology, Fourth Edition. Taylor & Francis
36. Ndukwe IE, Wang X, Reibarkh M, Isik M, Martin GE (2018) NMR characterization of microstates of SM14. Tech. rep., Merck NMR Structure Elucidation Group
37. Faber NKM (1999) Estimating the uncertainty in estimates of root mean square error of prediction: application to determining the size of an adequate test set in multivariate calibration. Chemometrics and Intelligent Laboratory Systems 49(1):79 – 89, DOI 10.1016/S0169-7439(99)00027-1

Electronic Supplementary Material

Blinded evaluation of farnesoid X receptor (FXR) ligands binding using molecular docking and free energy calculations

Edithe Selwa,[1,‡] *Eddy Elisee,*[1,‡] *Agustin Zavala,*[1,‡] *Bogdan I. Iorga*[1,*]

[1] Institut de Chimie des Substances Naturelles, CNRS UPR 2301, LabEx LERMIT, 91198 Gif-sur-Yvette, France

Corresponding Author

[*] Phone: +33 1 6982 3094; Fax: +33 1 6907 7247; Email: bogdan.iorga@cnrs.fr (B.I.I.).

Author Contributions

[‡] These authors contributed equally.

Table of contents

Protein Data Bank (PDB) structures available

27 structures were available in the PDB for FXR at the moment when the D3R Grand Challenge 2 took place. They were organized in 5 distinct groups, according to the type of ligand and the binding site conformation. The representative structure for each group (based on the crystal structure resolution and lack of missing residues) is colored in red.

Group 1 (isoxazoles): 3dct, 3dcu, 3gd2, **3hc5**, 3hc6, 3p88, 3p89, 3rut, 3ruu, 3rvf

Group 2 (benzimidazoles): 3okh, 3oki, **3olf**, 3omk, 3omm, 3oof, 3ook

Group 3 (FXR_5-like): 3l1b, **3fli**

Group 4 (steroid, FXR_34-like): 3bej, **1osv**, 1ot7, 4qe6

Group 5 (miscellaneous): 1osh, 4oiv, 4qe8, **4wvd**

FXR ligands regroupment according to their chemical structure

Isoxazoles: FXR_4, FXR_23, FXR_33, FXR_65

Benzimidazoles: FXR_6, FXR_7, FXR_8, FXR_9, FXR_13, FXR_14, FXR_19, FXR_20, FXR_21, FXR_22, FXR_24, FXR_25, FXR_26, FXR_27, FXR_28, FXR_29, FXR_30, FXR_31, FXR_32, FXR_35, FXR_36, FXR_37, FXR_39, FXR_40, FXR_42, FXR_50, FXR_51, FXR_52, FXR_53, FXR_54, FXR_55, FXR_56, FXR_57, FXR_58, FXR_59, FXR_60, FXR_61, FXR_62, FXR_63, FXR_64, FXR_66, FXR_67, FXR_68, FXR_69, FXR_70, FXR_71, FXR_72

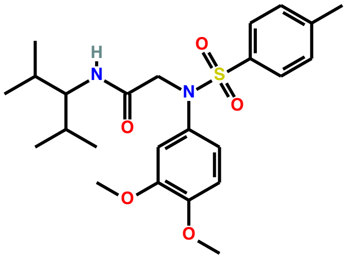
Spiro compounds: FXR_10, FXR_11, FXR_12, FXR_38, FXR_41, FXR_73, FXR_74, FXR_75, FXR_76, FXR_77, FXR_78, FXR_79, FXR_80, FXR_81, FXR_82, FXR_83, FXR_84, FXR_85, FXR_86, FXR_87, FXR_88, FXR_89

Sulfonamides: FXR_15, FXR_16, FXR_17, FXR_43, FXR_44, FXR_45, FXR_46, FXR_47, FXR_48, FXR_49, FXR_90, FXR_91, FXR_92, FXR_93, FXR_94, FXR_95, FXR_96, FXR_97, FXR_98, FXR_99, FXR_100, FXR_101, FXR_102

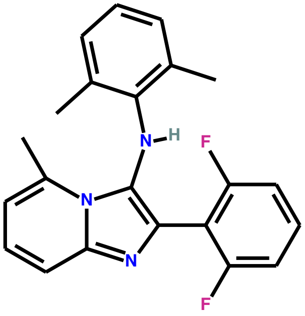
Miscellaneous: FXR_1, FXR_2, FXR_3, FXR_5, FXR_18, FXR_34

Figure S1. Chemical structures of the entire FXR dataset, containing 102 ligands used for ranking prediction.

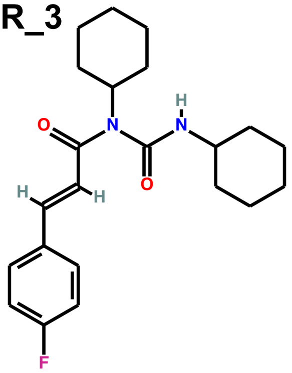
FXR_1



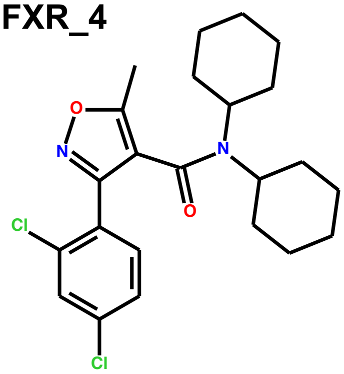
FXR_2



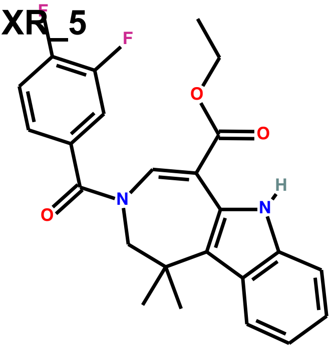
FXR_3



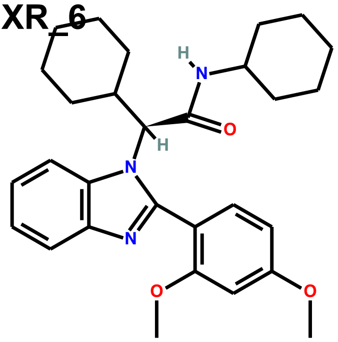
FXR_4



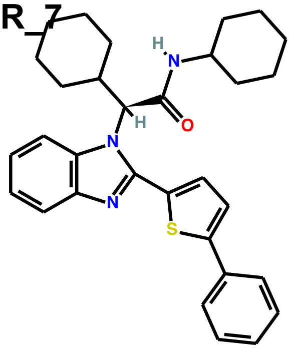
FXR_5



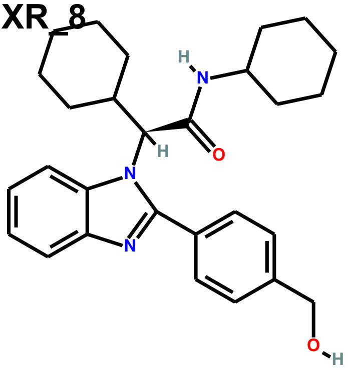
FXR_6



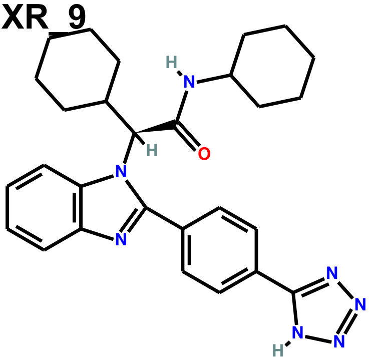
FXR_7



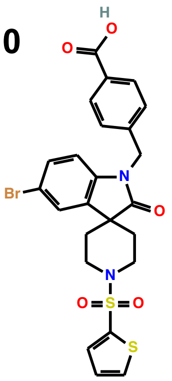
FXR_8



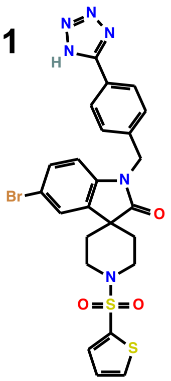
FXR_9



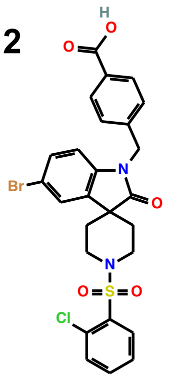
FXR_10



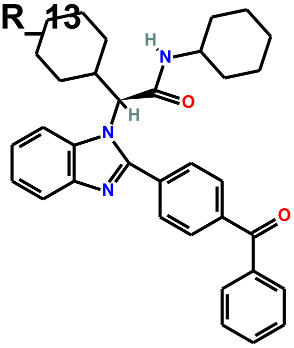
FXR_11



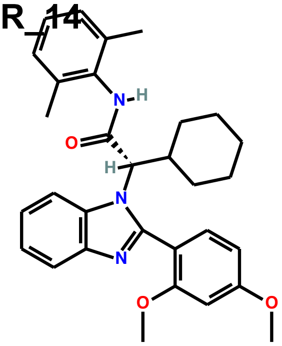
FXR_12



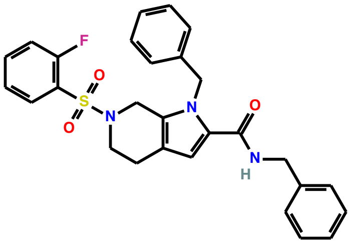
FXR_13



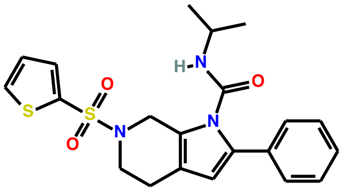
FXR_14



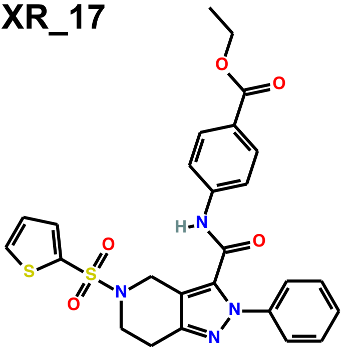
FXR_15



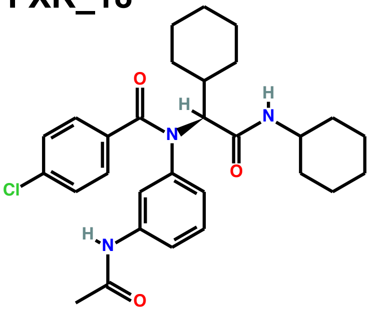
FXR_16



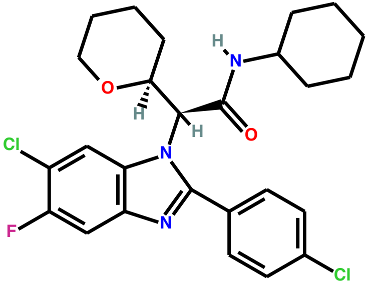
FXR_17



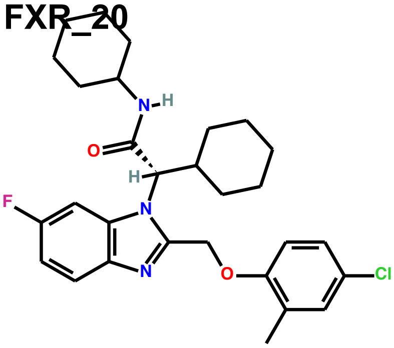
FXR_18



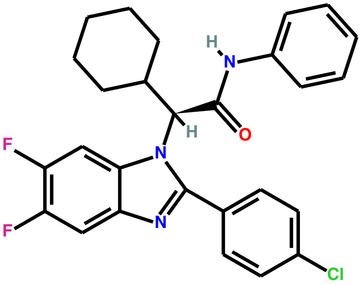
FXR_19



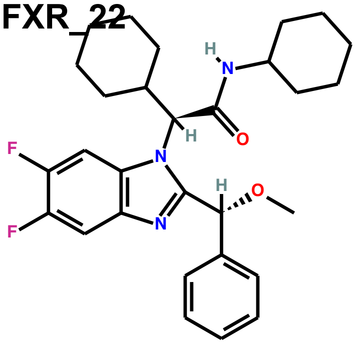
FXR_20



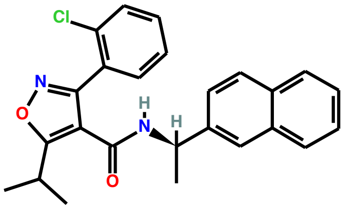
FXR_21



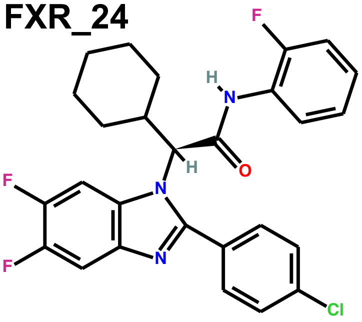
FXR_22



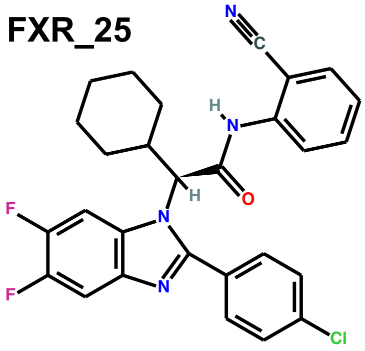
FXR_23



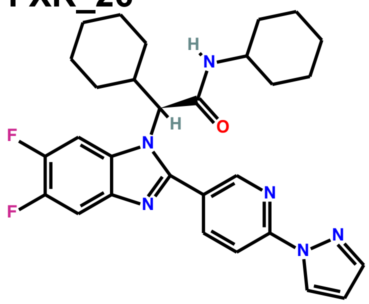
FXR_24



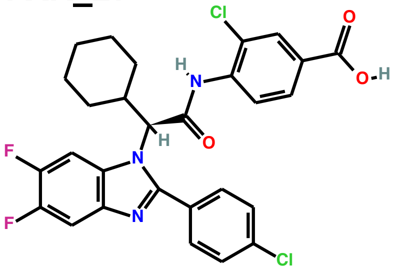
FXR_25



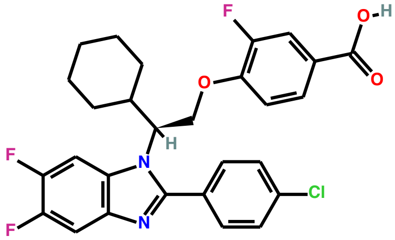
FXR_26



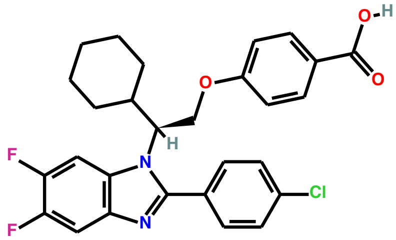
FXR_27



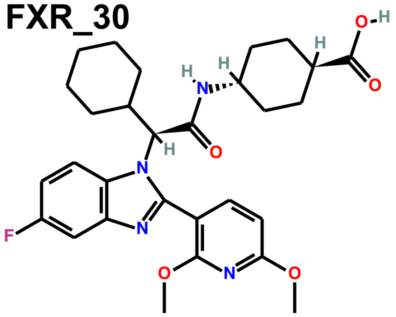
FXR_28



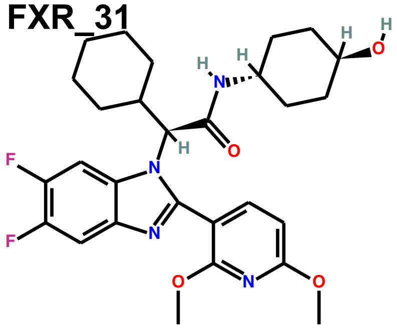
FXR_29



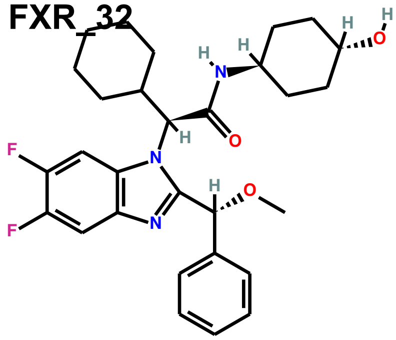
FXR_30



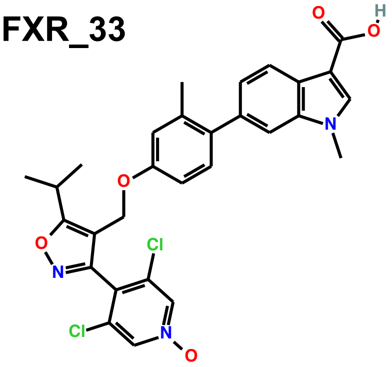
FXR_31



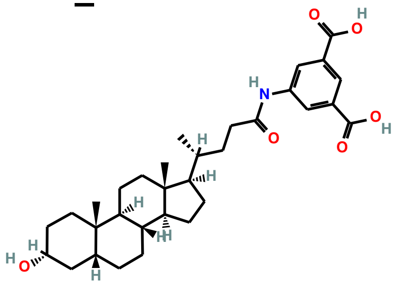
FXR_32



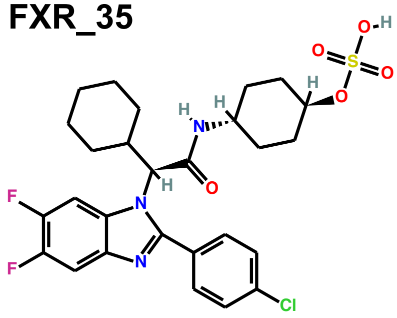
FXR_33



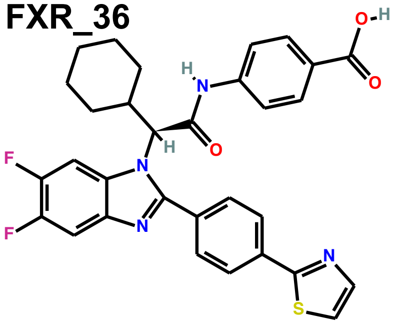
FXR_34



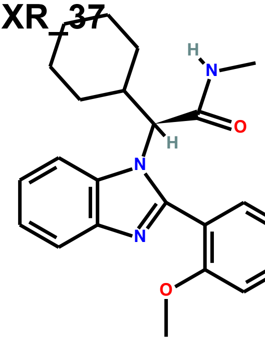
FXR_35



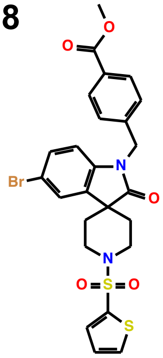
FXR_36



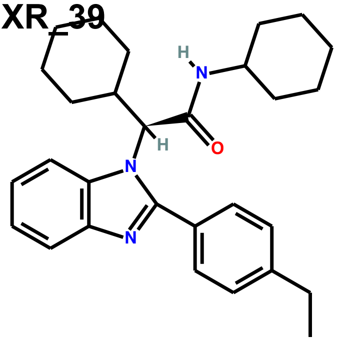
FXR_37



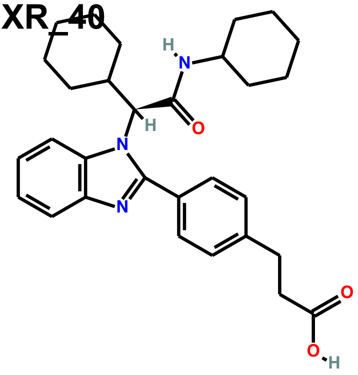
FXR_38



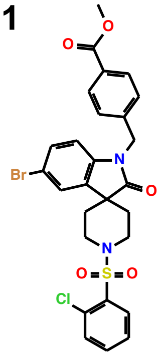
FXR_39



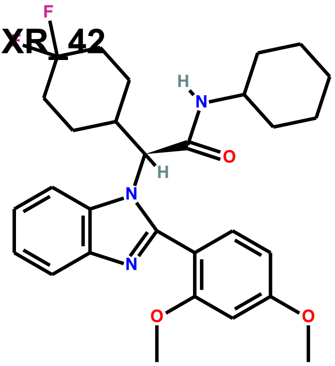
FXR_40



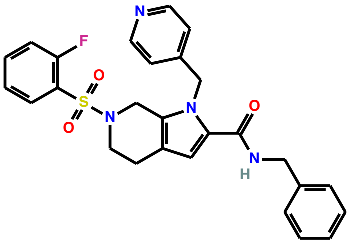
FXR_41



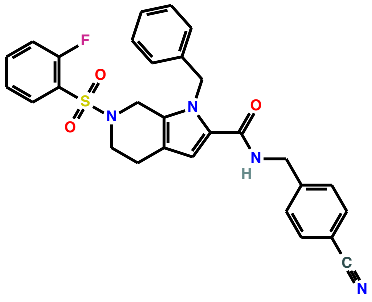
FXR_42



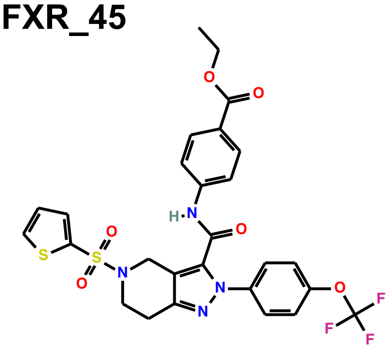
FXR_43



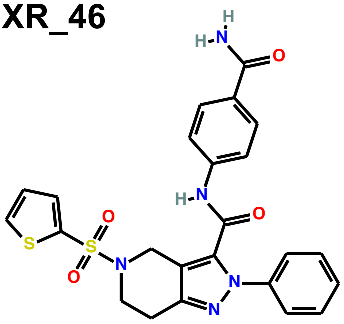
FXR_44



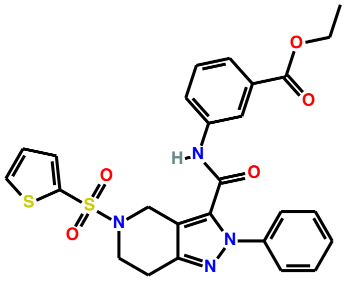
FXR_45



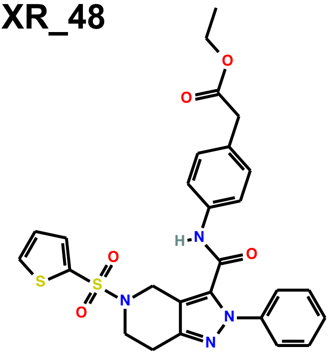
FXR_46



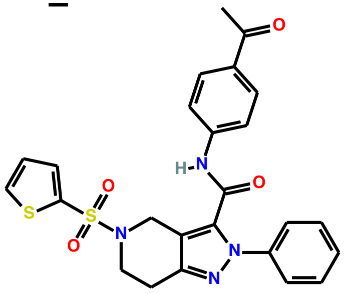
FXR_47



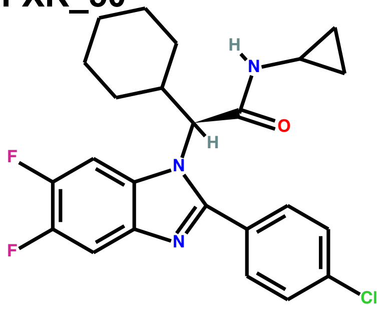
FXR_48



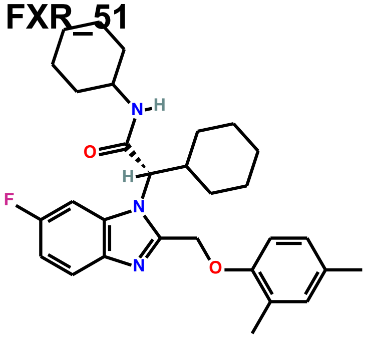
FXR_49



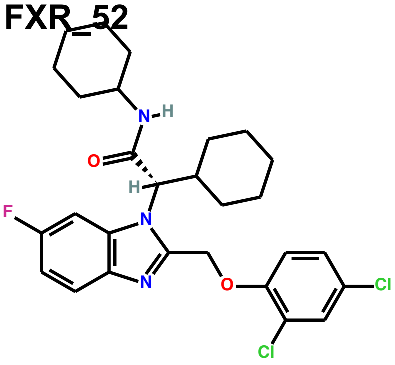
FXR_50



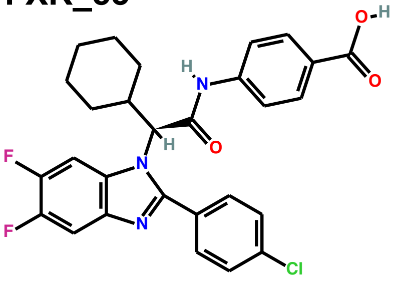
FXR_51



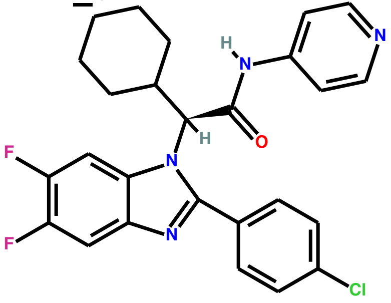
FXR_52



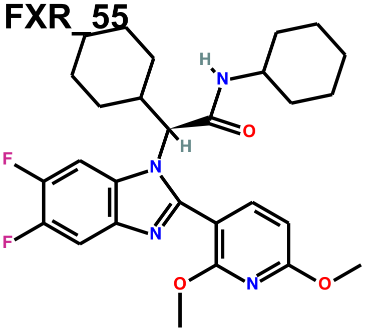
FXR_53



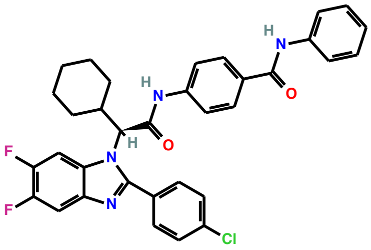
FXR_54



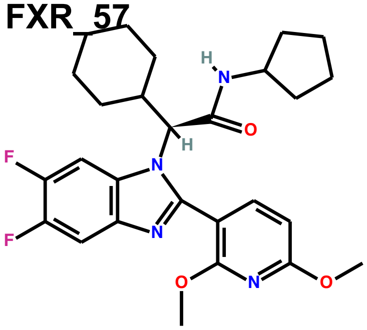
FXR_55



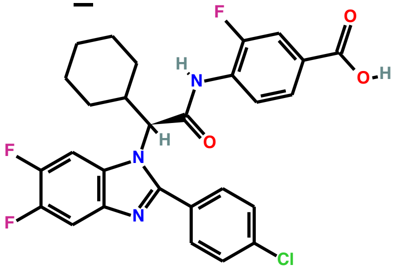
FXR_56



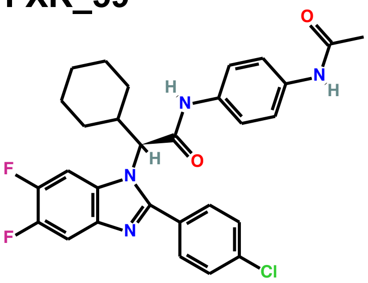
FXR_57



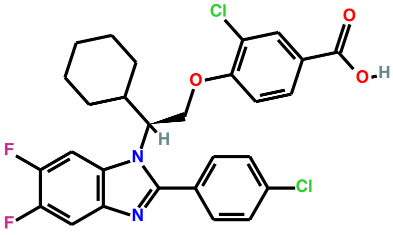
FXR_58



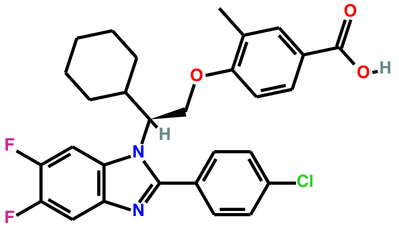
FXR_59



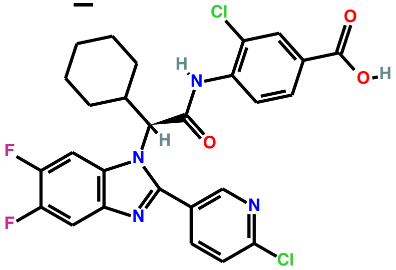
FXR_60



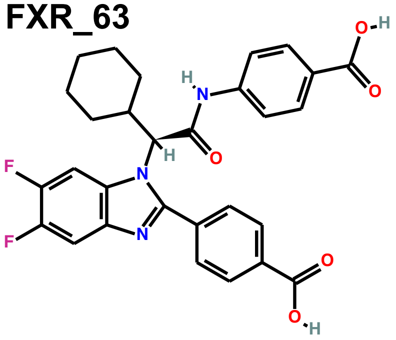
FXR_61



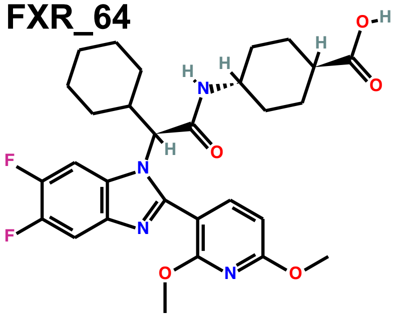
FXR_62



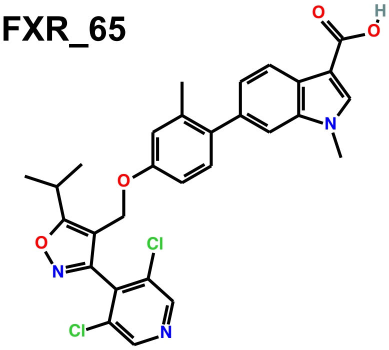
FXR_63



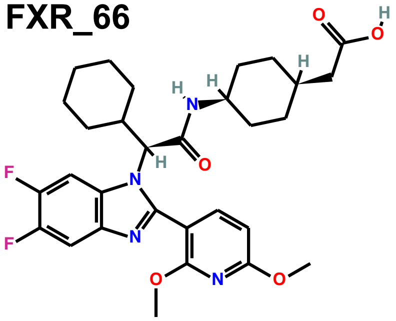
FXR_64



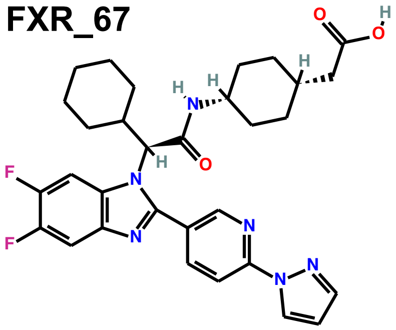
FXR_65



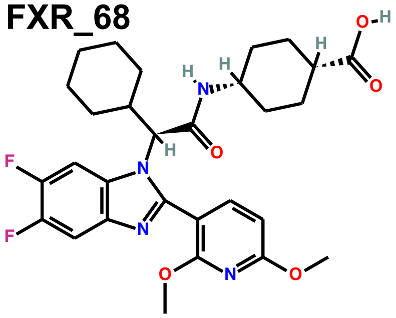
FXR_66



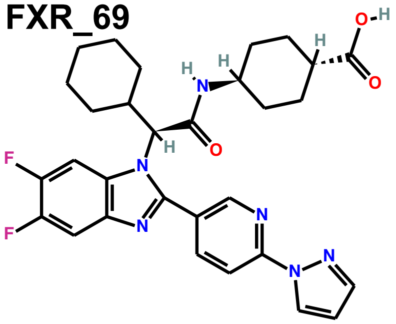
FXR_67



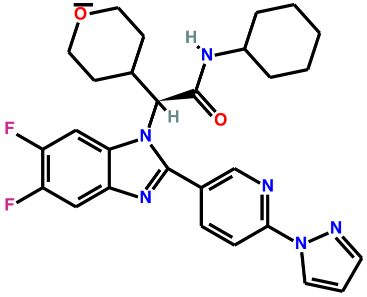
FXR_68



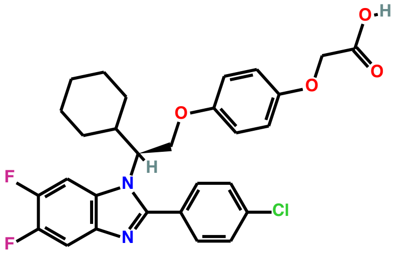
FXR_69



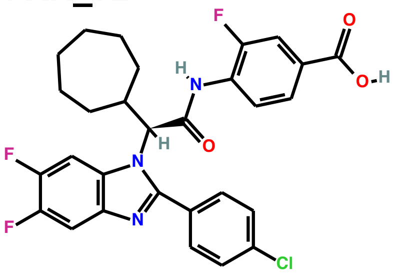
FXR_70



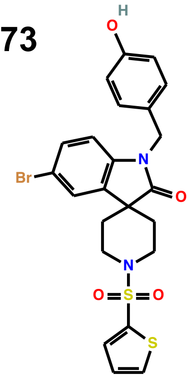
FXR_71



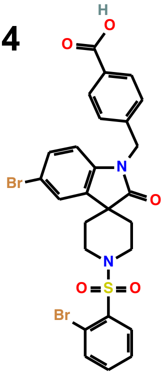
FXR_72



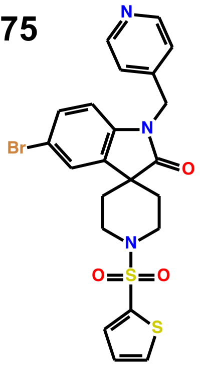
FXR_73



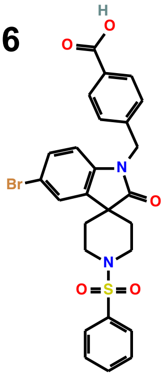
FXR_74



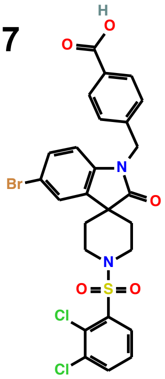
FXR_75



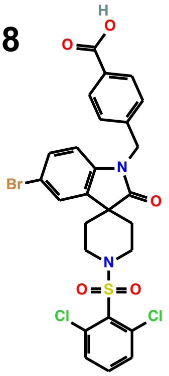
FXR_76



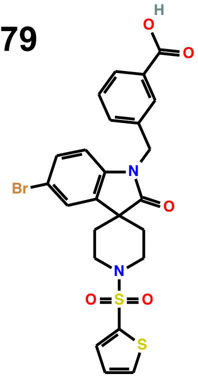
FXR_77



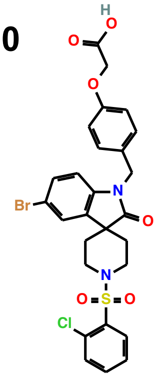
FXR_78



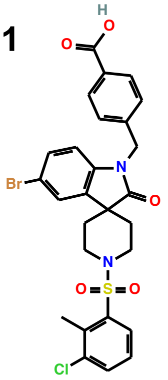
FXR_79



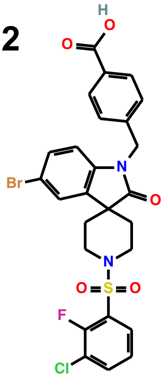
FXR_80



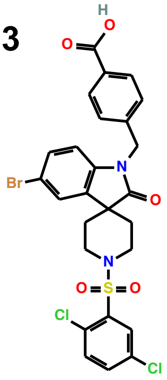
FXR_81



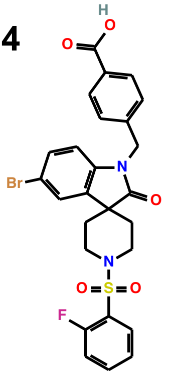
FXR_82



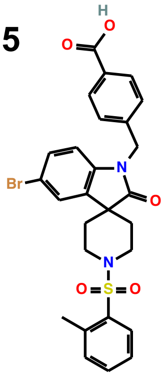
FXR_83



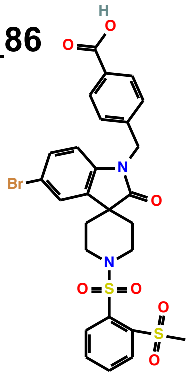
FXR_84



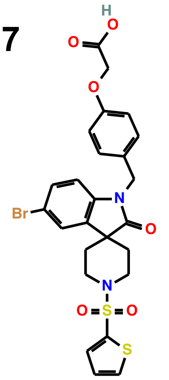
FXR_85



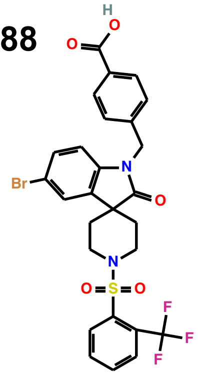
FXR_86



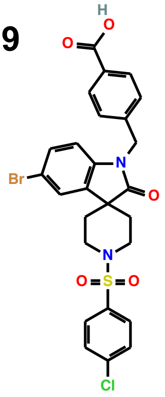
FXR_87



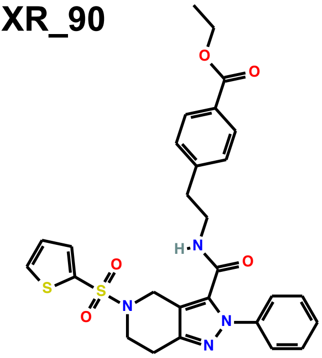
FXR_88



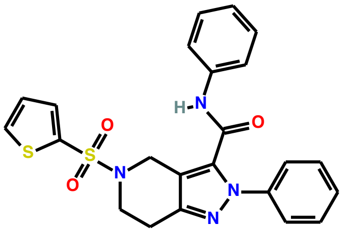
FXR_89



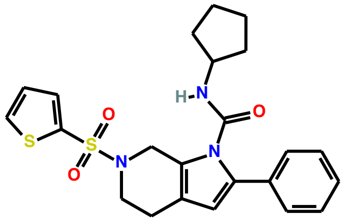
FXR_90



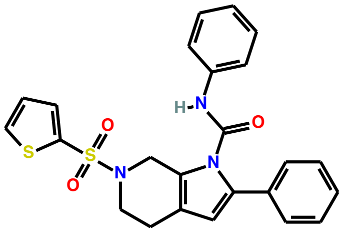
FXR_91



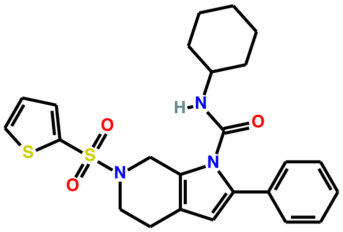
FXR_92



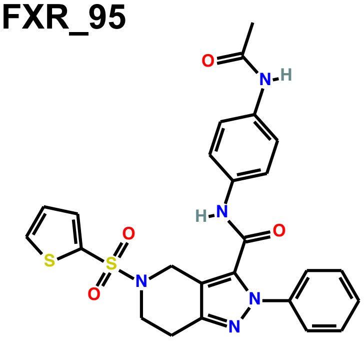
FXR_93



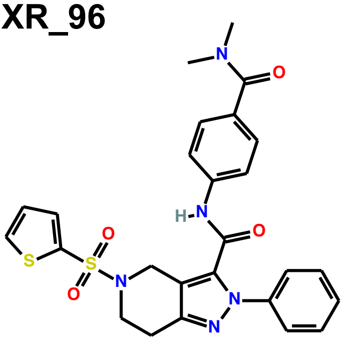
FXR_94



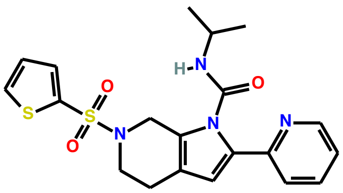
FXR_95



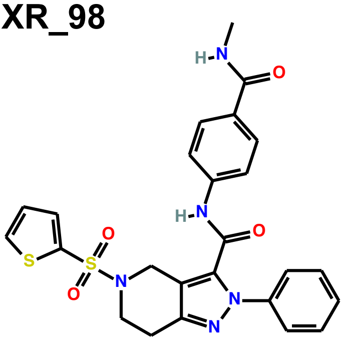
FXR_96



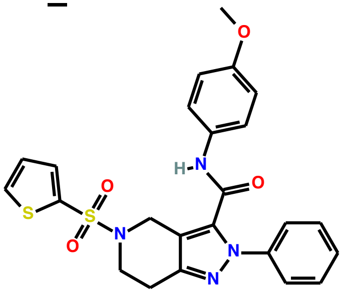
FXR_97



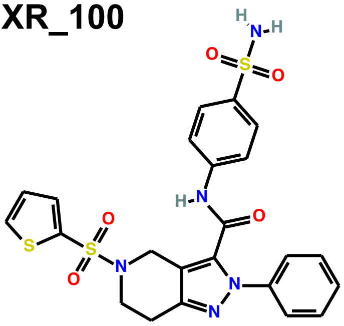
FXR_98



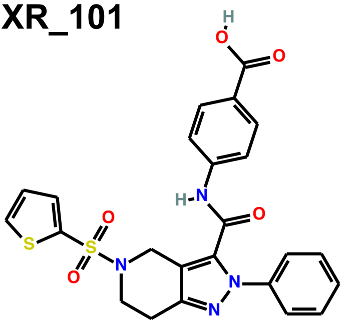
FXR_99



FXR_100



FXR_101



FXR_102

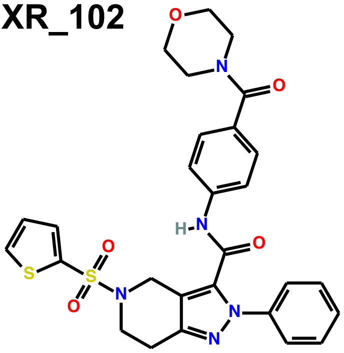


Figure S2. Thermodynamic cycle for the calculation of relative binding affinities of ligands L_1 and L_2 compared with the protein P.

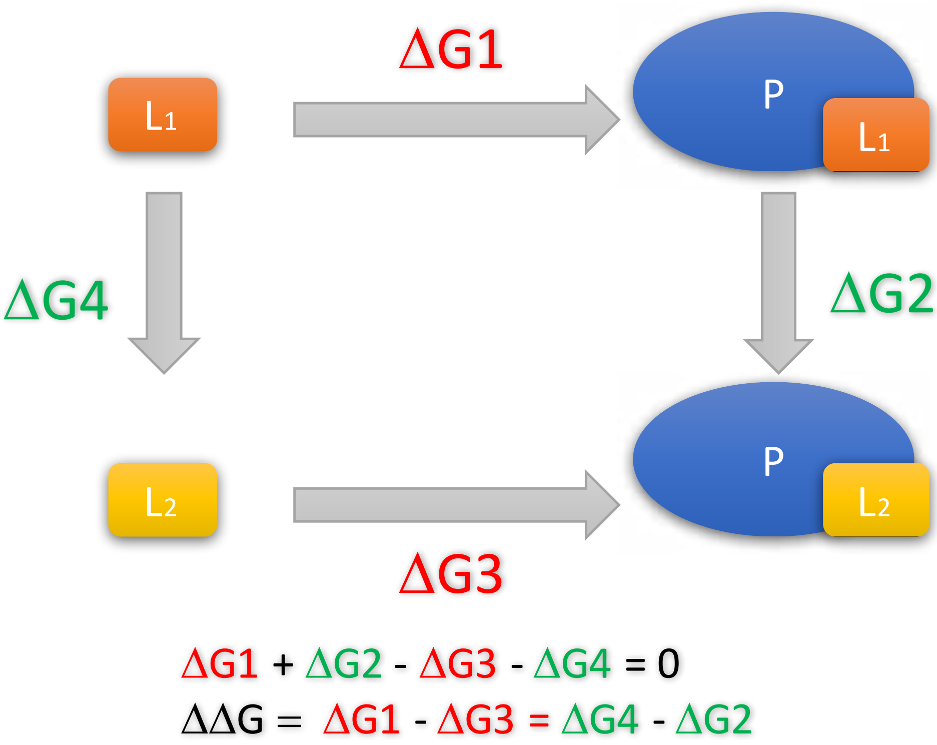


Figure S3. Schematic representation of the system used for the calculation of relative binding affinities of ligands L_1 and L_2 for the protein P, in the case of a charge conserving structural change on the ligand.

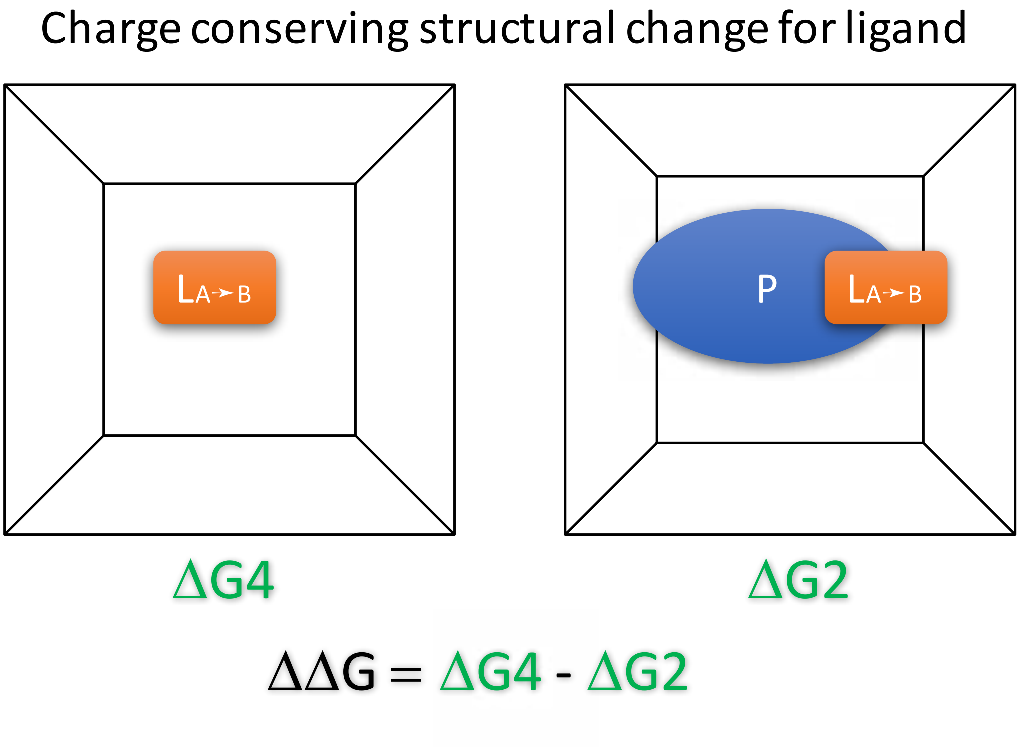


Figure S4. Schematic representation of the system used for the calculation of relative binding affinities of ligands L_1 and L_2 for the protein P, in the case of a charge modifying structural change on the ligand.

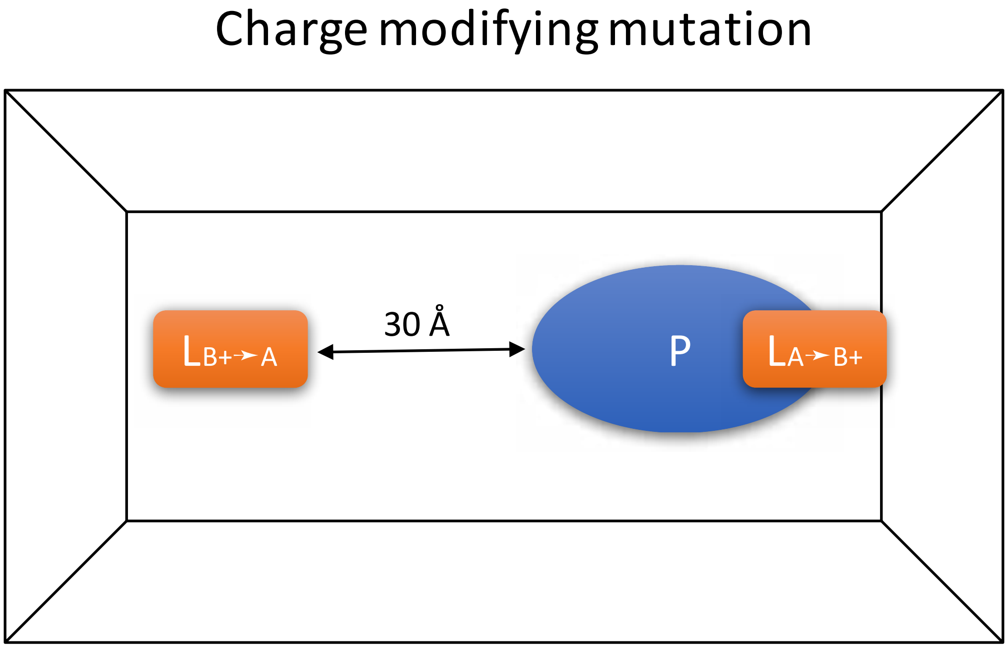
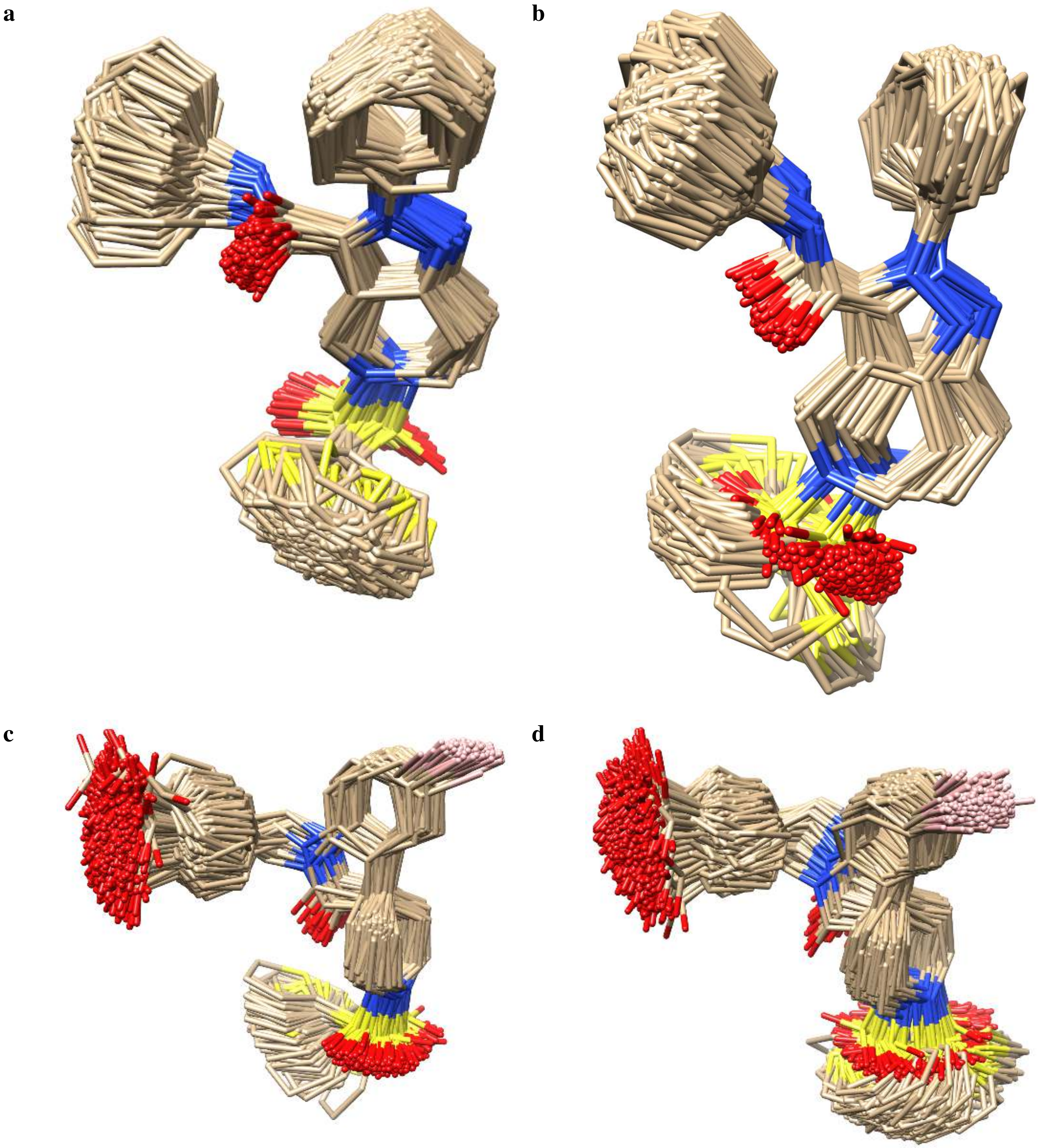


Figure S5. Conformational distribution of ligands FXR_17 (**a**, **b**), FXR_10 (**c**, **d**) and FXR_12 (**e**, **f**), as representative structures for set1 and set2. In each case, all 501 conformers extracted from the 10 ns molecular dynamics simulation of the ligand alone in water, using the OPLS-AA (**a**, **c**, **e**) and AMBER/GAFF (**b**, **d**, **f**) force fields, are represented. Hydrogen atoms are not shown for more clarity.



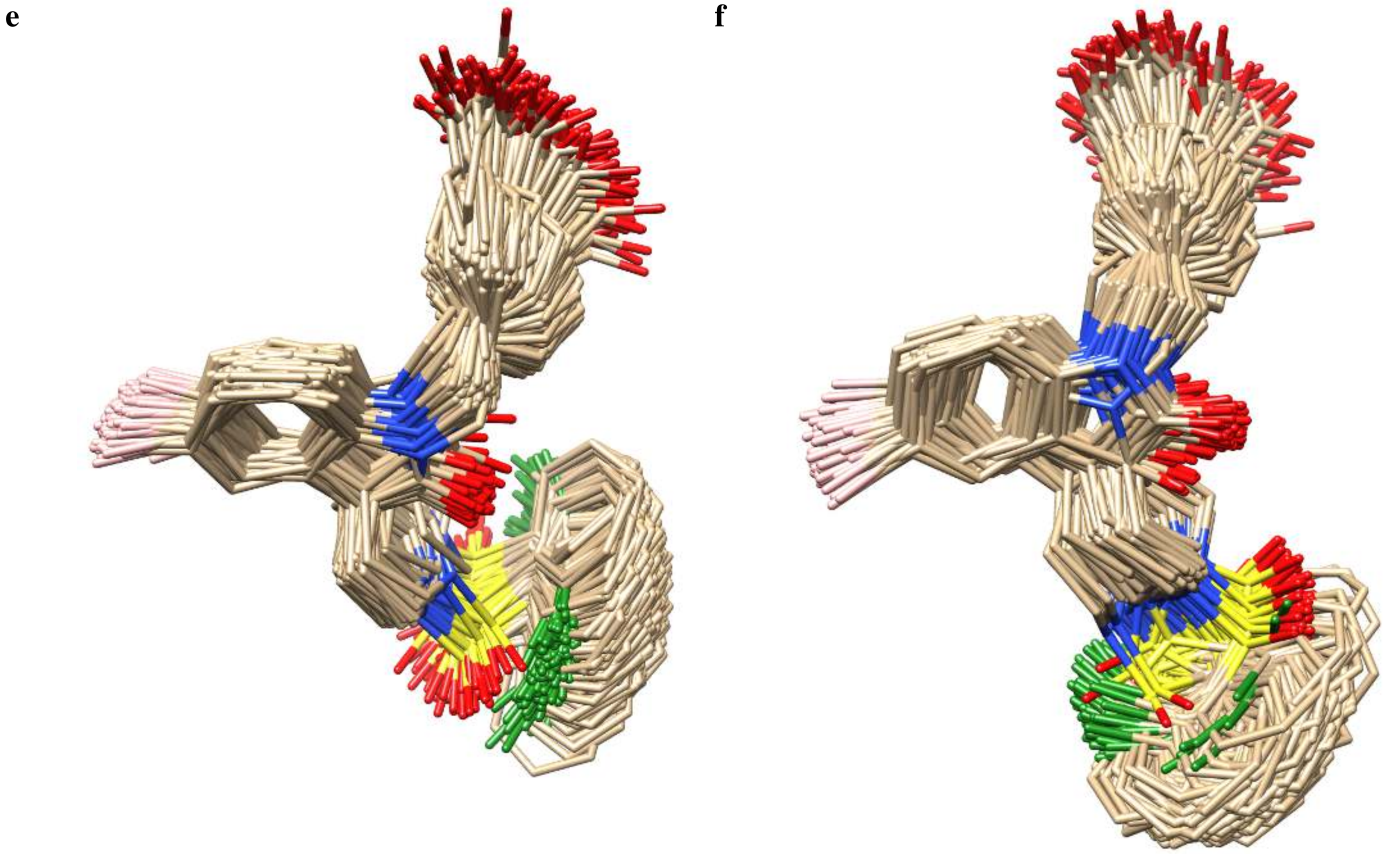

e
f

Table S1. Rank of the best RMSD poses. When no reference structure was available, the score of the first ranked pose was reported for the two submissions, therefore no rank of best RMSD pose is considered.

Ligand	Score of first ranked pose	Score of best RMSD pose	Rank of best RMSD pose (out of 10 poses)
FXR_1	57.25	57.25	–
FXR_2	57.27	57.27	–
FXR_3	59.03	59.03	–
FXR_4	63.35	64.13	6
FXR_5	58.53	62.45	7
FXR_6	49.10	52.44	4
FXR_7	48.07	49.36	4
FXR_8	53.28	54.13	3
FXR_9	43.55	43.55	1
FXR_10	51.19	51.19	–
FXR_11	44.97	44.97	–
FXR_12	47.66	47.66	–
FXR_13	41.27	41.27	1
FXR_14	45.85	46.23	2
FXR_15	49.65	49.65	–
FXR_16	56.47	56.47	–
FXR_17	42.05	42.05	–
FXR_18	54.34	54.34	–
FXR_19	49.24	52.23	9
FXR_20	48.05	49.73	2
FXR_21	46.07	46.26	2
FXR_22	51.95	51.95	1
FXR_23	53.66	57.20	10
FXR_24	44.42	45.46	7
FXR_25	46.74	47.83	7
FXR_26	39.43	39.43	1
FXR_27	39.50	41.40	5
FXR_28	40.42	42.27	7
FXR_29	41.16	43.78	6
FXR_30	41.54	44.58	7
FXR_31	44.30	45.09	4
FXR_32	52.50	53.19	2
FXR_33	40.94	42.95	5
FXR_34	52.87	59.51	10
FXR_35	40.64	40.64	1
FXR_36	32.84	35.97	7
FXR_37	55.97	58.21	10
FXR_38	50.16	50.16	–
FXR_39	51.58	51.66	4
FXR_40	48.89	51.89	5
FXR_41	45.89	45.89	–
FXR_42	45.92	50.25	7
FXR_43	46.24	46.24	–
FXR_44	48.13	48.13	–
FXR_45	39.38	39.38	–
FXR_46	42.54	42.54	–
FXR_47	51.05	51.05	–
FXR_48	43.75	43.75	–
FXR_49	43.65	43.65	–
FXR_50	52.93	54.39	6
FXR_51	48.25	48.74	2
FXR_52	45.76	54.79	3
FXR_53	39.81	41.12	7

FXR_54	46.19	48.11	4
FXR_55	45.06	45.37	4
FXR_56	43.69	43.69	1
FXR_57	47.05	48.29	8
FXR_58	38.19	38.19	1
FXR_59	39.35	41.05	4
FXR_60	41.27	43.93	2
FXR_61	40.38	43.31	5
FXR_62	38.54	38.63	7
FXR_63	40.07	40.55	9
FXR_64	39.42	41.64	4
FXR_65	38.25	40.65	7
FXR_66	41.96	43.83	5
FXR_67	36.46	37.87	2
FXR_68	41.87	44.72	3
FXR_69	33.62	33.66	2
FXR_70	39.28	45.34	6
FXR_71	36.91	38.68	7
FXR_72	38.51	38.77	2
FXR_73	49.38	49.38	–
FXR_74	48.03	48.03	–
FXR_75	50.52	50.52	–
FXR_76	48.20	48.20	–
FXR_77	44.43	44.43	–
FXR_78	46.49	46.49	–
FXR_79	48.64	48.64	–
FXR_80	37.81	37.81	–
FXR_81	44.02	44.02	–
FXR_82	45.46	45.46	–
FXR_83	45.72	45.72	–
FXR_84	47.15	47.15	–
FXR_85	46.19	46.19	–
FXR_86	44.59	44.59	–
FXR_87	41.90	41.90	–
FXR_88	41.68	41.68	–
FXR_89	47.94	47.94	–
FXR_90	41.55	41.55	–
FXR_91	48.44	48.44	–
FXR_92	55.18	55.18	–
FXR_93	49.39	49.39	–
FXR_94	53.68	53.68	–
FXR_95	41.40	41.40	–
FXR_96	42.61	42.61	–
FXR_97	55.88	55.88	–
FXR_98	42.50	42.50	–
FXR_99	45.73	45.73	–
FXR_100	40.67	40.67	–
FXR_101	45.12	45.12	–
FXR_102	40.80	40.80	–
Mean rank of best RMSD pose (from 53 values)			**4.68**

Table S2. Tanimoto similarity matrix for the compounds belonging to set1.

Tanimoto	FXR_17	FXR_45	FXR_46	FXR_47	FXR_48	FXR_49	FXR_91	FXR_93	FXR_95	FXR_96	FXR_98	FXR_99	FXR_100	FXR_101	FXR_102	Mean value per compound
FXR_17	1.00	0.91	0.88	0.95	0.93	0.87	0.84	0.68	0.85	0.88	0.88	0.86	0.84	0.94	0.87	0.88
FXR_45	0.91	1.00	0.81	0.88	0.85	0.81	0.78	0.63	0.78	0.81	0.81	0.91	0.77	0.86	0.80	0.83
FXR_46	0.88	0.81	1.00	0.86	0.91	0.97	0.95	0.76	0.94	0.98	0.98	0.83	0.95	0.90	0.94	0.91
FXR_47	0.95	0.88	0.86	1.00	0.93	0.86	0.86	0.69	0.86	0.86	0.86	0.88	0.86	0.90	0.85	0.87
FXR_48	0.93	0.85	0.91	0.93	1.00	0.91	0.90	0.73	0.90	0.90	0.90	0.85	0.89	0.90	0.89	0.89
FXR_49	0.87	0.81	0.97	0.86	0.91	1.00	0.93	0.76	0.93	0.97	0.97	0.83	0.93	0.88	0.92	0.90
FXR_91	0.84	0.78	0.95	0.86	0.90	0.93	1.00	0.78	0.97	0.94	0.94	0.85	0.98	0.86	0.90	0.90
FXR_93	0.68	0.63	0.76	0.69	0.73	0.76	0.78	1.00	0.77	0.76	0.75	0.68	0.77	0.69	0.73	0.75
FXR_95	0.85	0.78	0.94	0.86	0.90	0.93	0.97	0.77	1.00	0.94	0.94	0.85	0.97	0.85	0.89	0.90
FXR_96	0.88	0.81	0.98	0.86	0.90	0.97	0.94	0.76	0.94	1.00	0.98	0.82	0.93	0.89	0.94	0.91
FXR_98	0.88	0.81	0.98	0.86	0.90	0.97	0.94	0.75	0.94	0.98	1.00	0.82	0.93	0.89	0.93	0.91
FXR_99	0.86	0.91	0.83	0.88	0.85	0.83	0.85	0.68	0.85	0.82	0.82	1.00	0.85	0.83	0.80	0.84
FXR_100	0.84	0.77	0.95	0.86	0.89	0.93	0.98	0.77	0.97	0.93	0.93	0.85	1.00	0.86	0.89	0.89
FXR_101	0.94	0.86	0.90	0.90	0.90	0.88	0.86	0.69	0.85	0.89	0.89	0.83	0.86	1.00	0.85	0.87
FXR_102	0.87	0.80	0.94	0.85	0.89	0.92	0.90	0.73	0.89	0.94	0.93	0.80	0.89	0.85	1.00	0.88
Global mean value																**0.88**

Table S3. Tanimoto similarity matrix for the compounds belonging to set2.

Tanimoto	FXR_10	FXR_12	FXR_38	FXR_41	FXR_73	FXR_74	FXR_75	FXR_76	FXR_77	FXR_78	FXR_79	FXR_81	FXR_82	FXR_83	FXR_84	FXR_85	FXR_88	FXR_89	Mean value per compound
FXR_10	1.00	0.85	0.94	0.81	0.87	0.89	0.75	0.90	0.82	0.83	0.95	0.83	0.84	0.83	0.89	0.89	0.87	0.85	0.87
FXR_12	0.85	1.00	0.81	0.94	0.75	0.92	0.64	0.94	0.96	0.97	0.82	0.95	0.96	0.97	0.92	0.92	0.91	0.97	0.90
FXR_38	0.94	0.81	1.00	0.85	0.83	0.84	0.74	0.85	0.78	0.79	0.91	0.80	0.80	0.79	0.84	0.85	0.83	0.81	0.84
FXR_41	0.81	0.94	0.85	1.00	0.71	0.87	0.63	0.89	0.91	0.92	0.78	0.91	0.91	0.92	0.87	0.88	0.86	0.92	0.87
FXR_73	0.87	0.75	0.83	0.71	1.00	0.77	0.75	0.79	0.72	0.73	0.88	0.73	0.73	0.73	0.77	0.77	0.76	0.75	0.78
FXR_74	0.89	0.92	0.84	0.87	0.77	1.00	0.66	0.98	0.89	0.90	0.85	0.90	0.91	0.90	0.96	0.96	0.95	0.92	0.89
FXR_75	0.75	0.64	0.74	0.63	0.75	0.66	1.00	0.67	0.62	0.62	0.74	0.62	0.63	0.62	0.66	0.66	0.65	0.64	0.68
FXR_76	0.90	0.94	0.85	0.89	0.79	0.98	0.67	1.00	0.90	0.92	0.86	0.91	0.92	0.91	0.98	0.98	0.96	0.94	0.91
FXR_77	0.82	0.96	0.78	0.91	0.72	0.89	0.62	0.90	1.00	0.95	0.79	0.93	0.94	0.96	0.89	0.89	0.88	0.94	0.88
FXR_78	0.83	0.97	0.79	0.92	0.73	0.90	0.62	0.92	0.95	1.00	0.80	0.93	0.94	0.96	0.90	0.90	0.89	0.95	0.88
FXR_79	0.95	0.82	0.91	0.78	0.88	0.85	0.74	0.86	0.79	0.80	1.00	0.80	0.81	0.80	0.85	0.85	0.84	0.82	0.84
FXR_81	0.83	0.95	0.80	0.91	0.73	0.90	0.62	0.91	0.93	0.93	0.80	1.00	0.95	0.94	0.90	0.93	0.90	0.95	0.88
FXR_82	0.84	0.96	0.80	0.91	0.73	0.91	0.63	0.92	0.94	0.94	0.81	0.95	1.00	0.95	0.94	0.91	0.91	0.96	0.89
FXR_83	0.83	0.97	0.79	0.92	0.73	0.90	0.62	0.91	0.96	0.96	0.80	0.94	0.95	1.00	0.90	0.90	0.88	0.95	0.88
FXR_84	0.89	0.92	0.84	0.87	0.77	0.96	0.66	0.98	0.89	0.90	0.85	0.90	0.94	0.90	1.00	0.96	0.97	0.92	0.90
FXR_85	0.89	0.92	0.85	0.88	0.77	0.96	0.66	0.98	0.89	0.90	0.85	0.93	0.91	0.90	0.96	1.00	0.97	0.92	0.90
FXR_88	0.87	0.91	0.83	0.86	0.76	0.95	0.65	0.96	0.88	0.89	0.84	0.90	0.91	0.88	0.97	0.97	1.00	0.91	0.89
FXR_89	0.85	0.97	0.81	0.92	0.75	0.92	0.64	0.94	0.94	0.95	0.82	0.95	0.96	0.95	0.92	0.92	0.91	1.00	0.90
Global mean value																			**0.86**