Volume 1 | Number 1 | Jan 2013 | Pages 1–100
PCCP
Physical Chemistry Chemical Physics
www.rsc.org/pccp
ROYAL SOCIETY OF CHEMISTRY

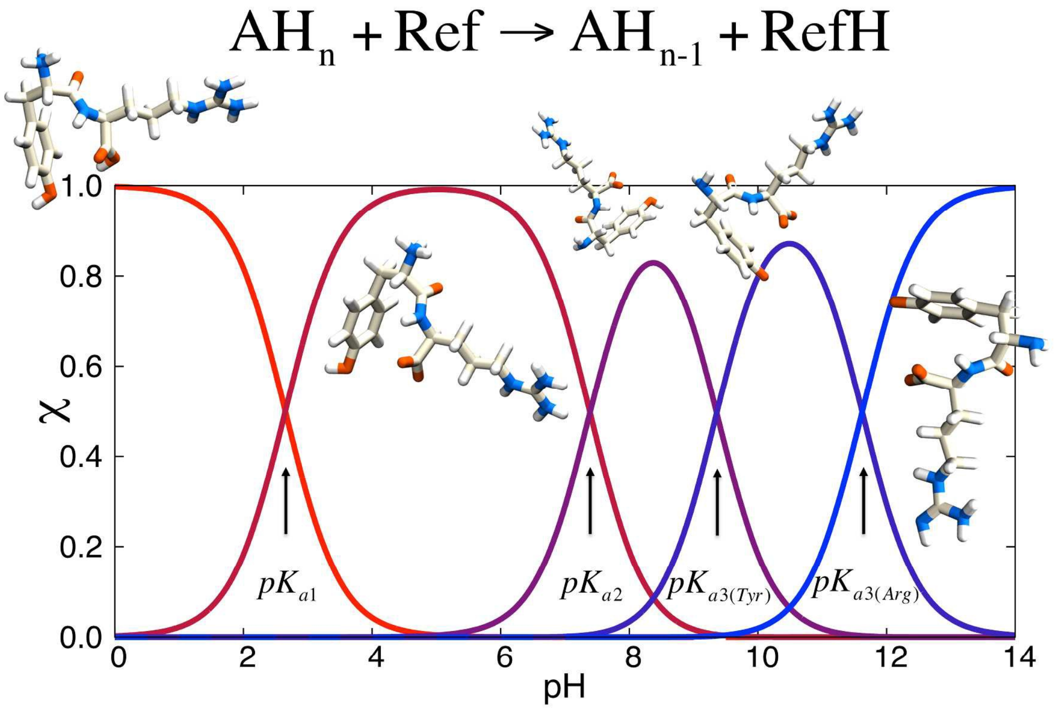
Isodesmic reaction for accurate theoretical pK_a calculations of amino acids and peptides

S. Sastre,[a] R. Casasnovas,[b] F. Muñoz,[a,c] and J. Frau[a,c]

Keywords: pK_a calculation, isodesmic reaction, DFT, continuum solvent model, amino acids

Abstract

Theoretical and quantitative prediction of pK_a values at low computational cost is a current challenge in computational chemistry. We report that the isodesmic reaction scheme provides semi-quantitative predictions (i.e. mean absolute errors of 0.5-1.0 pK_a unit) for the pK_{a1} (α-carboxyl) pK_{a2} (α-amino) and pK_{a3} sidechain groups) of a broad set of amino acids and peptides. This method fills the gaps of thermodynamic cycles for the computational pK_a calculation of molecules that are unstable in gas phase or undergo proton transfer reactions or large conformational changes from solution to gas phase. We also report the key criteria to choose a reference species to make accurate predictions. This method is computationally inexpensive and makes use of standard density functional theory (DFT) and continuum solvent models. It is also conceptually simple and easy to use for researchers not specialized in theoretical chemistry methods.



[a] Mr. Sebastià Sastre, Dr. Francisco Muñoz and Dr. Juan Frau Department de Química, Universitat de les Illes Balears Palma de Mallorca 07122, Spain
E-mail: juan.frau@uib.es
[b] Dr. Rodrigo Casasnovas
Institute of Neuroscience and Medicine INM-9, Institute for Advanced Simulations IAS-5.
Computational Biomedicine, Forschungszentrum Jülich 52428, Germany
[c] Dr. Francisco Muñoz and Dr. Juan Frau
Instituto de Investigación Sanitaria de Palma (IdISPa), 07010, Palma, Spain

1. Introduction

Acid-base reactions are one of the most fundamental and ubiquitous reactions in chemistry, organic, inorganic or biological. In the last years great effort has been devoted by many groups, including ours, to develop computational protocols for the accurate prediction of pK_a values [1-4 and references therein]. Our group in particular has tackled some “difficult” cases for pK_a calculations. These include the pK_a calculation of extremely weak carbon acids [5-7] pK_as of simple amino acids [5] and the combined calculation of pK_a of ligands and stability constants of metal complexes [8]. In all these works our approach pursued the maximum accuracy at the least computational cost. Nevertheless, the quantitative prediction of pK_as is still a challenge in many cases and there is room for improvement for the current methodologies to become practical for general cases.

The current protocols use continuum solvent models because they allow a coarse description of solvent effects at low computational cost. Typically, continuum solvent models are designed to reproduce the experimental solvation energies of a given set of molecules [9-13]. For this reason, the free energy associated to the acidity constant K_a has been typically calculated with a thermodynamic cycle that considers desolvation of the acid species, its deprotonation in gas phase and eventually solvation of the resulting products [1-4].

The pK_a calculation of amino acids is a paradigm of difficult cases for pK_a calculations because the most stable protonation states in gas phase and solution differ. In fact, there are scarce studies reporting theoretical calculations of the α-carboxylic (pK_a1), α-amino (pK_a2) and sidechain groups (pK_a3). Kiani et al [14] determined the pK_a1 and pK_a2 of few amino acids and short peptides with non-polar sidechains with mean absolute deviation (MAD) values of 0.32 and 0.40 units. They used Density Functional Theory (DFT) calculations combined with the PCM continuum solvent model to calculate the free energies of deprotonation of the α-carboxyl and α-amino groups against explicit water molecules. Gupta et al. [15] calculated the pK_{a1} and pK_{a2} of 10 amino acids and also pK_{a3} of 2 amino acids by using the thermodynamic cycle 1 (See Scheme 1 in Theory section) with only continuum solvent models or with explicit water solvent for the first solvation shell and a continuum for the bulk solvent. The latter approach provided the best predictions with mean absolute deviations (MAD) of 3.2, 1.8 and 1.6 units respectively for pK_a1, pK_a2 and pK_a3.

An alternative method for pK_a calculations that avoids gas phase calculations and the related problems is the isodesmic reaction. This method has been proven to provide very accurate pK_a values for a variety of organic functionalities [1,5-8,16]. In a previous work, we used the isodesmic reaction to calculate the pK_a1 and pK_a2 of several nonpolar amino acids [5]. We reported that the isodesmic reaction provides MAD values as low as 0.2 units without needing explicit solvent molecules. Recently, Ho [17] also used the isodesmic reaction to calculate pK_a1 and pK_a2 of several nonpolar amino acids concluding that this approach performs better than thermodynamic cycles.

The main objective of the present work is to provide an exhaustive assessment of the isodesmic reaction for the calculation of pK_a values of the α-carboxylic (pK_a1), α-amino (pK_a2) and the sidechain groups (pK_a3) of any amino acid or peptide. For comparison purposes, thermodynamic cycles as well as the ChemOffice pK_a prediction tool were also used when possible.

2. Theory

Next, three of the most used thermodynamic cycles and the isodesmic reaction are introduced. In all the thermodynamic cycles, the free energy of deprotonation of a given acid species in solution (ΔG_{soln}) is obtained as the sum of the deprotonation free energy of such acid in gas phase (ΔG_{gas}) and the solvation free energy difference between the products and reactants of the deprotonation reaction ($\Delta\Delta G_{solv}$)

$$\Delta \mathrm{G_{soln}} = \Delta \mathrm{G_{gas}} + \Delta\Delta \mathrm{G}_{\mathrm{solv}} \qquad (1)$$

If continuum solvent models are used, the solvation free energy of each species is calculated from equation 2,

$$\Delta \mathrm{G_{solv}} = \mathrm{E_{soln}} - \mathrm{E_{gas}} \qquad (2)$$

where E_{soln} corresponds to the potential energy of the solute in the presence of the reaction field of the continuum solvent and E_{gas} corresponds to the potential energy of the solute in gas phase. It is important to note that in this approach both E_{gas} and E_{soln} are calculated from the geometry of the solute optimized in the gas phase.

The two terms of equation 1 adopt different expressions according to the construction of each thermodynamic cycle. For example, the so-called “direct method” or cycle 1 in this paper (Scheme 1) considers the deprotonation of the acid species (AH^q) in its conjugated base (A^{q-1}) and an isolated proton (H^+).

The free energy of deprotonation in the gas phase (ΔG_{gas}) is given by equation 3.

$$\Delta \mathrm{G_{gas}} = \mathrm{G_{gas}(H^+)} + \mathrm{G_{gas}(A^{q-1})} - \mathrm{G_{gas}\,(AH^q)} + \Delta \mathrm{n\,RT\,ln\,24.46} \qquad (3)$$

Gas phase free energies are calculated for a standard state of 1 atm but the standard state considered for solvation free energies is 1 M in both gas phase and solution. Therefore, the last term accounts for the free energy increment associated to the change of standard state from 1 atm to 1 M in the gas phase. The free energy of the proton in the gas phase ($G_{gas}(H^+)$) is -6.28 kcal/mol at 298 K and 1 atm. This value is the sum of the entropic contribution, 7.76 kcal/mol at 298 K and 1 atm obtained from the Sackur-Tetrode equation [18], and the enthalpy contribution given by the translational motion of a monoatomic particle in the gas phase, 5/2 RT or 1.48 kcal/mol at 298 K and 1 atm.

The solvation free energy difference ($\Delta\Delta G_{solv}$) is calculated as

$$\Delta\Delta \mathrm{G_{solv}} = \Delta \mathrm{G_{solv}(H^+)} + \Delta \mathrm{G_{solv}(A^{q-1})} - \Delta \mathrm{G_{solv}\,(AH^q)} \qquad (4)$$

Equation 4 makes use of the solvation free energy of the proton ($\Delta G_{solv}(H^+)$). Several values of this term have been proposed [4] but currently the accepted value is -265.9 kcal/mol, reported by Tissandier et al [19] and confirmed by Kelly et al [20]. Eventually, the pK_a is calculated as

$$\mathrm{p}K_\mathrm{a} = \frac{\Delta \mathrm{G_{soln}}}{2.303\ \mathrm{RT}} \qquad (5)$$

An inconvenience of cycle 1 is that the solvation free energy of the proton $\Delta G_{solv}(H^+)$ introduces a large uncertainty. However this can be easily circumvented by using cycle 2 (Scheme 2) in which the proton is substituted by the water/hydronium pair H_2O/H_3O^+. In this case, the corresponding gas phase free energies and solvation free energies can be either taken from experiment or calculated [21, 22].

In cycle 2 the gas phase deprotonation energy (ΔG_{gas}) and the solvation free energy increment ($\Delta\Delta G_{solv}$) are given respectively by equations 6 and 7

$$\Delta G_{gas} = G_{gas}(H_3O^+) + G_{gas}(A^{q-1}) - G_{gas}(AH^q) - G_{gas}\ (H_2O) \qquad (6)$$

$$\Delta\Delta G_{solv} = \Delta G_{solv}(H_3O^+) + \Delta G_{solv}(A^{q-1}) - \Delta G_{solv}(AH^q) - \Delta G_{solv}\ (H_2O)(7)$$

As shown in equation 6, the free energy term corresponding to the change of standard state from 1 atm to 1 M vanishes because the number of molecules in both sides of the reaction is equal in cycle 2. The free energy of the reaction in solution given by cycle 2 corresponds to the equilibrium constant of the proton transfer reaction, K_{eq}, which is in turn related to the acidity constant K_a

$$K_{eq} = \frac{[H_3O^+]\,[A^{q-1}]}{[AH^q]\,[\,H_2O]} = \frac{K_a(AH^q)}{[\,H_2O]} \qquad (8)$$

Therefore the pK_a can be calculated as

$$pK_a = \frac{\Delta G_{soln}}{2.303\ RT} - \log[\,H_2O] \qquad (9)$$

The last cycle introduced here, cycle 3, is a variant of cycle 2 in which water and hydronium are substituted by another base (B) and conjugated acid (BH) pair (Scheme 3). Cycle 3 allows B and BH to be chosen in a way that their formal charges equal those of A and AH respectively. This entails a significant advantage over cycles 1 and 2 by improving the accuracy of the pK_a predictions because the cancelation of errors between the solvation free energies increases [1,4,23].

Similar to cycle 2, the resulting free energy calculated with cycle 3 is related to the equilibrium constant of the acid base reaction, which can be expressed in terms of the pK_a values of AH and BH.

$$K_{eq} = \frac{[A^{q-1}][BH^m]}{[AH^q][B^{m-1}]} = \frac{K_a\ (AH^q)}{K_a\ (BH^m)} \qquad (10)$$

Hence cycle 3 provides pK_a values of AH relative to BH. For this reason, BH is also known as reference acid species.

$$pK_a(AH^q) = \frac{\Delta G_{soln}}{2.303\ RT} + pK_a(BH^m\,) \qquad (11)$$

In this cycle, the free energy of the proton transfer reaction is calculated again as the sum of ΔG_{gas} and $\Delta\Delta G_{solv}$, given by equations 12 and 13.

$$\Delta\Delta G_{solv} = \Delta G_{solv}(BH^m) + \Delta G_{solv}(A^{q-1}) - \Delta G_{solv}(AH^q) - \Delta G_{solv}\ (B^{m-1})\ (12)$$

$$\Delta G_{gas} = G_{gas}(BH^m) + G_{gas}(A^{q-1}) - G_{gas}(AH^q) - G_{gas}\ (B^{m-1})\ (13)$$

Because of the way solvation energies are calculated, thermodynamic cycles cannot be used for species that are not stable in gas phase or for species that undergo large conformational changes upon solvation. Recently our group proposed the calculation of pK_a values by avoiding all gas phase calculations [1,5-8,16]. Our protocol is based on an isodesmic reaction, defined in the IUPAC's Gold Book as a reaction in which the types of bonds that are made in forming the products are the same as those which are broken in the reactants (Scheme 4).

This protocol has been successfully used in the pK_a calculation of common organic acids like aliphatic alcohols, carboxylic acids, amines, phenols, benzoic acids and pyridines [1,5,16], weak acids like carbon acids [5-7] and organic groups of ligands in metal complexes [8]. In these studies it is shown that the pK_a values calculated with the isodesmic reaction are as accurate as the best results obtained with thermodynamic cycles. For example, the mean absolute deviation (MAD) of the predicted p*K*$_a$ values lies between 0.5 and 1.0 units for common organic acids [16]. Besides, the isodesmic reaction is more robust and good accuracy is obtained independently of the formal charge of the reference acid species and without using microsolvation of explicit water molecules [5].

Like cycle 3, the isodesmic reaction describes an acid base reaction whose equilibrium constant is given by equation 10 and the p*K*$_a$ of AH is calculated from equation 11 with the use of a reference acid species BH. However, the free energy of the reaction ΔG_{soln} is calculated only with the free energies in solution of the reactants and products (equation 14).

$$\Delta \mathrm{G_{soln}} = \mathrm{G_{soln}}(\mathrm{BH^{m}}) + \mathrm{G_{soln}}(\mathrm{A^{q-1}}) - \mathrm{G_{soln}}(\mathrm{AH^{q}}) - \mathrm{G_{soln}}\,(\mathrm{B^{m-1}}) \quad (14)$$

A rigorous calculation of the free energies requires knowledge of the partition functions. However, these are unknown for a species in solution. Therefore, the best approach would involve statistical methods to sample the relevant states under the given conditions of pressure and temperature. However doing so would also break the philosophy of minimal computational cost of the current protocols. Still it is desirable to include temperature effects in some fashion but it is also true that the partition functions of the harmonic and rigid rotor approximations do not represent the physics of a species in solution. As a compromise, our group proposed the introduction of thermal effects by assuming that the harmonic approximation is valid to represent the vibrational motions in solution and that accounts for the largest thermal effects. Also, we assume that all the remaining contributions from the nuclei motion are similar between reactants and products and do not contribute to ΔG_{soln} as they cancel out in equation 14. Accordingly, the free energy in solution of each species G_{soln} is calculated as

$$\mathrm{G_{soln}} = \mathrm{E_{soln}} + \mathrm{G_{nes}} + \Delta \mathrm{G_{corr\_soln}} \quad (15)$$

where E_{soln} is the potential energy of the solute at 0 K including the electrostatic interactions with the dielectric continuum, ΔG_{corr_soln} corresponds to the thermal effects of vibrational motion at 298 K and G_{nes} includes all the non-electrostatic solute continuum interactions (i.e. dispersion, repulsion and cavitation).

3. Computational details

The calculations were performed with Density Functional Theory (DFT) methods by using the M052X [24], M062X [25] and B3LYP [26,27] exchange correlation functionals. The non-DFT PM6 semiempirical Hamiltonian was also used [28]. Solvent effects were introduced by using the SMD [13,29] continuum solvent model.

In previous publications [5,16] we made use of composite methods like CBS-QB3 and CBS-4B3* [26] for the calculation of deprotonation free energies and p*K*$_a$ values. From those studies we conclude that when used with continuum solvent models like CPCM or SMD such methods provide similar precision for relative free energies, from which p*K*$_a$ values are calculated, than DFT methods at higher computational cost.

The SMD solvent model was parametrized by using M05-2X/6-31+G(d,p) and M05-2X/cc-pVTZ calculations [13,29]. In the present work we used the 6-31+G(d,p) basis set.

All structures were optimized and characterized as energy minima by the absence of imaginary frequencies. All calculations were performed with the Gaussian09 software [30].

4. Results and discussion

4.1. pK_a calculation of α-carboxylic (pK_{a1}) and α-amino groups (pK_{a2}). The isodesmic reaction scheme was used to calculate the pK_a of the α-carboxylic and α-amino groups of 19 of the proteinogenic amino acids. Alanine in the full protonated state was used as a reference species for the calculations of the α-carboxylic group whereas alanine in the zwitterionic state was used as a reference species for the calculations of the α-amino group (Scheme 5).

Table 1 and Table 2 show the absolute errors, mean absolute deviation (MAD) and standard deviation (SD) of the calculated pK_{a1} and pK_{a2} with respect to the experimental values. In terms of absolute errors, most predictions of pK_{a1} and pK_{a2} have errors lower than 1.0 pK_a units for all the DFT functionals and the PM6 Hamiltonian. For pK_{a1} and pK_{a2} the MAD values are approximately 0.5 pK_a units with the exception of the predictions of pK_{a2} with PM6, which shows a higher MAD value of 0.9 units.

The pK_{a1} and pK_{a2} were also calculated for a series of peptides. In this case two reference species were used: the alanine amino acid and the glycylglycine dipeptide. Table 3 and Table 4 show the absolute errors, MAD and standard deviations of pK_{a1} and pK_{a2} calculated with the glygylglycine reference with respect to the corresponding experimental values.

When glycylglycine is used as a reference, the absolute errors of 46 predictions of pK_{a1} are lower than 0.5 pK_a units and 29 are between 0.5 to 1.0 pK_a units. For pK_{a2}, 81 predictions show absolute errors lower than 1.0 pK_a units.

If alanine is used as a reference, the MAD value of pK_{a1} shows little difference with the MAD obtained with glycylglycine as reference species (i.e. ~0.85-1.61 pK_a units, Table S1). However the MAD values of pK_{a2} increase by ~0.5 pK_a units when using alanine and DFT methods (Table S2). Instead, in the case of alanine as a reference the PM6 Hamiltonian improves the MAD by ~0.7 pK_a units (Supporting Information).

In the case of glycylvaline and glycylphenylalanine there are calculated values in the literature to compare with. We find that our calculations show similar errors than those reported by Kiani et al. [14].

The pK_a prediction tool of the ChemOffice software [31] was also used for pK_{a1} and pK_{a2} of the same amino acids and peptides. The MAD and SD of ChemOffice predictions are 3-4 times lower than those of the isodesmic reaction results for pK_{a1} (Table 1, Table 3). For pK_{a2}, the MAD and SD of ChemOffice predictions are similar for amino acids (Table 2) but better for peptides (Table 4). This case shows that well-parametrized methods can provide very good pK_a estimations. However, the isodesmic reaction does not require explicit parametrization. In fact, in the isodesmic reaction, the parametrization is performed *in situ* by the inclusion of the reference species. Therefore any pK_a of any molecule can be calculated, provided that a suitable reference species is chosen.

4.2. Influence of the reference species in pK_{a1} and pK_{a2}. A key point in the pK_a calculation with the isodesmic reaction is the choice of reference species. It is particularly important to show how dependent is the accuracy of the calculations on the reference species.

The calculation of pK_{a1} and pK_{a2} of amino acids were repeated by systematically using all of them one by one as reference species. In most cases the obtained MAD values fall between 0.5 and 1.0 pka units (Table S3). Note that the MAD of pK_{a1} when using histidine as reference species is significantly higher than for the other amino acids (Table S3).

Then, acetic acid and ethylamine were used as, arguably, the simplest reference species for carboxylic acids and amines. The obtained MAD for pK_{a1} and pK_{a2} of amino acids were ~2.6 and ~1.5 pK_a units respectively, while the MAD for pK_{a1} and pK_{a2} of peptides were ~2.1 and ~0.8 pK_a units respectively (Table 5).

To explain these results it should be considered that a good reference species should have similar solute-solvent interactions than the studied acid. Actually, and following the notation of Scheme 5, the interaction of BH^m with the continuum solvent should be as similar as possible to that of AH^q, and likewise for B^{m-1} and A^{q-1}. For this reason an amino acid reference species yields absolute errors of ~0.75 pK_a units for the calculation of pK_{a1} and pK_{a2} of other amino acids (Table S3). On the other hand, if the reference species contains the same functional group but shows significantly different charge distribution, the calculated errors raise. Such is the case when acetic acid and ethylamine are used as reference species for the calculation of pK_{a1} and pK_{a2} of amino acids (Table 5).

It should be noted that we refer to the charge distribution of the reference species because it has been shown that the electrostatic interaction with the continuum model is the major source of error in these calculations. However, the chemical environment should not be neglected. In fact, the large errors obtained for the pK_{a1} of histidine and aspartic acid can be attributed to that factor. In both cases, the sidechain functional group is close enough to interact with the α-carboxyl and, or the α-amino groups. Since these interactions are absent in the alanine reference species and are not constant upon deprotonation of the α-carboxyl group of aspartic acid and histidine, the cancelation of errors is worsened.

In the case of peptides, the use of glycylglycine dipeptide as a reference leads to a decrease of MAD down to .091 and 0.81 units for pK_{a1} and pK_{a2} (Table 5).

If acetic acid is used as a reference for the calculation of pK_{a1}, the MAD of peptides doubles but that of amino acids dramatically multiplies by 5 (Table 5). In the present case, in which the peptides are short and the chosen conformations are extended, this is attributable to the larger separation between the α-carboxylic and α-amino groups in peptides so, the acetic acid reference is a better descriptor of the solute-solvent interactions for a peptide than for an amino acid. For the same reason, ethylamine yields lower errors of pK_{a2} for peptides than for amino acids.

We also analyzed the errors (MAD and SD) of the calculated pK_{a1} and pK_{a2} depending on the global charge and separation between the α-carboxyl and α-amino groups (Table S4). The clearest result is that alanine and glycylglycine are better reference species for amino acids and peptides respectively. Regarding the effect of the global charge, the MAD values of pK_{a1} or pK_{a2} fluctuate around 1 pK_a unit, without clear trend.

These results are in agreement with previous works of our group, which show that rather than the global charge, the local charge distribution of the acid and the reference species are key to obtain low errors [1,5,16]. This is due to the fact that when using continuum solvent models, the electrostatic is the largest solute-solvent interaction and such interactions are computed locally [9-13,29]. The isodesmic reaction scheme exploits the local design of continuum solvent models. This is why there is a clear distinction between the MADs of amino acids and peptides but no clear trend between dipeptides and tripeptides or between differently charged species.

In summary, the two main criteria to consider when choosing the reference acid species are the functional acid group and its neighboring charge distribution as well as other important interactions like hydrogen bonds.

4.3. pK_a calculation of sidechain groups (pK_{a3}). The isodesmic reaction was also used to calculate the pK_a of acid functionalities in the sidechains of peptides (i.e. pK_{a3}), namely: ε-amino of lysine, guanidinium of arginine, sulfhydryl of cysteine, phenol of tyrosine, imidazole of histidine and carboxyl of glutamic and aspartic acids. In these cases the reference species was the acid group of the sidechain in the isolated amino acids lysine, arginine, cysteine, tyrosine, histidine and glutamic acid respectively.

Table 6 reports the absolute errors, MAD and SD of the calculated pK_{a3} with respect to the experimental values. The resulting MADs are below 1.0 pK_a units for all residues but histidines, for the DFT calculations. The PM6 Hamiltonian performs somewhat worse for lysines, cysteines and tyrosines but better for histidines. The absolute error is lower than 1.0 pK_a units for 47 cases independently of the residue type. The remaining 15 cases, in which the errors are larger than 1.0 pK_a unit, are mostly predictions for histidines and then a few lysines, arginines, tyrosines and aspartic acids. Therefore, with the exception of histidines, the isodesmic reaction performs satisfactorily (i.e. errors lower than 1.0 pK_a unit) for all types of residues.

In these cases, the ChemOffice software [31] was also used for the prediction of pK_{a3} (Table 6). The pK_a prediction tool of this software cannot calculate imidazole and thiol groups. However, for the remaining functionalities, the obtained errors are similar to the ones obtained for the Isodesmic reaction scheme for lysines, tyrosines and aspartic or glutamic acids but much worse for arginines.

4.4. Influence of the reference species in pK_{a3}. To evaluate the influence of the reference species in pK_{a3}, non-amino acid references were used for each kind of residue sidechain. The pK_{a3} of lysines, arginines, histidines, cysteines, tyrosine and glutamic or aspartic acids were calculated by using ethylamine, ethylguanidinium, 4-methylimidazole, ethanethiol, phenol and acetic acid respectively. The absolute errors, MAD and SD are reported in Table 7.

For lysines, arginines, tyrosines and glutamic or aspartic acids the MAD values with the amino acid and non-amino acid references differ in less than 0.5 pK_a units. For cysteines, the MAD increase dramatically but given that there are only this two values, it is difficult to make conclusions about the change of reference species in this case.

The case of histidines is worth mentioning as an example in which the reference species is non-intuitive. As can be seen from the absolute errors, MAD and SD in Table 7, 4-methylimidazole is a better reference species than histidine for pK_{a3} of other histidine residues in peptides. We attribute this effect to interactions established between the imidazole ring and the neighboring α-carboxyl group in the histidine reference that are not present in other peptides. In fact, the lowest errors are obtained for those dipeptides in which the histidine is C-terminal and, therefore, such interactions can be also established. Oppositely, when the histidine residue is N-terminal or internal, the obtained errors are higher.

4.5. Thermodynamic cycles for the calculation of pK_{a3}. The use of thermodynamic cycles implies inconveniences for chemical species that are unstable in gas phase or undergo large conformational changes during solvation/desolvation. In a previous work we reported that the thermodynamic cycle approach is not practical for the calculation of pK_{a1} and pK_{a2} of amino acids because the zwitterionic species are unstable in gas phase and there are spontaneous proton transfers between the α-amino and α-carboxyl groups [5]. Our purpose here is to compare the isodesmic reaction with some of the thermodynamic cycles introduced in the theory section for the calculation of pK_{a3}.

The calculation of pK_{a3} was unfeasible for many peptides due to one or more of the following events: a) proton transfer involving the functional group for which pK_{a3} is being calculated. Typically proton transfers from the protonated lysine, histidine, arginine sidechains to the deprotonated α-carboxylate group; b) proton transfer involving other groups than the object of study; c) large conformational changes between gas phase and solution.

Table 8 shows the experimental values of pK_{a3} and the absolute errors, MAD and SD of the peptides for which the calculation was possible with thermodynamic cycles. As can be seen in Table 8, only the isodesmic reaction systematically provides low errors (i.e. approximately 1 pK_a unit on average). Oppositely, none of the thermodynamic cycles is a real alternative for the calculation of pK_as in peptides.

4.6. Conformational sampling and p*K*$_a$ calculations. An important aspect of peptides is their capability to adopt an enormous range of conformations at room temperature by rotations of the backbone dihedrals and the sidechain dihedrals. This entails that each acid functional group in the peptide sequence can potentially establish many intramolecular interactions. For the same reason, the folding state of the peptide also modulates the degree of solvent exposure of the acidic groups. These effects can be major contributions to the pK_a shifts of each residue.

These effects are also expected to gain importance in long peptides and proteins. However, we performed a conformational search on each protonation state of some amino acids and recalculated their pK_{a1}, pK_a and pK_{a3} values. In these cases, the pK_a values were calculated as

$$\mathrm{p}K_{\mathrm{a(AH)}} = \mathrm{p}K_{\mathrm{a(RefH)}} - \log\frac{Q_A Q_{RefH}}{Q_{AH} Q_{Ref}} \qquad (16)$$

Where Q_i stands for the partition function of the species *i*.

$$Q_{\mathrm{i}} = \sum \exp(-E/k_b T) \qquad (17)$$

Between 10 and 18 initial conformations were generated for each protonation state of each molecule with the OpenBabel 2.3.1 software [40]. The conformational search was performed to optimize the root mean square deviation (RMSD) diversity [40]. The resulting conformers were subsequently minimized with the PM6 Hamiltonian in aqueous solution modeled with the SMD method [13, 29]. In some cases, the starting geometry was poor and the geometry optimization lead to a chemical chimera or did not converge. The final number of conformations employed for each amino acid is reported in Table 9. As can be seen in Table 9, using a conformational ensemble also leads to good predictions but does not entail a systematic improvement of pK_{a1}, pK_{a2} and pK_{a3}.

The variations of the potential energies (*E*) within the conformational ensemble, measured as standard deviation, are of ~0.5-1 kcal/mol in most cases. Comparison with the experimental pK_a values, suggest that these fluctuations cancel out in equation 16, as a result of the isodesmic scheme. On other hand, such energy differences fall in the accuracy limit of the employed DFT and semiempirical Hamiltonians and the typical continuum solvent models. The positive message drawn from our results is that for amino acids and other small molecules there is no need of an exhaustive search for the absolute lowest energy conformation.

However, the conformational space of peptides grows rapidly as more residues are in the polymer chain. A conformational search was carried out on the peptide structures to evaluate the effect of the employed conformation on the pK_a calculations. In this case, the OpenBabel 2.3.1 [40] software and the mmff94 force field [41] were employed to generate an ensemble of conformations and choose the most stable one for each protonation state of each peptide. Then, the geometries of the resulting conformers were optimized with the PM6 Hamiltonian and the SMD solvent model.

The mean absolute deviations (MAD) and standard deviations (SD) of pK_{a1}, pK_{a2} and pK_{a3} calculated by using the most stable conformers as initial configurations are reported in Table 10. Comparison with the pK_a values calculated with PM6 on manually generated initial structures shows that the conformational search improved the predictions of pK_{a2} by ~0.2 units but the predictions of pK_{a1} and pK_{a3} worsened by ~1 unit (Table10). The predictions of pK_{a3} worsened for all sidechain functional groups. While the errors for lysines, cysteines and tyrosines increased moderately (i.e. 0.21, 0.25 and 0.7 units respectively), the errors of glutamic/aspartic acids, histidines and particularly arginines increased significantly (i.e. 1.05, 1.42 and 2.49 units) (Table 10).

A deeper analysis of the individual pK_{a3} values of each peptide (Table S5) shows that the errors increase for the largest and more flexible peptides. For instance, the error on the pK_{a3} of the lysine residue in the LysGlu dipeptide is 0.44 pK_a units while the corresponding error on the GlyGlyLysAla is 5.48 pK_a units. Similar trends are observed for histidines, tyrosines, arginines and, although in a less extent, aspartic and glutamic acids (Table S5).

Examination of the peptide structures shows that the most stable generated conformations of large peptides tended to be packed. In this way the non-covalent Van der Waals, Coulomb and H-bond interactions were maximized because OpenBabel performs the conformational search in the gas phase. Therefore, this conformational study cannot validate the peptide structures generated manually as representative conformers in solution. However, it clearly indicates that using a structure that is a bad representative of the most populated conformation in solution can lead to large errors in the pK_a calculations (Table 10). The fact that the pK_a values predicted for peptides (Tables 2-6) show errors similar to those of small molecules (i.e. amino acids) in which conformations are less important, suggests that the structures generated manually are decent representatives of the solution conformations.

We intend to perform further investigations to include solvent effects and conformational sampling for pK_a calculations.

5. Conclusions and further challenges

We have shown that the isodesmic reaction scheme shows significant advantages with respect to thermodynamic cycles, mainly due to the inconveniences resulting from gas phase calculations.

The isodesmic reaction provides accurate results for the pK_a calculation of the α-carboxylic, α-amino groups and sidechains of amino acids and peptides, resulting mean absolute deviations (MAD) of 1 pK_a unit or lower.

The accuracy shows to be robust regarding the choice of the DFT functional. In fact, simpler semiempirical calculations also provide good results. The achieved accuracy in the isodesmic reaction is similar to that of available empirical pK_a estimators. So, even though using a quantum method is slower than other estimators, it is much less limited regarding the chemical composition and structure of the acid of interest.

The choice of the reference species is important for the precision of the pK_a calculations. However, the cancelation of errors intrinsic to the isodesmic reaction allows more flexibility for the choice of such species. As in previous works, we confirm that it is key to choose a reference species for which the local charge distribution neighboring the acid group is similar to that of the studied species. In most cases, this is fulfilled by choosing a molecule with the same functional acid group.

Conformational sampling is not a major source of error in the prediction of pK_a values of small molecules like the amino acids but it can have a large impact on the pK_a calculations as the peptide size increases.

As a final remark, we would like to mention that this scheme is applicable to the calculation of non-aqueous solvents in a simple manner as long as there is an available reference species with a known pK_a value in such solvent. The solvent environment can be changed in many continuum models simply by setting the correspondent static dielectric constant of the desired solvent. However, for this reason dealing with solvent mixtures can be more challenging.

We conclude that the isodesmic reaction is a suitable methodology for the theoretical calculation of pK_a values, especially in those species implying difficulties for thermodynamic cycles. We expect that in the near future this work can be expanded to address more of the current difficulties.

Acknowledgements

This work was cofunded by the Ministerio de Economía y Competitividad (MINECO) and by the European Fund for Regional Development (FEDER) (CTQ2014-55835-R). It was also funded by the Conselleria d'Educació, Cultura i Universitats (Ajuts a accions especials d'R+D AAEE027/2014). The authors are grateful to "Centro de Cálculo de

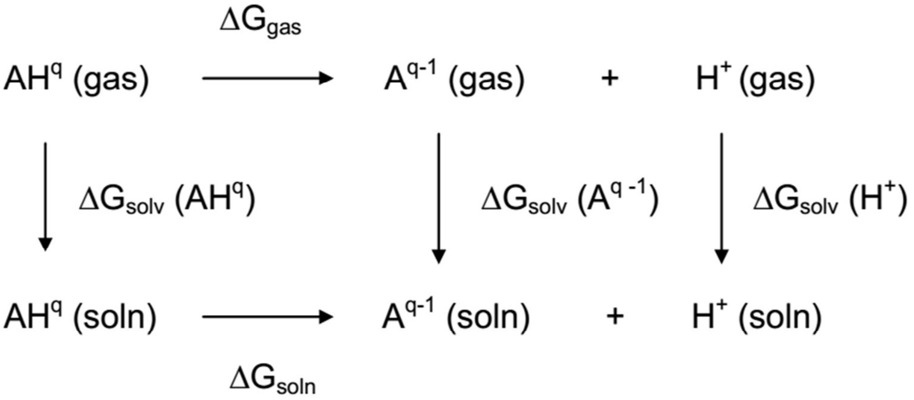
Supercomputación de Galicia (CESGA)" and to ''Consorci de Serveis Universitaris de Catalunya (CSUC)" for providing access to their computational facilities.

References

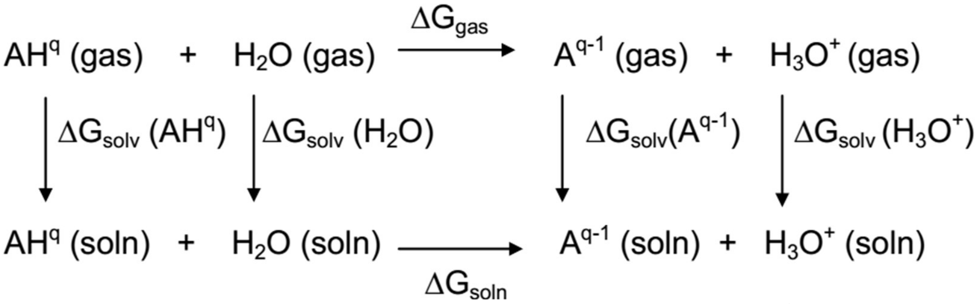
1. R. Casasnovas, J. Ortega-Castro, J. Frau, J. Donoso, F. Muñoz, *Int. J. Quantum Chem*. 2014, **114**, 1350-1363.
2. P. G. Seybold, G. C. Shields, *WIREs Comput. Mol. Sci*. 2015, **5**, 290-297.
3. J. Ho, M. L. Coote, *WIREs Comput. Mol. Sci*. 2011, **1**, 649-660.
4. J. Ho, M. L. Coote, *Theor. Chem. Acc*. 2010, **125**, 3-21.
5. R. Casasnovas, D. Fernández, J. Ortega-Castro, J. Frau, J. Donoso, F. Muñoz, *Theor. Chem. Acc*., 2011, **130**, 1-13.
6. R. Casasnovas, M. Adrover, J. Ortega-Castro, J. Frau, J. Donoso, F. Muñoz *J. Phys. Chem. B*, 2012, **116**, 10665-10675.
7. R. Casasnovas, J. Ortega-Castro, J. Frau, J. Donoso, F. Muñoz, *J. Phys. Chem. B*, 2013, **117**, 2339-2347.
8. R. Casasnovas, J. Ortega-Castro, J. Donoso, J. Frau, F. Muñoz, *Phys. Chem. Chem. Phys.* 2013, **15**, 16303-16313.
9. J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev*. 2005, **105**, 2999-3093.
10. C. J. Cramer, D. G. Truhlar, *Chem. Rev*. 1999, **99**, 2161-2200.
11. A. Klamt, G. Schuurman, *J. Chem. Soc. Perkin Trans*. 1993, **2**, 799-805.
12. M. Cossi, N. Rega, G. Scalmani, V. Barone, *J. Comput. Chem*. 2003, **24**, 669-681.
13. A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B* 2009, **113**, 6378-6396.
14. F. Kiani, A. A. Rostami, S. Sharifi, A. Bahadori, M. J. Chaichi *J. Chem. Eng. Data* 2010, **55**, 2732-2740.
15. M. Gupta, E. F. da Silva, H. F. Svendsen *J. Chem. Theory Comput.* 2013, *9*, 5021−5037.
16. S. Sastre, R. Casasnovas, F. Muñoz, J. Frau *Theor. Chem. Acc.* 2013, 132:1310.
17. Ho, J. Phys. Chem. Chem. Phys. 2015, 17, 2859-2868.
18. McQuarrie, D.M. Statistical Mechanics, Harper and Row, New York, 1970.
19. M. D. Tissandier, K. A. Cowen, W. Y. Feng, E. Gundlach, M. H. Cohen, A. D. Earhart, J. V. Coe, T. R. Tuttle, Jr. *J. Phys. Chem. A* 1998, **102***,* 7787-7794.
20. C. P. Kelly, C. J. Cramer, D. G. Truhlar *J. Phys. Chem. B* 2006, **110**, 16066-16081.
21. M. W. Palascak, G. C. Shields *J. Phys. Chem. A* 2004, **108**, 3692-3694.
22. J. R. Pliego Jr., J. M. Riveros *Phys. Lett*. 2000, **332**, 597-602.
23. R. Casasnovas, J. Frau, J. Ortega-Castro, A. Salvà, J. Donoso, F. Muñoz *J. Mol. Struct. Theochem* 2009, **912**, 5-12.
24. Y. Zhao, N. E. Schultz, D. G. Truhlar *J. Chem. Theory Comput.* 2006, **2**, 364-382.
25. Y. Zhao, D. G. Truhlar *Theor. Chem. Acc*. 2008, **120**, 215–241.
26. A. D. Becke *J. Chem. Phys*. 1993, **98**, 5648–5652.
27. C. Lee, W. Yang, G. Parr *Phys. Rev. B* 1988, **37**, 785–789.
28. J. J. P. Stewart *J. Mol. Model*. 2007, **13,** 1173-1213.
29. A. V. Marenich, C. J. Cramer, D. G. Truhlar *J. Phys. Chem. B* 2009, **113**, 4538-4543.
30. Gaussian 09, Revision B.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K.

Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2009.

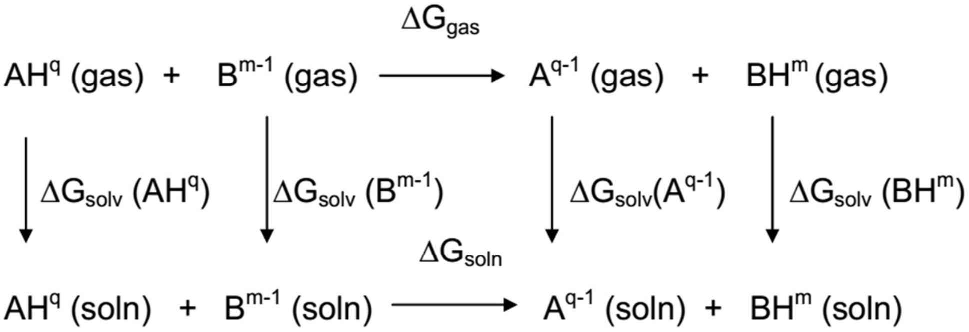
31. ChemOffice, Cambridgesoft, Copyright 1998-2016 PerkinElmer Inc.
32. J. A. Dean (1999) Mc Graw-Hill, inc. Lange's Handbook of Chemistry, Fourteenth edition
33. W. P. Jencks, J. Regenstein *Ionization Constants of Acids and Bases. Handbook of Biochemistry and Molecular Biology*. CRC Press: Cleveland, 1970, 67.
34. H. Schmidt, I. Andersson, D. Rehder, L. Pettersson *Chemistry*. 2001, **7**, 251.
35. T. T. Tominaga, H. Imasato, O. R. Nascimento, M. Tabak *Anal. Chim. Acta* 1995, **315**, 217.
36. M. Remelli, C. Conato, A. Agarossi, F. Pulidori, P. Mlynarz, H. Kozlowski *Polyhedron* 2000, **19**, 2409–2419.
37. D. L. Rabenstein, M. S. Greenberg, C. A. Evans *Biochemistry* 1977, **16**, 977-981.
38. M. R. Arnold, W. Kremer, H. D. Ludemann, H. R. Kalbitzer *Biophys. Chem*. 2002, **96**, 129–140.
39. G. Vistoli, V. Straniero, A. Pedreti, L. Fumagalli, C. Bolchi, M. Pallavicini, E. Valoti, B. Testa *Chirality* 2012, 24, 566–576.
40. a. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison *J. Cheminf*. 2011, **3**, 33. b. The Open Babel Package, version 2.3.1 http://openbabel.org
41. a. T. A. Halgren *J. Comput. Chem*. 1996, **17**, 490-519. b. T. A. Halgren *J. Comput. Chem*. 1996, **17**, 520-552. c. T. A. Halgren *J. Comput. Chem*. 1996, **17**, 553-586. d. T. A. Halgren *J. Comput. Chem*. 1996, **17**, 587-615. e. T. A. Halgren *J. Comput. Chem*. 1996, **17**, 616-641. f. T. A. Halgren *J. Comput. Chem*. 1999, **20**, 720-729.



Scheme 1. Thermodynamic cycle 1 in which an acid species AH^q is dissociated in its conjugated base A^{q-1} and a proton. ΔG_{gas}, ΔG_{soln} and ΔG_{solv} are respectively the free energies of deprotonation in gas phase, in solution and the free energy of solvation. The formal charge of the acid AH is represented by q.


ΔG_{gas}
AH^q (gas) + H_2O (gas)
A^{q-1} (gas) + H_3O^+ (gas)
ΔG_{solv} (AH^q)
ΔG_{solv} (H_2O)
ΔG_{solv}(A^{q-1})
ΔG_{solv} (H_3O^+)
AH^q (soln) + H_2O (soln)
A^{q-1} (soln) + H_3O^+ (soln)
ΔG_{soln}

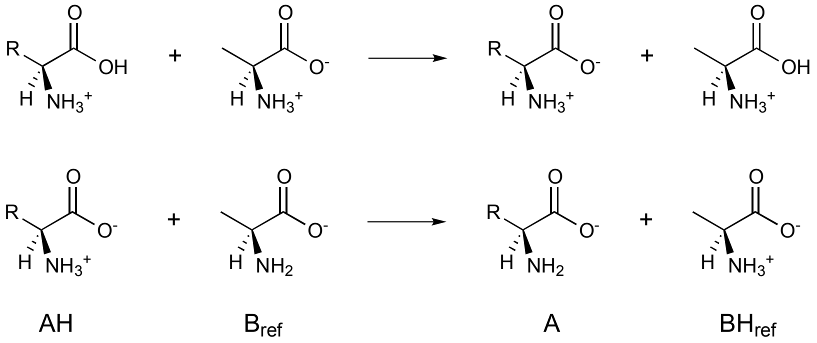
Scheme 2. Thermodynamic cycle 2 in which an acid species AH^q donates a proton to a water molecule to yield its conjugated base A^{q-1} and a hydronium cation. ΔG_{gas}, ΔG_{soln} and ΔG_{solv} are respectively the free energies of deprotonation in gas phase, in solution and the free energy of solvation. The formal charge of the acid AH is represented by q.


ΔG_{gas}
AH^q (gas) + B^{m-1} (gas)
A^{q-1} (gas) + BH^m (gas)
ΔG_{solv} (AH^q)
ΔG_{solv} (B^{m-1})
ΔG_{solv}(A^{q-1})
ΔG_{solv} (BH^m)
ΔG_{soln}
AH^q (soln) + B^{m-1} (soln)
A^{q-1} (soln) + BH^m (soln)

Scheme 3. Thermodynamic cycle 3 in which an acid species AH^q donates a proton to a base B^{m-1} to yield the conjugated base A^{q-1} and acid AH^m. ΔG_{gas}, ΔG_{soln} and ΔG_{solv} are respectively the free energies of deprotonation in gas phase, in solution and the free energy of solvation. The formal charge of the acids AH and BH are represented by q and m.

$$AH^{q}(soln) + B^{m-1}(soln) \xrightarrow{\Delta G_{soln}} A^{q-1}(soln) + BH^{m}(soln)$$

Scheme 4. Isodesmic reaction employed for the calculation of pK_a (AH). ΔG_{soln} is the free energy of the acid-base reaction in solution. The formal charge of the acids AH and BH are represented by q and m.



Scheme 5. Isodesmic reaction for the pK_a calculation of pK_{a1} (top) and pK_{a2} (bottom) with alanine as reference.

Table 1. Mean absolute deviation and standard deviation of pK_{a1} of amino acids calculated with the Isodesmic reaction compared to the experimental values. Alanine was used as the reference species.

	pK_{a1} (exptl.^{a})	ΔpK_{a1} (M052X)	ΔpK_{a1} (M062X)	ΔpK_{a1} (B3LYP)	ΔpK_{a1} (PM6)	ΔpK_{a1} (ChemOffice)
Alanine (Ref)	2.34	-	-	-	-	0.08
Glycine	2.34	0.42	0.18	0.42	0.09	0.03
Valine	2.32	0.40	0.64	0.26	0.48	0.14
Iso-Leucine	2.35	0.28	0.11	0.13	0.36	0.06
Leucine	2.33	0.02	0.02	0.35	0.29	0.04
Methionine	2.28	0.37	0.13	0.04	0.15	0.04
Proline	1.99	0.37	0.10	0.36	0.43	0.53
Phenylalanine	1.83	0.08	0.03	0.33	0.62	0.38
Tryptophan	2.38	0.26	0.28	0.32	0.11	0.29
Serine	2.21	0.67	0.60	0.31	1.37	0.05
Threonine	2.09	0.57	0.49	0.87	1.35	0.09
Asparagine	2.01	0.76	0.97	0.16	1.70	0.09
Glutamine	2.17	0.44	0.28	0.19	0.27	0.02
Tyrosine	2.18	0.06	0.08	0.12	0.09	0.16
Lysine	2.18	0.36	0.44	0.63	0.49	0.11
Arginine	2.17	0.92	1.06	0.46	0.04	0.02
Cysteine	1.96	0.28	0.70	0.21	0.02	0.14
Histidine	1.82	2.34	2.78	2.48	1.89	0.38
Aspartic acid	1.89	1.35	1.31	0.85	1.09	0.03
Glutamic acid	2.19	0.75	0.71	0.28	0.34	0.04
MAD		0.56	0.57	0.46	0.59	0.14
SD		0.54	0.65	0.54	0.59	0.14

[a] Experimental values taken from Reference [32].

Table 2. Mean absolute deviation and standard deviation of pK_{a2} of amino acids calculated with the Isodesmic reaction compared to the experimental values. Alanine was used as the reference species.

	pK_{a2}	ΔpK_{a2}	ΔpK_{a2}	ΔpK_{a2}	ΔpK_{a2}	ΔpK_{a2}
	(exptl.[a])	(M052X)	(M062X)	(B3LYP)	(PM6)	(ChemOffice)
Alanine (Ref)	9.69	-	-	-	-	0.5
Glycine	9.6	1.13	0.10	0.61	0.17	0.45
Valine	9.79	0.63	0.74	0.30	0.30	0.6
Iso-Leucine	9.68	0.64	0.69	0.56	0.63	0.54
Leucine	9.6	0.14	0.04	0.26	0.59	0.62
Methionine	9.21	0.03	0.87	0.00	1.20	0.85
Proline	10.60[b]	2.13	1.79	1.55	2.36	0.78
Phenylalanine	9.12	0.06	0.14	0.21	0.00	0.81
Tryptophan	9.39	0.22	0.10	0.20	0.56	0.62
Serine	9.15	0.71	0.75	1.36	1.24	0.22
Threonine	9.1	0.90	0.99	0.89	0.84	0.35
Asparagine	8.8	0.22	0.50	0.29	0.40	1.35
Glutamine	9.13	0.30	0.41	0.19	1.40	0.38
Tyrosine	9.11	0.06	0.18	0.48	0.84	1.56
Lysine	8.94	0.32	0.71	0.08	0.52	0.84
Arginine	9.04	0.36	0.55	0.05	0.84	1.14
Cysteine	10.28	0.69	0.82	0.65	1.28	0.44
Histidine	9.16	0.49	0.39	0.15	0.83	0.82
Aspartic acid	9.6	0.62	0.85	0.82	1.06	0.55
Glutamic acid	9.67	0.31	0.77	0.01	1.13	0.52
MAD		0.52	0.60	0.46	0.85	0.70
SD		0.49	0.42	0.44	0.53	0.34

[a]Experimental values taken from Reference [32]. [b]Ref. [33]

Table 3. Absolute errors of pK_{a1} of peptides calculated with the Isodesmic reaction compared to the experimental values. GlycylGlycine was used as reference species.

	pK_{a1}	ΔpK_{a1}	ΔpK_{a1}	ΔpK_{a1}	ΔpK_{a1}	ΔpK_{a1}
	(exptl.[a])	(M052X)	(M062X)	(B3LYP)	(PM6)	(ChemOffice)
GlyGly	3.13	-	-	-	-	0.17
GlyVal	3.18	0.16	0.40	0.29	0.20	0.12
GlyPhe	3.23	0.33	0.47	0.10	1.70	0.39
GlyAla	3.15	0.85	0.72	0.64	1.45	0.10
AlaGly	3.16[b]	0.10	0.43	0.34	0.93	0.22
AlaAla	3.32	2.08	1.95	1.01	0.35	0.29
AsnGly	2.90[b]	0.97	0.68	0.18	2.34	0.24
ValGly	3.23	1.37	1.56	1.22	1.12	0.24
SerGly	3.10[b]	0.31	0.63	0.26	0.80	0.07
SerLeu	3.08	0.87	1.94	0.45	3.32	0.01
AlaHis	2.64[c]	2.45	2.23	2.13	0.44	0.16
GlyTyr	2.93	2.39	2.39	0.25	1.31	0.28
HisGly	2.40	1.04	0.57	0.21	0.36	0.42
GlyAsp	2.81	0.78	0.76	0.55	1.54	0.27
CysAsn	2.97	0.60	0.63	1.68	0.83	0.37
PheArg	2.66	0.37	0.14	0.53	1.01	0.06
LysAla	3.22[b]	2.26	1.87	1.89	0.91	0.20
LeuTyr	3.46	1.90	1.37	0.73	1.53	0.85
TyrGly	2.98[d]	0.28	0.97	0.33	0.52	0.11
LysGlu	2.93	2.93	1.29	1.86	2.33	1.10
TyrArg	2.65	0.15	0.12	0.54	0.66	0.07
HisLys	2.50[e]	2.13	2.12	1.08	1.21	0.28
AspHis	-	-	-	-	-	-
GlyHis	-	-	-	-	-	-
GlyLys	2.96[e]	0.01	0.07	0.31	0.47	0.04
AspGly	2.10[b]	0.08	0.01	0.20	0.35	0.42
AlaGlyGly	3.19[b]	0.12	0.18	0.07	4.26	0.42
GlyAlaAla	3.38	1.61	2.44	0.64	1.97	0.30
GlySerGly	3.32	0.41	0.35	0.96	0.21	0.36
GlyGlyGly	3.23	0.09	0.20	0.71	1.69	0.39
CysGlyGly	3.13[b]	0.14	0.05	0.80	2.79	0.47
AlaLysAla	3.15[b]	0.24	0.69	1.32	1.47	0.52
PheAlaArg	2.60[b]	0.27	0.02	0.81	1.01	0.83
GlyHisLys	-	-	-	-	-	-
MAD		0.91	0.91	0.74	1.30	0.32
SD		0.89	0.79	0.57	0.96	0.25

[a]Experimental values taken from Reference [32] unless otherwise noted. [b]Ref. [33], [c]Ref. [34], [d]Ref. [35], [e]Ref. [36]

Table 4. Absolute errors of pK_{a2} of peptides calculated with the Isodesmic reaction compared to the experimental values. GlycylGlycine was used as reference species.

	pK_{a2} (exptl.)	ΔpK_{a2} (M052X)	ΔpK_{a2} (M062X)	ΔpK_{a2} (B3LYP)	ΔpK_{a2} (PM6)	ΔpK_{a2} (ChemOffice)
GlyGly	8.25	-	-	-	-	0.71
GlyVal	8.18	0.34	0.22	0.59	1.68	0.61
GlyPhe	8.11	1.64	1.52	1.54	0.77	0.58
GlyAla	8.33	0.03	0.37	0.08	2.14	0.77
AlaGly	8.24[b]	0.27	0.16	0.53	2.17	0.57
AlaAla	8.13	1.24	0.99	0.92	1.47	0.43
AsnGly	7.25[b]	0.10	0.65	0.16	2.15	0.98
ValGly	8.00	0.91	1.17	1.61	0.49	0.31
SerGly	7.33[b]	0.49	0.76	1.04	0.64	0.47
SerLeu	7.45	0.34	0.51	0.61	2.59	0.55
AlaHis	9.40[c]	0.16	0.21	0.45	4.26	0.37
GlyTyr	8.45	0.72	0.11	0.17	0.18	0.91
HisGly	7.82	0.51	0.80	2.56	3.74	0.36
GlyAsp	8.60	0.68	0.64	0.39	1.34	1.04
CysAsn	8.47	2.43	3.02	3.31	0.66	1.25
PheArg	7.57	0.14	0.15	0.25	2.20	0.16
LysAla	8.47[b]	0.41	0.69	0.47	2.70	1.18
LeuTyr	7.84	1.05	0.73	1.32	0.03	0.14
TyrGly	8.00[d]	0.12	0.18	0.32	1.44	0.56
LysGlu	7.75	0.88	0.47	0.08	0.82	0.64
TyrArg	7.39	0.54	0.77	0.26	0.89	0.05
HisLys	7.41[e]	0.23	0.45	0.70	2.01	0.07
AspHis	7.98[b]	2.49	2.43	1.23	3.60	0.64
GlyHis	8.20	2.49	2.43	1.23	3.60	0.32
GlyLys	8.01[e]	1.35	1.56	1.29	0.56	0.46
AspGly	9.07[b]	1.69	1.84	1.93	4.28	1.43
AlaGlyGly	8.15[b]	0.36	0.47	0.55	3.29	0.83
GlyAlaAla	8.10	0.33	1.03	1.29	1.24	0.89
GlySerGly	7.99	0.04	0.28	0.46	2.49	0.87
GlyGlyGly	8.09	0.29	0.18	0.13	3.31	0.9
CysGlyGly	6.95[b]	4.43	4.71	4.58	0.46	0.04
AlaLysAla	7.65[b]	0.50	0.62	0.49	1.53	0.32
PheAlaArg	7.54[b]	0.12	0.49	0.20	2.25	0.46
GlyHisLys	8.06[f]	0.84	0.09	0.23	0.96	0.85
MAD		0.81	0.88	0.92	1.76	0.61
SD		0.91	0.96	0.99	1.12	0.35

[a] Experimental values taken from Reference [32] unless otherwise noted. [c] Ref. [33], [d] Ref. [34], [e] Ref. [35], [f] Ref. [36], [g] Ref. [37]

Table 5. Mean absolute deviation (MAD) of pK_{a1} and pK_{a2} of amino acids and peptides calculated with M05-2X with various reference species.

	pK_{a1}			pK_{a2}	
Reference	Amino acids	Peptides	Reference	Amino acids	Peptides
Alanine	0.56	1.02	Alanine	0.52	1.44
GlyGly	0.87	0.91	GlyGly	2.10	0.81
Acetic acid	2.63	2.06	Ethylamine	1.46	0.81

Table 6. Absolute errors of pK_{a3} calculated with the Isodesmic reaction compared to the experimental values.

	pK_{a3} (exptl.)	ΔpK_{a3} (M052X)	ΔpK_{a3} (M062X)	ΔpK_{a3} (B3LYP)	ΔpK_{a3} (PM6)	ΔpK_{a3} (ChemOffice)
Lysine						
Lys (ref.)	10.53[a]	-	-	-	-	0.24
LysGlu	10.50[a]	0.38	0.58	0.60	2.13	0.36
GlyLys	10.50[b]	0.24	0.27	0.06	0.00	0.35
HisLys	10.49[b]	0.02	0.28	0.58	0.50	0.34
LysAla	10.70[a]	0.14	0.05	0.95	2.14	0.55
AlaLys	10.70[b]	0.26	0.39	0.03	1.00	0.36
AlaLysAla	10.30[c]	0.29	0.39	0.48	2.18	0.30
LysD-Ala	10.63[b]	0.65	0.78	0.59	1.78	0.48
GlyHisLys	10.71[d]	1.68	1.10	1.30	2.71	0.57
GlyGlyLysAla	11.10[e]	2.28	1.08	0.38	3.54	1.10
MAD		**0.66**	**0.55**	**0.55**	**1.78**	**0.47**
SD		**0.78**	**0.37**	**0.40**	**1.11**	**0.25**
Arginine						
Arg (ref.)	12.47[a]	-	-	-	-	3.88
PheArg	12.40[a]	0.00	0.00	0.00	0.00	3.76
TyrArg	11.62[a]	1.64	2.20	0.71	1.42	2.98
PheAlaArg	12.43[c]	0.38	0.31	2.29	0.59	3.79
MAD		**0.78**	**1.66**	**0.72**	**0.73**	**3.60**
SD		**0.93**	**1.39**	**1.24**	**0.92**	**0.42**
Histidine						
His (ref.)	6.00[a]	-	-	-	-	-
AlaHis	6.72[f]	1.10	1.81	2.31	0.78	-
HisGly	5.80[a]	1.38	2.51	4.58	2.84	-
HisLys	5.91[b]	1.93	2.73	2.18	2.45	-
AspHis	6.82[a]	0.52	0.98	1.01	0.02	-
GlyHis	6.77[a]	0.48	0.99	0.64	0.70	-
GlyHisLys	6.60[d]	2.00	2.23	2.54	0.89	-
GlyHisGly	6.62[d]	1.93	2.24	2.69	0.13	-
GlyGlyHisAla	7.00[e]	1.51	1.01	2.71	0.38	-
TyrHisOMe	6.41[f]	2.17	2.59	3.05	1.95	-
GluHisOMe	6.44[f]	2.34	2.94	2.76	1.81	-
MAD		**1.53**	**2.00**	**2.45**	**1.19**	-
SD		**0.66**	**0.76**	**1.08**	**1.00**	-
Cysteine						
Cys (ref.)	8.18[a]	-	-	-	-	-
CysAsn	7.09[a]	0.24	0.20	0.09	1.49	-
CysGlyGly	6.36[c]	0.88	1.24	1.06	0.93	-
MAD		**0.56**	**0.72**	**0.57**	**1.21**	-
SD		**0.45**	**0.73**	**0.69**	**0.39**	-
Tyrosine						
Tyr (ref.)	10.60[a]	-	-	-	-	1.25
GlyTyr	10.49[a]	2.45	1.97	0.23	0.63	1.08
TyrArg	9.36[a]	1.10	1.32	0.81	0.82	0.17
LeuTyr	10.09[a]	0.86	1.02	0.71	1.82	0.69
TyrGly	10.51[g]	0.18	0.01	0.12	1.43	0.97
GlyGlyTyrAla	10.30[e]	0.80	0.22	0.33	2.03	0.88
D-LeuTyr	10.35[a]	0.62	0.30	1.00	1.01	0.95
TyrHisOMe	9.69[f]	0.20	0.17	0.74	0.13	0.16
MAD		**0.89**	**0.72**	**0.56**	**1.13**	**0.77**
SD		**0.77**	**0.74**	**0.34**	**0.67**	**0.41**
Glu/Asp						
Glu (ref.)	4.25[a]	-	-	-	-	1.20
Asp	3.65[a]	1.07	1.76	0.42	0.60	1.79
LysGlu	4.47[a]	0.14	0.74	0.16	0.42	0.44
GlyGlyGluAla	4.30[e]	0.22	0.56	0.27	0.31	0.06
GlyAsp	4.45[a]	1.58	2.15	1.33	0.14	0.80
GlyGlyAspAla	3.90[e]	0.53	1.70	1.83	0.87	0.24
GluHisOMe	3.79[f]	0.04	0.18	1.03	0.93	0.66
AspGly	4.53[c]	0.84	0.27	0.93	0.21	2.76
MAD		**0.63**	**1.05**	**0.85**	**0.50**	**0.99**
SD		**0.56**	**0.80**	**0.61**	**0.31**	**0.90**
All						
MAD		**0.94**	**1.13**	**1.16**	**1.17**	**1.11**
SD		**0.75**	**0.87**	**1.05**	**0.89**	**1.14**

[a] Ref. [32], [b] Ref. [36], [c] Ref. [33], [d] Ref. [37], [e] Ref. [38], [f] Ref. [39], [g] Ref. [35]

Table 7. Mean absolute deviations and standard deviations of pK_{a3} calculated with the Isodesmic reaction and simple organic molecules as reference species[a].

	ΔpK_{a3} (M052X)	ΔpK_{a3} (M062X)	ΔpK_{a3} (B3LYP)	ΔpK_{a3} (PM6)
Lysine				
MAD	0.59	0.48	0.48	1.54
SD	0.76	0.49	0.35	1.13
Arginine				
MAD	1.52	0.99	1.67	0.69
SD	0.89	0.84	1.28	0.54
Histidine				
MAD	0.63	0.77	0.92	1.90
SD	0.43	0.52	0.81	1.04
Cysteine				
MAD	2.34	2.46	1.94	2.58
SD	0.46	0.66	0.59	0.75
Tyrosine				
MAD	1.46	0.71	0.77	2.17
SD	0.61	0.71	0.47	1.25
Glu/Asp				
MAD	0.70	1.22	0.89	2.24
SD	0.42	0.87	0.52	0.53
All				
MAD	**0.98**	**0.91**	**0.93**	**1.87**
SD	**0.77**	**0.79**	**0.75**	**1.05**

[a] Organic molecule reference species (i.e. ethylamine, ethylguanidinium, 4-methylimidazole, ethanethiol, phenol and acetic acid).

Table 8. Absolute errors of pK_{a3} calculated with thermodynamic cycles or the isodesmic reaction compared to the experimental values.

	pK_{a3} (exptl.)	C1	C2	C3	Isodesmic reaction
Tyr	10.60[c]	5.15	6.09	-	-
Gly**Tyr**[b]	10.49[c]	1.12	2.06	4.03	2.46
Leu**Tyr**	10.09[c]	3.81	2.87	8.96	0.86
TyrGly	10.12[d]	5.09	6.03	0.06	0.21
GlyGly**Tyr**Ala	10.30[e]	3.32	4.26	1.83	0.80
D-Leu**Tyr**	10.35[c]	2.23	1.28	7.37	0.62
TyrHisOMe	9.69[f]	0.37	0.57	5.52	0.19
Phe**Arg**	12.40[c]	22.43	21.48	-[a]	1.65
PheAla**Arg**	12.43[g]	31.10	30.15	-[a]	0.79
Tyr**His**OMe	6.41[f]	0.98	1.93	-[a]	2.18
MAD		**7.56**	**7.67**	**4.63**	**1.08**
SD		**10.46**	**9.95**	**3.35**	**0.82**

[a] Omitted values because restraints to one or more species in the gas phase calculations were required. [b] Residues of which pK_a are calculated are shown in bold. [c] Ref. [32], [d] Ref. [35], [e] Ref. [38], [f] Ref. [39], [g] Ref. [33].

Table 9. Absolute errors of pK_{a1}, pK_{a2} and pK_{a3} calculated with the Isodesmic reaction and PM6 by considering a conformational ensemble.

	N[a]	pK_{a1} (exptl.[d])	ΔpK_{a1} (PM6)	pK_{a2} (exptl.[d])	ΔpK_{a2} (PM6)	pK_{a3} (exptl.[d])	ΔpK_{a3}[b] (PM6)
Alanine	25	2.34	-	9.69	-		
Aspartic acid	38	1.89	1.17	9.6	1.21	3.65	1.94
Glutamic acid	57	2.19	0.16	9.67	2.25	4.25	2.50
Histidine	67	1.82	0.18	9.16	0.81	6.0	0.93
Lysine	61	2.18	0.36	8.94	0.17	10.53	0.57
Arginine	60	2.17	0.03	9.04	0.18	12.47	0.12
Tyrosine	46	2.18	0.30	9.11	0.84	10.6	1.71
Cysteine	40	1.71	0.07	10.78	2.49	8.18	2.65
MAD			**0.33**		**1.14**		**1.49**
SD			**0.39**		**0.92**		**0.97**
MAD[c]			**0.59**		**0.9**		**1.87**
SD[c]			**0.59**		**0.62**		**1.05**

[a]Total number of conformations used in the calculation. [b]Organic molecules were used as reference species for the calculation of pK_{a3}. [c]MAD and SD values calculated with a single structure for each protonation state. [d] Ref. [32].

Table 10. Mean absolute deviations and standard deviations of pK_{a1}, pK_{a2} and pK_{a3} of peptides calculated with the Isodesmic reaction and PM6 by using the minimum energy conformer as an initial structure.

	PM6	PM6 minimum energy conformer in vacuum
ΔpK_{a1}[a]	1.30 ± 0.96	2.40 ± 1.77
ΔpK_{a2}	1.80 ± 1.18	1.62 ± 1.51
ΔpK_{a3} total[b]	1.17 ± 0.89	2.13 ± 1.51
ΔpK_{a3} lys.	1.78 ± 1.11	1.99 ± 1.75
ΔpK_{a3} arg.	0.92 ± 0.44	3.41 ± 2.97
ΔpK_{a3} his.	1.19 ± 1.00	2.61 ± 1.15
ΔpK_{a3} cys.	1.21 ± 0.39	1.46 ± 1.57
ΔpK_{a3} tyr.	1.13 ± 0.67	1.83 ± 1.09
ΔpK_{a3} glu./asp.	0.50 ± 0.31	1.55 ± 1.27

[a]Errors reported as the mean absolute deviations (MAD) ± standard deviation (SD). [b]Amino acids were used as reference species for the calculation of pK_{a3}.