A peer-reviewed version of this preprint was published in PeerJ on 11 August 2016.

Kromann JC, Larsen F, Moustafa H, Jensen JH. 2016. Prediction of pKa values using the PM6 semiempirical method. PeerJ 4:e2335 https://doi.org/10.7717/peerj.2335

Prediction of pKa values using the PM6 semiempirical method

The PM6 semiempirical method and the dispersion and hydrogen bond-corrected PM6-D3H+ method are used together with the SMD and COSMO continuum solvation models to predict pKa values of pyridines, alcohols, phenols, benzoic acids, carboxylic acids, and phenols using isodesmic reactions and compared to published \latin{ab initio} results. The pKa values of pyridines, alcohols, phenols, and benzoic acids considered in this study can generally be predicted with PM6 and \latin{ab initio} methods to within the same overall accuracy, with average mean absolute differences of 0.6 - 0.7 pH units. For carboxylic acids the accuracy (0.7 - 1.0 pH units) is also comparable to \latin{ab initio} results if a single outlier is removed. For primary, secondary, and tertiary amines the accuracy is, respectively, similar (0.5 - 0.6), slightly worse (0.5 - 1.0), and worse (1.0 - 2.5), provided that di- and triethylamine are used as reference molecules for secondary and tertiary amines. When applied to a drug like molecule where an empirical pKa predictor exhibits a large (4.9 pH unit) error, we find that the errors for PM6-based predictions are roughly the same in magnitude but opposite in sign. As a result most of the PM6-based methods predict the correct protonation state at physiological pH, while the empirical predictor does not. The computational cost is around 2-5 minutes per conformer per core processor, making PM6-based pKa prediction computationally efficient enough to be used for high-throughput screening using on the order of 100 core processors.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

Prediction of pKa Values Using the PM6 Semiempirical Method

Jimmy C. Kromann[1], Frej A. N. Larsen[1], Hadeel Moustafa[1], and Jan H. Jensen[1,*]

[1]**Department of Chemistry, University of Copenhagen, Copenhagen, Denmark**
[*]**jhjensen@chem.ku.dk; Twitter @janhjensen**

ABSTRACT

The PM6 semiempirical method and the dispersion and hydrogen bond-corrected PM6-D3H+ method are used together with the SMD and COSMO continuum solvation models to predict pKa values of pyridines, alcohols, phenols, benzoic acids, carboxylic acids, and phenols using isodesmic reactions and compared to published *ab initio* results. The pKa values of pyridines, alcohols, phenols, and benzoic acids considered in this study can generally be predicted with PM6 and *ab initio* methods to within the same overall accuracy, with average mean absolute differences of 0.6 - 0.7 pH units. For carboxylic acids the accuracy (0.7 - 1.0 pH units) is also comparable to *ab initio* results if a single outlier is removed. For primary, secondary, and tertiary amines the accuracy is, respectively, similar (0.5 - 0.6), slightly worse (0.5 - 1.0), and worse (1.0 - 2.5), provided that di- and triethylamine are used as reference molecules for secondary and tertiary amines. When applied to a drug like molecule where an empirical pKa predictor exhibits a large (4.9 pH unit) error, we find that the errors for PM6-based predictions are roughly the same in magnitude but opposite in sign. As a result most of the PM6-based methods predict the correct protonation state at physiological pH, while the empirical predictor does not. The computational cost is around 2-5 minutes per conformer per core processor, making PM6-based pKa prediction computationally efficient enough to be used for high-throughput screening using on the order of 100 core processors.

Keywords: pKa prediction, electronic structure, semiempirical methods, drug design

INTRODUCTION

A large proportion of organic molecules relevant to medicine and biotechnology contain one or more ionizable groups, which means that fundamental physical and chemical properties, such as the charge of the molecule, depend on the pH of the solution via the corresponding pKa values of the molecules. As drug- and material design increasingly is being done through high throughput screens, fast - yet accurate - computational pKa prediction methods are becoming crucial to the design process.

There are several empirical pKa prediction tools, such as ACD pKa DB (ACDLabs, Toronto, Canada), Chemaxon (Chemaxon, Budapest, Hungary), and Epik (Schrödinger, New York, USA), that offer predictions in less than a second and can be used by non-experts. These methods are generally quite accurate but can fail for classes of molecules that are not found in the underlying databases. Settimo et al. (2013) have recently shown that the empirical methods are particularly prone to failure for amines, which represent a large fraction of drugs currently on the market or in development. The underlying databases are not public and it is therefore difficult to anticipate when empirical methods will fail. Furthermore, the user is generally not able to augment the databases for cases where the empirical methods are found to fail.

pKa values can be predicted with significantly less empiricism using electronic structure theory (QM) (for a review see Ho (2014)). The accuracy of these QM-based predictions appear to rival that of the empirical approaches, but a direct comparison to empirical methods on a common set of molecules has not appeared in the literature and most QM-based pKa prediction studies have focused on relatively small sets of simple benchmark molecules. Two notable exceptions are the studies by and Klicić et al. (2002) and Eckert and Klamt (2005) who computed pKa values for sets of drug-like molecules. Klicić et al.

(2002) computed the standard free energy change for

$$BH^+ \rightleftharpoons B + H^+ \quad (1)$$

using B3LYP/cc-pVTZ//B3LYP/6-31G(d), with diffuse functions added to negative functional groups, and the Poisson-Boltzmann continuum solvation model implemented in the Jaguar software package. The gas phase deprotonation stardard free energy is computed without vibrational corrections. The pKa values are computed by

$$pK_a = A\frac{\Delta G^\circ}{RT\ln(10)} + B \quad (2)$$

where A and B are found by a linear fit to experimental pKa values for a training set of 200 molecules. Atomic radii for the ions used in the calculation of solvation free energies were optimized as part of the fitting procedure. When applied to the prediction of pKa values for 16 drug like molecules the mean absolute difference relative to experiment was 0.6 pH units.

Eckert and Klamt (2005) computed the standard free energy change for

$$BH^+ + H_2O \rightleftharpoons B + H_3O^+ \quad (3)$$

using BP/TZVP and the COSMOtherm continuum solvation model. The gas phase deprotonation standard free energy is computed without vibrational corrections and the pKa values are computed using Eq 2 where where A and B are found by a linear fit to experimental pKa values for a training set of 43 amines. Eckert and Klamt (2005) observed that the method systematically underestimates the pKa of secondary and tertiary aliphatic amines by ca 1 and 2 pH units, respectively, so an additional empirical correction is added for these two molecule types. Using this approach the pKa values of 58 drug-like molecules containing one or more ionizable N atoms can be reproduced with a root mean square deviation (RMSD) of 0.7 pH units.

While quite accurate, both methods rely on DFT calculations which are computationally too expensive for routine use in high-throughput screening and design. Semiempirical QM (SQM) methods are many orders of magnitude faster than conventional QM but their application to small molecule pKa prediction has been very limited and have focused mainly indirect prediction using atomic charges (Stewart, 2008; Ugur et al., 2014). The most likely reason for this is that semiempirical methods give significantly worse pKa predictions if used with an arbitrary reference molecule such as H_2O. However, we (Li et al., 2004) and others (Li et al., 1997; Govender and Cukrowski, 2010; Sastre et al., 2012) have shown that a judicious choice of reference molecule is a very effective way of reducing the error in pKa predictions. Here we show that this approach is the key to predict accurate pKa values using PM6 and continuum solvation methods.

COMPUTATIONAL METHODOLOGY

The pKa values are computed by

$$pK_a = pK_a^{ref} + \frac{\Delta G^\circ}{RT\ln(10)} \quad (4)$$

where ΔG° denotes the change in standard free energy for the isodesmic reaction

$$BH^+ + B_{ref} \rightleftharpoons B + B_{ref}H^+ \quad (5)$$

where the standard free energy of molecule X is computed as the sum of the PM6 heat of formation, the rigid rotor, harmonic oscillator (RRHO) free energy correction, and the solvation free energy

$$G^\circ(X) = \Delta H_f(X) + [G^\circ_{RRHO}(X)] + \Delta G^\circ_{solv}(X) \quad (6)$$

In some calculation the $G^\circ_{RRHO}(X)$ term is neglected, which will be indicated by an *. Nominally the standard state for $G^\circ_{RRHO}(X)$ has been corrected to 1 M, but this effect cancels out for isodesmic reactions.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

All energy terms are computed using gas phase geometries. $\Delta H_f(X)$ is computed using either PM6 (Stewart, 2007) or PM6-D3H+ (Kromann et al., 2014) while $\Delta G^{\circ}_{solv}(X)$ is computed using either the SMD (Marenich et al., 2009) or COSMO (Klamt and Schüürmann, 1993) solvation method. The PM6-D3H+ and SMD calculations are performed with the GAMESS program (Schmidt et al., 1993), the latter using the semiempirical PCM interface developed by Steinmann et al. (2013), while the COSMO calculations are performed using MOPAC2012. The pKa of dimethylamine is also calculated at the M05-2X/6-311++G(d,p)/SMD* level of theory using Gaussian09 (Frisch et al., 2014). Geometry optimizations were performed in GAMESS using a convergence criterion of 5×10^{-4} au, which is five times higher than default. In cases where imaginary frequencies were found this criterion was reduced to 1×10^{-4} and, again, to 5×10^{-5}. Structures with imaginary frequencies found using the lowest convergence criterion were then ignored when computing the PM6-D3H+/SMD pKa values.

A conformational search was done for each molecule using Open Babel (O'Boyle et al., 2011) version 2.3.90 compiled from their GitHub repository. Conformations was genererated using genetic algorithm and RMSD diversity with the following settings for obabel;

```
obabel start.xyz -O finish.xyz --conformer --nconf 30 --score rmsd --writeconformers
```

Open Babel does not consider C-NH_2 and C-OH bonds to be rotatable so several different start configuration for these sites were prepared manually. Similarly new conformations due to nitrogen inversion for deprotonated secondary amines and protonated and deprotonated tertiary amines, are generated manually were applicable. All start geometries are made available as supplementary material. When computing the pKa values the structures with the lowest free energies ($G^{\circ}(X)$) are chosen.

For compound **1** (Figure 2)) Open Babel failed to find any conformations and Balloon (Vainio and Johnson, 2007) was used for the conformational search instead. The Balloon config file can be found in the supplementary information.

RESULTS AND DISCUSSION

Comparison of pKa values predicted using PM6 and *ab initio* methods

Sastre et al. (2012) have computed the pKa values using isodesmic reactions and a several *ab initio* method for a variety of molecules containing six types of ionizable groups. Table 1 lists the molecules from Sastre et al. (2012) used in this study. The molecules in the first row are the reference molecules (ref) with the corresponding pK_a^{ref} value in parenthesis. Molecules containing chlorine have been eliminated because PM6 calculations for this elements involves *d*-integrals, which have not yet been implemented in GAMESS.

Table 1. List of molecules used in this work. The first row indicates the reference molecules used for each of the functional group and the corresponding experimental reference pKa values.

Pyridines	Alcohols	Carboxylic acids	Amines	Phenols	Benzoic acids
Pyridine (5.23)	Ethanol (15.90)	Acetic acid (4.76)	Ethyl amine (10.63)	Phenol (9.98)	Benzoic acid (4.20)
2-Methylpyridine	Methanol	Formic	Methylamine	p-Cyanophenol	p-Methylbenzoic
3-Methylpyridine	Propanol	Benzoic	Propylamine	m-Cyanophenol	m-Methylbenzoic
4-Methylpyridine	i-Propanol	Hexanoic	i-Propylamine	m-Fluorophenol	p-Fluorobenzoic
2,3-Dimethylpyridine	2-Butanol	Propanoic	Butylamine	p-Fluorophenol	
2,4-Dimethylpyridine	tert-butanol	Pentanoic	2-Butylamine	m-Methylphenol	
3-Fluoropyridine		Trimethylacetic	tert-Butylamine	p-Methylphenol	
3-Cyanopyridine			Trimethylamine	o-Methylphenol	
			Dimethylamine		

Columns 2 - 4 of Table 2 lists mean absolute differences (MAD) and maximum absolute differences (Max AD) relative to experiment for pKa values calculated by Sastre et al. (2012) using B3LYP and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

M05-2X/6-311++G(d,p) as well as the CBS-4B3* composite method (Casasnovas et al., 2010) and the SMD solvation method. The data shows that all three *ab initio* methods perform roughly equally well, with all three methods giving a MAD below 1 pH unit, with the exception of alcohols where the MAD ranges from 1.0 to 1.3 pH units. The Max ADs are lowest for amines (0.6 - 0.8 pH units) and highest for alcohols (2.3 - 2.9 pH units).

Table 2. Mean absolute differences (MADs) and maximum absolute difference (Max AD) of predicted pKa values relative to experimental values for the molecules listed in Table 1. CBS-4B3*, B3LYP, and M05-2X refer to predictions made by Sastre et al. (2012) using a modified CBS-4B3 composite method and the SMD solvation method, B3LYP/6-311++G(d,p)/SMD and M05-2X/6-311++G(d,p)/SMD, respectively. The ”*”s in the last three columns indicate that the rigid rotor-harmonic oscillator free energy term is neglected.

	CBS-4B3*/ SMD	B3LYP/ SMD	M05-2X/ SMD	PM6-D3H+/ SMD	PM6-D3H+/ SMD*	PM6/ SMD*	PM6/ COSMO*
Amines							
MAD	0.2	0.4	0.3	1.2	1.2	1.3	0.7
Max AD	0.6	0.8	0.7	3.9	4.0	4.1	1.9
MAD**	0.2	0.4	0.3	0.5	0.6	0.6	0.6
Max AD**	0.6	0.8	0.7	1.2	1.4	1.4	1.4
Carboxylic acids							
MAD	0.7	0.7	0.6	1.4	1.3	1.2	1.0
Max AD	1.1	1.5	1.3	3.5	3.3	3.3	2.3
Pyridines							
MAD	0.5	0.6	0.6	0.2	0.3	0.3	0.4
Max AD	0.8	1.0	1.0	0.4	0.4	0.5	1.0
Alcohols							
MAD	1.3	1.0	1.3	0.7	0.8	0.8	0.8
Max AD	2.8	2.3	2.9	1.7	1.9	1.8	1.9
Phenols							
MAD	0.6	0.9	0.9	1.3	1.2	1.2	1.3
Max AD	1.7	2.2	2.1	2.4	2.5	2.4	2.4
Benzoic Acids							
MAD	0.4	0.5	0.3	0.3	0.3	0.3	0.3
Max AD	1.1	1.4	0.7	0.7	0.7	0.7	0.7

The fifth column lists the corresponding values computed using PM6-D3H+ with the SMD solvation method. For pyridines, alcohols, phenols, and benzoic acids the overall accuracy of PM6-D3H+ is comparable to the ab initio methods: the MADs are within 0.5 pH units of the *ab initio* values and while the Max ADs range from 0.4 (pyridines) to 2.4 (phenols). For carboxylic acids the results are dominated by a 3.5 pH unit error for trimethylacetic acid, without which the MAD is 1.0 pH units. Thus, different reference molecules should be used to predict pKa values for carboxylic acid groups bonded to secondary and tertiary carbons, using PM6 based methods. For amines the MAD and Max AD is 1.2 and 3.9 pH units, respectively. If only primary amines, which are most similar to the reference compound, are considered the MAD and Max AD drops to 0.5 and 1.2 pH units, respectively. We investigate this point further in the next subsection.

The sixth column of Table 2 lists PM6-D3H+/SMD* pKa values computed with the $G^{\circ}_{RRHO}(X)$ term

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

in Eq 6 removed (denoted by the "*"). In all cases the change in MAD and Max AD is $\leq$ 0.2 and 0.3 pH units, respectively. This small change is not surprising the use of isodesmic reactions and approach has been used in pKa prediction before (Li et al., 2004). Neglecting the dispersion correction (PM6/SMD*) has an even smaller effect on the pKa values, changing the MAD and Max AD by at most 0.1 pH units. It is important to note that the molecules used in this part of the study are relatively small and contain only one functional group. The effect of neglecting vibrational free energies and dispersion corrections may have a bigger effect on the pKa values computed for larger molecules with, for example, intramolecular interactions where both dispersion and vibrational effects can play and important role.

The final column of Table 2 lists PM6/COSMO* pKa values. The pKa values for alcohols, phenols, and benzoic acids are very similar to PM6/SMD with MAD and Max ADs changing by at most 0.1 pH units. In the case of pyridines and carboxylic acids Max AD changes by 0.5 and -1.0 pH units, respectively although this only changes the MAD by at most 0.2 pH units. In the case of pyridines the PM6/SMD* and PM6/COSMO* Max AD is observed for 2,3-dimethylpyridine and 2,4-dimethylpyridine, respectively, while in the case of carboxylic acids the Max AD is observed for trimethylacetic acid. In the case of amines the accuracy of PM6/SMD* and PM6/COSMO* is very similar for primary amines, but the error for di- and trimethylamine is reduced by 1.9 and 2.2 pH units, respectively, by using the COSMO solvation method implemented in MOPAC. To understand these differences we look more closely at dimethylamine and compare the results to corresponding M05-2X/6-311++G(d,p)/SMD calculations, which is one of the methods used by Sastre et al. (2012), but used here without the $G^{\circ}_{RRHO}(X)$ contribution to make the results directly comparable to PM6/SMD* and PM6/COSMO*. Both M05-2X/6-311++G(d,p)/SMD* and PM6/COSMO* yield pKa values for dimethylamine that are virtually identical in accuracy: 10.1 and 11.2 compared to the experimental value of 10.6 pH units. In the case of M05-2X/6-311++G(d,p)/SMD ΔE_{ele} (which replaces ΔH_f in Eq 6) and $\Delta\Delta G^{\circ}_{solv}$ are 11.4 and -10.7 kcal/mol, while the corresponding values for PM6/COSMO* are 3.5 and -4.2 kcal/mol. Taking M05-2X/6-311++G(d,p)/SMD* as a reference, the good performance of PM6/COSMO* is thus a result of significant error cancellation. The corresponding $\Delta\Delta G^{\circ}_{solv}$ computed using PM6/SMD* is -6.8 kcal/mol. While this value is closer to the M05-2X/6-311++G(d,p)/SMD* value it leads to worse error cancellation with the electronic energy contribution and therefore a less accurate pKa prediction (8.2 pH units).

In summary, the pKa values of the pyridines, alcohols, phenols, and benzoic acids considered in this study can generally be predicted with PM6 and *ab initio* methods to within the same overall accuracy, with average MADs for these four functional groups are 0.7 - 0.8 and 0.6 - 0.7 pH units, for the *ab initio* and PM6-based predictions. Similarly, the corresponding Max ADs ranges are 1.6 - 1.7 and 1.3 - 1.5 pH units, respectively. For carboxylic acids the PM6-based results are dominated by 2.3 - 3.5 pH unit errors for trimethylacetic acid, without which the MAD is 0.7 - 1.0 pH units and comparable to the corresponding ab initio results (0.6 - 0.7 pH units). Similarly, for amines the PM6-based results are dominated by a 1.9 - 4.1 pH unit errors for di- and trimethylamine, without which the MAD is 0.5 - 0.6 pH units and comparable to the corresponding *ab initio* results (0.2 - 0.3 pH units). For these simple molecules dispersion corrections and vibrational free energy make a negligible contribution to the predicted pKa values.

Secondary and Tertiary Amines

Here we investigate whether the accuracy of PM6-based predictions of amines can be improved by using different reference molecules for primary, secondary, and tertiary amines. Table 3 lists experimental and predicted pKa values for six secondary and tertiary amines shown in Figure 1 using di- and triethylamine as respective reference. The accuracy of the predicted pKa values for secondary amines is slightly worse compared to primary amines (Tabel 2): the MADs and Max ADs are 0.5 - 1.0 and 1.0 - 1.6 pH units, respectively, compared to 0.5 - 0.6 and 1.2 - 1.4 pH units. The lowest MAD and Max AD is observed for PM6/COSMO*. The contributions of vibrational and dispersion effects are larger than for primary amines, with respective changes of up to 0.8 and 0.9 pH units - both observed for diallylamine. This is presumably due to the fact that the secondary amines are structurally more different from the reference (diethylamine) than for the primary amines. For example, if piperidine is taken as a reference for the prediction of the pKa of morpholine and piperazine then the effects of vibrations and dispersion contribute at most 0.1 pH units. For the SMD-based predictions the lowest MAD is observed for PM6-D3H+ without vibrational

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

contributions.

Table 3. Predicted pKa values for the secondary and tertiary amines shown in Figure 1, using di- and triethylamine as a reference, respectively. In the case or piperazine and DABCO the pKa value corresponds to the singly protonated species.

	Exp	PM6-D3H+/ SMD	PM6-D3H+/ SMD*	PM6/ SMD*	PM6/ COSMO*
Secondary amines					
diethylamine	11.1				
morpholine	8.4	7.3	7.8	7.2	7.9
Piperidine	11.2	10.9	11.3	10.8	10.9
Piperazine	9.8	8.8	9.0	8.4	9.1
Pyrrolidine	11.3	11.3	11.1	10.6	11.3
Diallylamine	9.3	8.0	8.7	7.9	8.3
Diisopropylamine	11.0	12.6	12.4	11.7	11.4
MAD		0.9	0.6	1.0	0.5
Max AD		1.6	1.4	1.4	1.0
Tertiary amines					
triethylamine	10.7				
N-methyl morpholine	7.4	4.9	5.8	4.6	7.4
quinuclidine	11.0	8.1	8.7	7.5	9.4
DABCO	8.8	5.1	5.6	4.3	6.7
N-Ethylpyrrolidine	10.4	9.0	9.5	8.6	10.4
Triallylamine	8.3	4.8	6.9	5.2	6.9
Diisopropylmethylamine	10.5	11.8	12.4	11.3	11.5
MAD		2.5	1.9	2.7	1.0
Max AD		3.7	3.2	4.5	2.1

The accuracy of the predicted pKa values for tertiary amines is significantly worse than for primary and secondary amines with MADs and Max ADs of 1.0 - 2.8 and 2.1 - 4.4 pH units, respectively. As observed for secondary amines the lowest and next-lowest MAD is observed for PM6/COSMO and PM6-D3H+/SMD*. For these two methods the largest error is observed for DABCO: 3.2 and 2.1 pH units for PM6-D3H+/SMD* and PM6/COSMO, respectively. With the exception of diisopropylmethylamine both methods underestimate the pKa values and using the 2 pH unit correction proposed by Eckert and Klamt (2005) reduces the MAD and Max AD to 0.7 and 1.2 for PM6-D3H+/SMD* for these molecules, although the Max AD increases to 3.8 pH units if diisopropylmethylamine is included. Alternatively, the accuracy can be improved by changing the reference molecule. For example, using quinuclidine as a reference, the pKa of DABCO is predicted to within 0.9 and 0.5 pH units using PM6-D3H+/SMD* and PM6/COSMO, respectively.

In summary, the large errors observed for secondary and tertiary amines in Table 2 (i.e. di- and tri-ethylamine) can be decreased by using di- and tri-ethylamine as a reference. The MAD and Max AD for secondary amines (0.5 - 1.0 and 1.0 - 1.6 pH units) are only a little larger than those observed for primary amines (0.5 - 0.6 and 1.2 - 1.4). The MAD and Max AD for tertiary amines (1.0 - 2.5 and 2.1 - 4.5 pH units) are significantly larger than those observed for primary amines and secondary amines. As observed by Eckert and Klamt (2005) the pKa values tend to be underestimated and the error can be reduced somewhat by adding a 2 pH unit correction factor. Alternatively, the error can be reduced for individual molecules by choosing reference molecules with similar structures. PM6/COSMO results in the lowest errors, followed by PM6-D3H+/SMD* for both secondary and tertiary amines.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

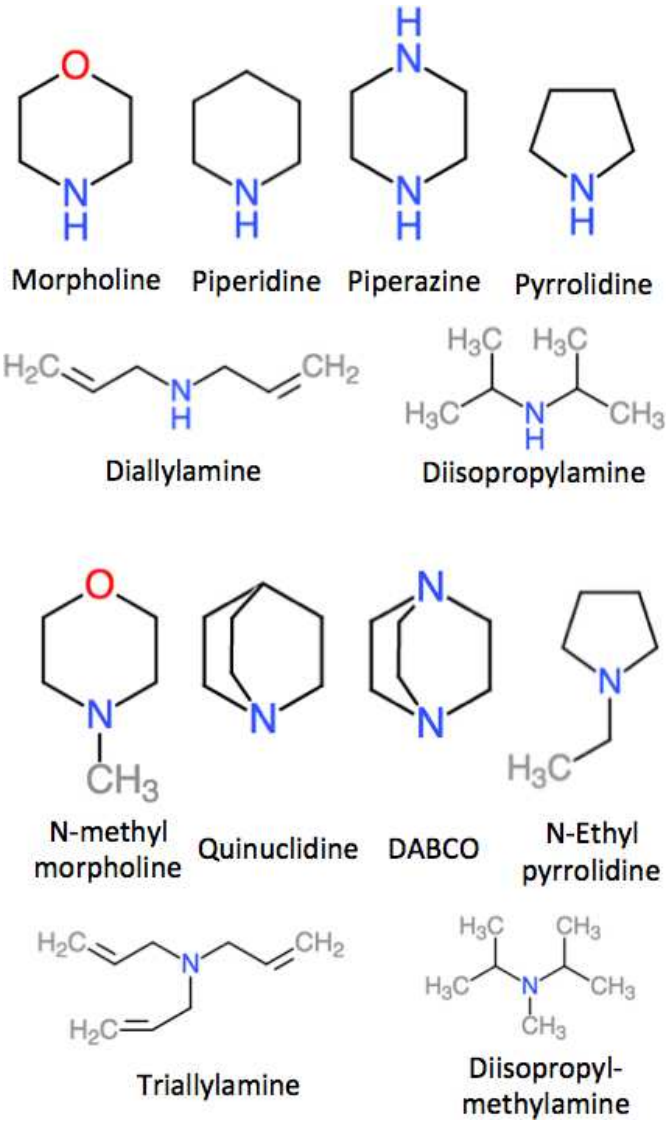

Figure 1. Depiction of the secondary and tertiary amines used used in this study

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

Application to a drug-like molecule

We explore the effect of using different reference molecules further for compound **1** shown in Figure 2. Settimo et al. (2013) have shown that the Chemaxon pKa predictor predicts a pKa value of 9.1 for compound **1**, which is significantly higher than the experimental value of 4.2, i.e. Chemaxon predicts that **1** is charged as physiological pH when, in fact, it is neutral. Table 4 list the pKa values for **1** predicted using PM6-based methodologies using three different reference molecules (cf. Table 2). The absolute errors range from 1.7 to 8.5 with the error being smallest for PM6/SMD using triethylamine as a reference. This agreement may be fortuitous as the error increases for reference molecules more closely related to **1**, while the opposite is seen for PM6-D3H+/SMD(*). Furthermore, the PM6-D3H+/SMD(*) results are consistent with the near systematic pKa-underestimation observed for the tertiary amines in Table 3 and if the 2 pH unit correction suggested by Eckert and Klamt (2005) is used the error decreases to 3.7 - 4.1 pH units when benzylpyrrolidine or heliotridane are used as references. While these error are substantial they lead to the correct qualitative prediction that **1** is neutral at physiological pH. However, whether PM6-based pKa predictions are sufficiently accurate to be useful in drug-design will require a great deal of additional study (see the outlook section for further information).

The computational cost of computing the free energy of a single conformation of **1** is ca 5 minutes on a single Intel Xeon 2.67GHz X5550 core processor with the time roughly equally split between geometry optimization and vibrational frequency calculations. Thus, if the vibrational contributions to the standard free energy can be neglected the time requirement drops to 2-3 minutes per conformer per core processor. For **1** we computed the free energy of roughly 200 conformers. Thus, PM6-based pKa prediction is computationally efficient enough to be used for high throughput screening using on the order of 100 core processors.

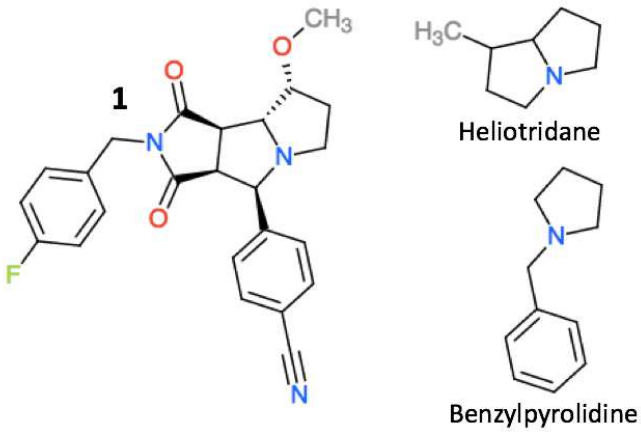

Figure 2. The structure of compound **1**, heliotridane, and benzylpyrolidine

Table 4. Predicted pKa values for compound **1** shown in Figure 2, using triethylamine, heliotridane, and bezylpyrrolidine as a reference, respectively. The pKa values of heliotridane, and bezylpyrrolidine are taken from (Morgenthaler et al., 2007). Note that the latter is estimated and not measured experimentally.

	pK_a^{ref}	PM6-D3H+/ SMD	PM6-D3H+/ SMD*	PM6/ SMD*	PM6/ COSMO*
triethylamine	10.7	-4.3	-3.6	5.9	-0.2
benzylpyrrolidene	8.9	-1.9	-1.5	7.8	0.1
heliotridane	11.4	-1.6	-1.8	8.7	0.7

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

SUMMARY AND OUTLOOK

The PM6 semiempirical method and the dispersion and hydrogen bond-corrected PM6-D3H+ method are used together with the SMD and COSMO continuum solvation models to predict pKa values of pyridines, alcohols, phenols, benzoic acids, carboxylic acids, and phenols using isodesmic reactions. The results are compared to *ab initio* results published by Sastre et al. (2012).

The pKa values of the pyridines, alcohols, phenols, and benzoic acids considered in this study can generally be predicted with PM6 and *ab initio* methods to within the same overall accuracy, with average MADs for these four functional groups of 0.7 - 0.8 and 0.6 - 0.7 pH units, for the *ab initio* and PM6-based predictions. Similarly, the corresponding Max ADs ranges are 1.6 - 1.7 and 1.3 - 1.5 pH units, respectively. For carboxylic acids the PM6-based results are dominated by 2.3 - 3.5 pH unit errors for trimethylacetic acid, without which the MAD is 0.7 - 1.0 pH units and comparable to the corresponding *ab initio* results (0.6 - 0.7 pH units). Similarly, for amines the PM6-based results are dominated by a 1.9 - 4.1 pH unit errors for di- and trimethylamine, without which the MAD is 0.5 - 0.6 pH units and comparable to the corresponding *ab initio* results (0.2 - 0.3 pH units). For these simple molecules dispersion corrections and vibrational free energy make a negligible contribution to the predicted pKa values.

The large errors observed for secondary and tertiary amines in Table 2 (i.e. di- and tri-ethylamine) can be decreased by using di- and tri-ethylamine as a reference. The MAD and Max AD for secondary amines (0.5 - 1.0 and 1.0 - 1.6 pH units) are only a little larger than those observed for primary amines (0.5 - 0.6 and 1.2 - 1.4). The MAD and Max AD for tertiary amines (1.0 - 2.5 and 2.1 - 4.5 pH units) are significantly larger than those observed for primary amines and secondary amines. As observed by Eckert and Klamt (2005) the pKa values tend to be underestimated and the error can be reduced somewhat by adding a 2 pH unit correction factor. Alternatively, the error can be reduced for individual molecules by choosing reference molecules with similar structures. PM6/COSMO results in the lowest errors, followed by PM6-D3H+/SMD* for both secondary and tertiary amines.

When applied to a drug like molecule where the empirical pKa predictor from Chemaxon exhibits a large error, we find that the error is roughly the same in magnitude but opposite in sign. As a result most of the PM6-based methods predict the correct protonation state at physiological pH, while the empirical predictor does not. The computational cost is around 2-5 minutes per conformer per core processor making PM6-based pKa prediction computationally efficient enough to be used for high throughput screening using on the order of 100 core processors.

While the accuracy found for PM6-based pKa prediction is encouraging, the performance needs to be tested for a much larger set of molecules with larger pKa shifts. However, several steps need to be automated to make this feasible. Many conformational search algorithms do not consider C-NH2 and C-OH single bonds rotatable and will leave the start orientation, which is often arbitrarily assigned, unchanged and this can lead to relatively large errors in the predicted pKa values. If such a conformational search algorithm is employed one needs to prepare all possible start conformations for these sites. Similarly, conformational search algorithms do not consider inversion of secondary and tertiary amines meaning that all possible start conformations of deprotonated secondary amines and deprotonated and protonated tertiary amines must be prepared. For molecules with several ionizable sites all relevant combinations of protonation states must be generated and apparent pKa values must be extracted from the calculations. Finally, a library of reference molecules and their experimental pKa values must be created and the most suitable reference molecules must be identified for each ionizable site in the target molecule. Work on all these steps are either currently ongoing or in the planning stages (Jensen, 2015).

Acknowledgments

JCK acknowledges support from the University of Copenhagen.

Supplementary materials

The following are made available at https://dx.doi.org/10.6084/m9.figshare.c.3259513.v1: a list of pKa values used for Table 2, all input and output files, a config file for the Balloon program used in the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

conformational search, various submit and analysis scripts.

REFERENCES

Casasnovas, R., Frau, J., Ortega-Castro, J., Salvà, A., Donoso, J., and Muñoz, F. (2010). Simplification of the CBS-QB3 method for predicting gas-phase deprotonation free energies. *International Journal of Quantum Chemistry*, 110(2):323–330.

Eckert, F. and Klamt, A. (2005). Accurate prediction of basicity in aqueous solution with COSMO-RS. *J. Comput. Chem.*, 27(1):11–19.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, Jr., J. A., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, J. M., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, O., Foresman, J. B., Ortiz, J. V., Cioslowski, J., and Fox, D. J. (2014). *Gaussian09 Revision D.01*. Gaussian Inc. Wallingford CT 2009.

Govender, K. K. and Cukrowski, I. (2010). Density Functional Theory and Isodesmic Reaction Based Prediction of Four Stepwise Protonation Constants as log K H (n) , for Nitrilotriacetic Acid. The Importance of a Kind and Protonated Form of a Reference Molecule Used. *J. Phys. Chem. A*, 114(4):1868–1878.

Ho, J. (2014). Predicting pKa in Implicit Solvents: Current Status and Future Directions. *Aust. J. Chem.*, 67(10):1441.

Jensen, J. H. (2015). High Througput pKa Prediction Using Semi Empirical Methods. *arXiv:1512.00701 [physics]*. arXiv: 1512.00701.

Klamt, A. and Schüürmann, G. (1993). COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *Journal of the Chemical Society, Perkin Transactions 2*, (5):799–805.

Klicić, J. J., Friesner, R. A., Liu, S.-Y., and Guida, W. C. (2002). Accurate Prediction of Acidity Constants in Aqueous Solution via Density Functional Theory and Self-Consistent Reaction Field Methods. *The Journal of Physical Chemistry A*, 106(7):1327–1335.

Kromann, J. C., Christensen, A. S., Steinmann, C., Korth, M., and Jensen, J. H. (2014). A third-generation dispersion and third-generation hydrogen bonding corrected PM6 method: PM6-D3H+. *PeerJ*, 2:e449.

Li, G.-S., Ruiz-López, M. F., and Maigret, B. (1997). Ab Initio Study of 4(5)-Methylimidazole in Aqueous Solution. *J. Phys. Chem. A*, 101(42):7885–7892.

Li, H., Robertson, A. D., and Jensen, J. H. (2004). The determinants of carboxyl pKa values in turkey ovomucoid third domain. *Proteins*, 55(3):689–704.

Marenich, A. V., Cramer, C. J., and Truhlar, D. G. (2009). Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B*, 113(18):6378–6396.

Morgenthaler, M., Schweizer, E., Hoffmann-Röder, A., Benini, F., Martin, R., Jaeschke, G., Wagner, B., Fischer, H., Bendels, S., Zimmerli, D., Schneider, J., Diederich, F., Kansy, M., and Müller, K. (2007). Predicting and Tuning Physicochemical Properties in Lead Optimization: Amine Basicities. *ChemMedChem*, 2(8):1100–1115.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33.

Sastre, S., Casasnovas, R., Muñoz, F., and Frau, J. (2012). Isodesmic reaction for pK a calculations of common organic molecules. *Theor Chem Acc*, 132(2).

Schmidt, M. W., Baldridge, K. K., Boatz, J. A., Elbert, S. T., Gordon, M. S., Jensen, J. H., Koseki, S., Matsunaga, N., Nguyen, K. A., Su, S., and others (1993). General atomic and molecular electronic structure system. *Journal of Computational Chemistry*, 14(11):1347–1363.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016

Settimo, L., Bellman, K., and Knegtel, R. M. A. (2013). Comparison of the Accuracy of Experimental and Predicted pKa Values of Basic and Acidic Compounds. *Pharm Res*, 31(4):1082–1095.

Steinmann, C., Bl\a edel, K. L., Christensen, A. S., and Jensen, J. H. (2013). Interface of the polarizable continuum model of solvation with semi-empirical methods in the GAMESS program. *PloS one*, 8(7):e67725.

Stewart, J. J. P. (2007). Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling*, 13(12):1173–1213.

Stewart, J. J. P. (2008). Application of the PM6 method to modeling proteins. *Journal of Molecular Modeling*, 15(7):765–805.

Ugur, I., Marion, A., Parant, S., Jensen, J. H., and Monard, G. (2014). Rationalization of the p K a Values of Alcohols and Thiols Using Atomic Charge Descriptors and Its Application to the Prediction of Amino Acid p K a 's. *Journal of Chemical Information and Modeling*, 54(8):2200–2213.

Vainio, M. J. and Johnson, M. S. (2007). Generating conformer ensembles using a multiobjective genetic algorithm. *Journal of Chemical Information and Modeling*, 47(6):2462–2474.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2075v1 | CC-BY 4.0 Open Access | rec: 25 May 2016, publ: 25 May 2016