Competition Reaction-Based Prediction of Polyamines' Stepwise Protonation Constants: a Case Study Involving 1,4,7,10-tetrazadecane (2,2,2-tet)

Adedapo S. Adeyinka, Bryan W. Bulling and Ignacy Cukrowski*.

Department of Chemistry, Faculty of Natural and Agricultural Sciences, University of Pretoria, Lynnwood Road, Hatfield, 0002 Pretoria. South Africa

*Corresponding author: ignacy.cukrowski@up.ac.za

Acknowledgments. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Numbers 87777) and the University of Pretoria.

ABSTRACT

Theoretical prediction of four stepwise protonation constants of 1,4,7,10-tetraazadecane (2,2,2-tet) in correct order and with the smallest (largest) deviation of about 0.1 (–0.8) log unit from experimental values was achieved by an explicit application of a competition reaction (CRn) methodology in discrete-continuum solvation model involving four explicit water molecules. This methodology performs best when (i) tested ($L^{(1)}$) and reference ($L^{(2)}$) molecules are structurally similar, (ii) lowest energy conformers (LECs, selected from all possible tautomers) are used and (iii) a CRn, which assures a balanced charge distribution between reactants and products, $H_{n-1}L^{(1)} + H_nL^{(2)} = H_nL^{(1)} + H_{n-1}L^{(2)}$, is implemented. A 5-step *EEBGB*-protocol was developed to effectively and in shortest time possible select LECs (*E*, *B* and *G* stands for electronic-energy-, Boltzmann-distribution- and Gibbs-free-energy-based stepwise selection of conformers). The *EEBGB*-protocol (i) reduced (by 94%) the number of conformers subjected to the frequency calculations (to obtain *G*-values) from 420 MM-selected to 25 used to compute four protonation constants and (ii) is of general-purpose as it is applicable to any flexible and poly-charged molecules. Moreover, in search for LECs, a rapid pre-screening protocol was developed and tested; it was found efficient for the purpose of this study. Additional research protocols, aimed at even better prediction of protonation constants, are also suggested.

Keywords: Protonation constants, competition reaction, isodesmic reaction, aliphatic polyamines, DFT, thermodynamic cycle, 2,2,2-tet, 3,2,3-tet.

1. INTRODUCTION

Aliphatic polyamines (APs) are well known chelating ligands and extensive studies of their chemical properties have been carried out [1–3]. They are ubiquitous in cells and some of the biogenic ones can reduce proliferation of cells making them suitable drug candidates investigated by medicinal chemists for various therapeutic purposes [2]. As an example, their strong chelating ability has been utilized for the preparation of metal complexes which have been tested for anticancer properties [3]. Specifically, 1,4,7,10-tetraazadecane (2,2,2-tet) or *trien* is known to be a copper chelator used for the treatment of Wilson disease and is a possible drug candidate to prevent diabetic heart failure [4]. The biological activity of polyamines depends on their protonation state; hence, their protonation behaviour is of immense interest to both experimental and theoretical chemists [5–7].

Proton transfer is one of the most important processes in chemical and biochemical systems [8–13]. Consequently, the ability of a molecule to accept or donate a proton is crucial and fundamental to our understanding of the pathways/mechanisms for several important reactions in living systems [9,12]. Several experimental techniques, such as mass spectrometry and ion-cyclotron resonance techniques in the gas phase as well as UV-visible spectroscopy, potentiometry, and NMR titration procedures in the solvent phase, have been used to obtain protonation constants [8,10,12]. Using most accurate experimental technique, glass electrode potentiometry, it is possible to obtain protonation constants with typical uncertainty on the second decimal place of the log unit. However, several experimental techniques (*e.g.*, ^{13}C NMR titration) are only capable of giving results to within a fraction of a log unit, in best cases with uncertainty on the first decimal place. In spite of the fact that experimental results for thousands of molecules are available (*e.g.*, those compiled by Martell and Smith [14] or IUPAC [15]), generating theoretical predictions is still of interest because (i) this would allow assessing biological activity of molecules yet to be synthesized, (ii) many biomolecules might be difficult to investigate due to solubility and stability issues and (iii) valuable insights one might gain from theoretical/computational modelling [12,13].

Many papers have focused on theoretical prediction of protonation constants for diverse biologically important compounds such as amines, amides, carboxylic acids, bicarbonates and proteins, amongst others [11,16–34]. Most of these studies made use of various thermodynamic cycles (TCs) and mainly focused on neutral or singly charged molecules. These TCs involve a two-stage process; (i) full gas-phase energy minimization of components involved in the protonation reaction, followed by (ii) a single point calculation in solvent (water) from which $\Delta G_{(aq)}$ is computed and used to calculate protonation constants at room

temperature [12,13,16]. Typically, the TC-based methods are able to give protonation constants within ±2 log units of experimental value for neutral or singly charged molecules but this accuracy depreciates as the charge on the studied molecule increases [9]. There are several sources of errors which contribute to an inherent uncertainty of results obtained from TCs, such as (i) uncertainty in the solvation free energy of a proton, (ii) inaccuracy in evaluating the solvation free energy of ionic species by continuum solvation models (this might range between 0.5–1 kcal/mol for neutral molecules and 3–4 kcal/mol for ions) [8] and (iii) errors inherent in state-of-the-art quantum chemistry methods (about 1 kcal/mol) [8] used to compute free energies in the gas phase. In order to minimize errors, several modifications [10,12,13,35] have been developed, a prominent example of which is the incorporation of an isodesmic reaction within a TC. In other cases, results obtained from TCs have been empirically corrected using parameters obtained from linear regression analysis of experimentally measured protonation constants.[16] In some instances, these modifications have made the prediction of protonation constants to within 1–2 log units possible, though this is still far from what is obtainable experimentally.

It has been demonstrated that making use of a competition reaction (CRn) based methodology may result in more accurate predictions of protonation constants [8,12,13,35–38]. For example, it was used recently to predict four protonation constants, as $\log K_{\mathrm{H}}^{(\mathrm{n})}$ where *n* stands for the order of the stepwise protonation reaction, of highly negatively charged molecules, such as NTPA and NTA [12,13] in simulated solvent with DFT, a relatively low and cheap level among electronic structure methods. There are many factors which contributed to high quality computationally predicted four $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values, among them (i) an inherent property of error cancellation which is typical for CRn (or isodesmic reaction in general), (ii) structural similarity of and charge distribution on the studied and reference molecules and (iii) simplicity of the continuum solvation model (CSM) used,[12,13] which performs well when used at the level of theory for which it was parameterized [9]. Many existing computational methodologies can be seen as well established (or routine) now in the field and they are described in details in recent reviews by Ho and Coote [9,10] and Casasnovas et al. [8]

In spite of unquestionable successes in this area, it is somewhat surprising that there is still little [39] (in case of diamines) or no information about the theoretical prediction of protonation constants of polyamines, such as, *e.g.*, tetramines. This observation might be attributed to specific properties of polyamines, particularly with more than two N-atoms:

(i) They are extremely flexible, resulting in almost an infinite number of possible conformers; this makes discovery of required for computing protonation constants low energy conformers (LECs) a herculean task due to amount of computational resources required (energy optimisation and frequency calculations).

(ii) The difference between their first two stepwise protonation constants [14] is a fraction of a log unit in most cases and is well below typical 'resolution' reported from computational work.

(iii) Just considering the title compound, 2,2,2-tet, the HL^{+} and H_2L^{2+} protonated forms have two tautomers and their preferences in the gas and solvent phases differ. Hence, this rules out the use of commonly utilized thermodynamic cycles.

We reported recently the first extensive conformational analysis of protonated polyamines using *trien* as a case study [40] and showed that the developed protocol was able to identify representative sets of LECs. These were used to predict %-fraction of each tautomeric form of the singly and doubly protonated *trien* which were in good agreement with experimental data obtained from the ^{13}C NMR spectrometry. Usefulness of that conformational search protocol and the CRn methodology in predicting four protonation constants of highly negatively charged molecules [12,13] have motivated us to undertake this investigation where four stepwise protonation constants of 2,2,2-tet will be predicted using 1,5,8,12-tetraazadodecane (3,2,3-tet) as a reference molecule. As far as we could establish, this is the first attempt to combine the two methodologies we referred to above [12,40]. Hence, we have selected tetramines of similar structures for which experimental values are known as this allowed us verification of theoretically predicted $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values and different protocols developed in this work.

2. COMPUTATIONAL METHODS

It has been emphasized [41–44] that an appropriate description of the solvation environment is critical for best theoretical prediction of protonation constants. Because continuum solvation models, CSMs, are known to suffer from errors due to their omission of discrete hydrogen bonding and inadequate treatment of short-range electrostatics, [41,45,46] the so-called discrete-continuum solvation model (DCSM) was also used in this work. DCSM involves placing explicit solvent molecules around the solute to simulate the first solvation shell. The resultant 'supermolecule' [41,45] is immersed in a cavity that is surrounded by a dielectric continuum to model bulk solvent effects. Unfortunately, there is no generally applicable theoretical method to determine the appropriate number of explicit water molecules needed to represent the first solvation shell. Hence, we have decided to use four water molecules (i) to

facilitate the formation of maximum number of possible hydrogen bonding between water molecules and the solute with two pairs of –NH and $–NH_2$ groups and (ii) to keep the computational resources needed for this work affordable.

Conformational search was performed in Spartan [47] to generate a large set of representative conformers of the various protonated (H_nL^{n+}) forms of 2,2,2-tet and 3,2,3-tet using molecular mechanics with the MMFF force field. Furthermore, to account for the aqueous solvent effects, the Monte Carlo algorithm in combination with MMFF(aq) option, as implemented in Spartan, were utilised. It was necessary to employ MMFF(aq) because the sets of LECs discovered in the gas phase (using MMFF) were significantly different. This was done by a systematic variation of the torsional angle of each rotatable bond as described previously [40] with slight modifications implemented in the case of 3,2,3-tet (see PART 1 of the SI for a full description of the conformational search procedure used). We have also performed conformational search on the same ligands with explicitly added four water molecules which were placed (i) randomly in relation to their orientation toward a backbone structure of a ligand, but (ii) quite evenly along a molecule; for illustration, free ligands with water molecules are shown in Figure 1. Linear structures of all possible forms the 2,2,2,-tet ($L^{(1)}$) and 3,2,3-tet ($L^{(2)}$) ligands, shown in PART 2 of the SI as Figures S1-S2, were used as inputs for the MM-based conformational search.

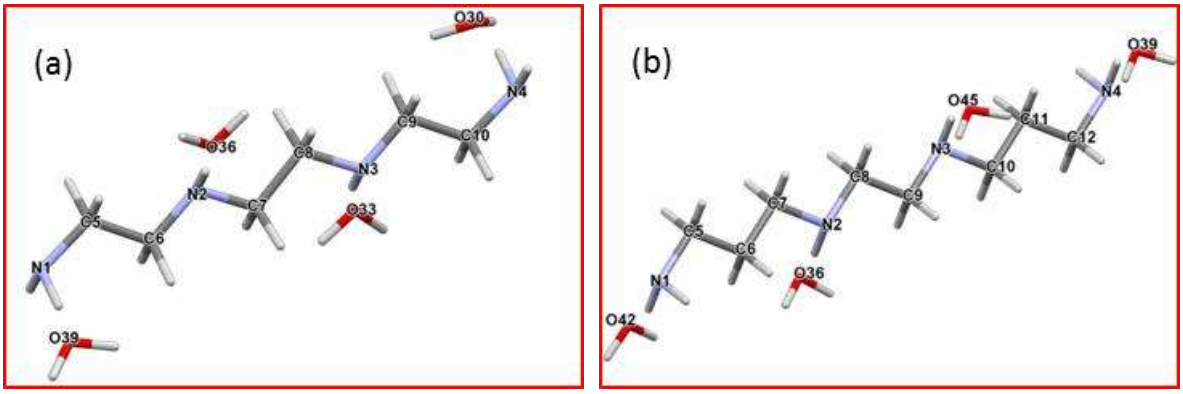


Fig. 1 Capped-stick representation of free ligand linear input structures with explicit water molecules used for conformational search by MMFF(aq): 2,2,2-tet in part a; 3,2,3-tet in part b.

A maximum of thirty unique and lowest in energy conformers was retained after each conformational search; they were energy optimised in Gaussian 09, revision D01, [48] at the RB3LYP/6-311++G(d,p) level of theory in conjunction with default settings of the Polarizable Continuum solvation Model (PCM) using water as solvent (ε = 78.3553). Vibrational frequencies were computed using the rigid rotor harmonic (RRHO) approximation, as implemented in Gaussian 09, in order to (i) obtain Gibbs free energies needed for computing

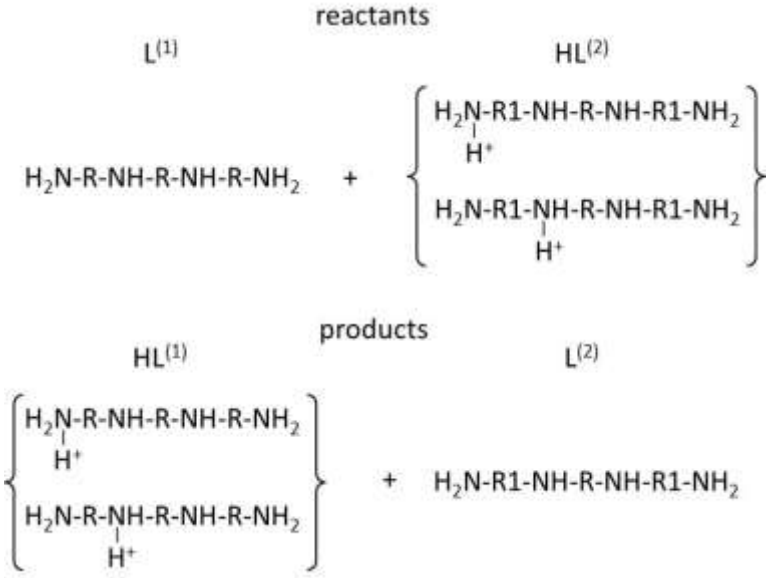
protonation constants and (ii) to verify that all minimum energy structures reported in this study were true stationary points on the potential energy surface. Furthermore, a tight gradient criterion was used along with an ultrafine integration grid to ensure acceptable convergence of frequencies computed. [49]

3. COMPETITION REACTION BASED PROTOCOL

A general concept of the CRn methodology was described previously when it was used to determine protonation constants of polycarboxylic acids [12,13] as well as formation constants [50]. As such, a competition for a proton between a polyamine (ligand) under investigation $L^{(1)}$ and a structurally analogous reference molecule $L^{(2)}$ is explored here to compute the free energy change $\Delta G^{(n)}_{\mathrm{CRn}}$ needed to calculate an *n*th protonation constant, as $\log K^{(n)}_{\mathrm{H}}$, in aqueous solution. An example of a CRn reaction for the first protonation reaction can be written as (for simplicity, the (aq) notation and charges were omitted throughout),

$$L^{(1)} + HL^{(2)} = HL^{(1)} + L^{(2)} . \qquad \Delta G^{(1)}_{\mathrm{CRn}} \qquad (1)$$

However, because this work is concerned with tetramines with several possible protonation sites, additional and important aspects had to be considered. As depicted in Scheme 1, which involves all possible protonation sites of 2,2,2-tet (here $L^{(1)}$) and 3,2,3-tet (here $L^{(2)}$) when they are singly protonated, one is faced with nine possible competition reactions shown in PART 1 in the SI, which indeed might take place in a real competition experiment in a solution.



Scheme 1 Possible tautomers of 2,2,2-tet ($L^{(1)}$) and 3,2,3-tet ($L^{(2)}$) (R = $–C_2H_4–$; R1 = $–C_3H_6–$) which were considered in the competition reaction based protocol to compute $\log K^{(1)}_{\mathrm{H}}$ of $L^{(1)}$.

It is obvious that each possible CRn generates different $\Delta G^{(1)}_{\mathrm{CRn}}$ value because inequality $G(\mathrm{HL_p}) \neq G(\mathrm{HL_s})$ holds for any polyamine. Furthermore, only in few instances one is able to predict most likely and the only one possible protonation site in polyamines with certainty. Hence, one must, in principle, use most general expression for a CRn,

$$\mathrm{L}^{(1)} + \mathrm{HL}^{(2)}_{\mathrm{p}} + \mathrm{HL}^{(2)}_{\mathrm{s}} = \mathrm{HL}^{(1)}_{\mathrm{p}} + \mathrm{HL}^{(1)}_{\mathrm{s}} + \mathrm{L}^{(2)} \qquad \Delta G^{(1)}_{\mathrm{CRn}} \qquad (2)$$

where subscripts 's' and 'p' denote a primary and secondary N-atom of a ligand being protonated. This is an additional complication because to calculate G_{products} and $G_{\mathrm{reactants}}$ an exact %-fraction of the two possible tautomers, $\mathrm{HL_p}$ and $\mathrm{HL_s}$, for both polyamines must be known. As a matter of fact, this is still not a sufficient requirement to calculate $\Delta G^{(1)}_{\mathrm{CRn}}$; note that to compute, *e.g.*, $G_{\mathrm{reactants}}$ of reaction 2, all possible conformers (there are thousands of them when linear aliphatic polyamines are considered), or at least the LECs of L, $\mathrm{HL_p}$ and $\mathrm{HL_s}$ for both polyamines must be considered, hence

$$G_{\mathrm{reactants}} = \sum_{k=1}^{\mathrm{LEC}} w_k G_k(\mathrm{L}^{(1)}) + \sum_{m=1}^{\mathrm{LEC}} w_m G_m(\mathrm{HL}^{(2)}_{\mathrm{p}}) + \sum_{n=1}^{\mathrm{LEC}} w_n G_n(\mathrm{HL}^{(2)}_{\mathrm{s}}) \qquad (3)$$

or, in more general form when *t* tautomers are possible, one can write

$$G_{\mathrm{reactants}} = \sum_{k=1}^{\mathrm{LEC}} w_k G_k(\mathrm{L}^{(1)}) + \sum_{l=1}^{t} \sum_{m=1}^{\mathrm{LEC}} w_m G_m(\mathrm{HL}^{(2)}_{\mathrm{t}}) \qquad (4)$$

where *w* stands for the population fraction obtained from the Boltzmann distribution calculated for selected LECs of $\mathrm{L}^{(1)}$, $\mathrm{HL}^{(2)}_{\mathrm{p}}$ and $\mathrm{HL}^{(2)}_{\mathrm{s}}$ (or in general *t* tautomers of $\mathrm{HL}^{(2)}_{\mathrm{t}}$). Note that *w* can be seen and must be used as the weight factor which assures proportional (to this structure contribution to the entire population) free energy contribution to the computed free energy change of the competition reaction. The same considerations equally apply to products of the reaction 1, hence one can write

$$G_{\mathrm{products}} = \sum_{x=1}^{\mathrm{LEC}} w_x G_x(\mathrm{L}^{(2)}) + \sum_{y=1}^{\mathrm{LEC}} w_y G_y(\mathrm{HL}^{(1)}_{\mathrm{p}}) + \sum_{z=1}^{\mathrm{LEC}} w_z G_z(\mathrm{HL}^{(1)}_{\mathrm{s}}) \qquad (5)$$

or, as a general expression,

$$G_{\mathrm{products}} = \sum_{x=1}^{\mathrm{LEC}} w_x G_x(\mathrm{L}^{(2)}) + \sum_{k=1}^{t} \sum_{y=1}^{\mathrm{LEC}} w_y G_y(\mathrm{HL}^{(1)}_{\mathrm{t}}) \qquad (6)$$

As shown previously [12], the free energy change for the competition reaction of the first protonation step can be obtained from a general expression

$$\Delta G^{(1)}_{\mathrm{CRn}} = G(\mathrm{HL}^{(1)}) + G(\mathrm{L}^{(2)}) - G(\mathrm{L}^{(1)}) - G(\mathrm{HL}^{(2)}) = \Delta G(\mathrm{L}^{(1)}) - \Delta G(\mathrm{L}^{(2)}) \quad (7)$$

where $\Delta G(\mathrm{L})$ is calculated for a direct protonation reaction L + H = HL involving all tautomers. Our aim is to compute $\Delta G(\mathrm{L}^{(1)})$ which is needed to obtain the first protonation constant, as $\log K^{(1)}_{\mathrm{H}}$, of $\mathrm{L}^{(1)}$ from $\Delta G = -RT \ln K^{(1)}_{\mathrm{H}}$. Because protonation constants of $\mathrm{L}^{(2)}$ are known, the $\Delta G(\mathrm{L}^{(2)})$ term can be easily obtained ($\Delta G = -\mathrm{RT}\ln K$) and one is left, in general, with four G values for all reactants and products in reaction 1. However, when polyamines investigated here are considered, one must combine expressions 3 and 5 to compute $\Delta G^{(1)}_{\mathrm{CRn}}$,

$$\Delta G^{(1)}_{\mathrm{CRn}} = \sum_{x=1}^{\mathrm{LEC}} w_x G_x(\mathrm{L}^{(2)}) + \sum_{y=1}^{\mathrm{LEC}} w_y G_y(\mathrm{HL}^{(1)}_{\mathrm{p}}) + \sum_{z=1}^{\mathrm{LEC}} w_z G_z(\mathrm{HL}^{(1)}_{\mathrm{s}})$$

$$- \sum_{k=1}^{\mathrm{LEC}} w_k G_k(\mathrm{L}^{(1)}) + \sum_{m=1}^{\mathrm{LEC}} w_m G_m(\mathrm{HL}^{(2)}_{\mathrm{p}}) + \sum_{n=1}^{\mathrm{LEC}} w_n G_n(\mathrm{HL}^{(2)}_{\mathrm{s}}) . \quad (8)$$

As it is seen from expression 8, computing $\Delta G^{(1)}_{\mathrm{CRn}}$ is a formidable and almost an impossible task when, at least, time and computational resources needed to achieve our main goal are considered. Because of that, we explored different options (they will be discussed in sections that follow) to simplify the protocol without compromising the quality of computed protonation constants.

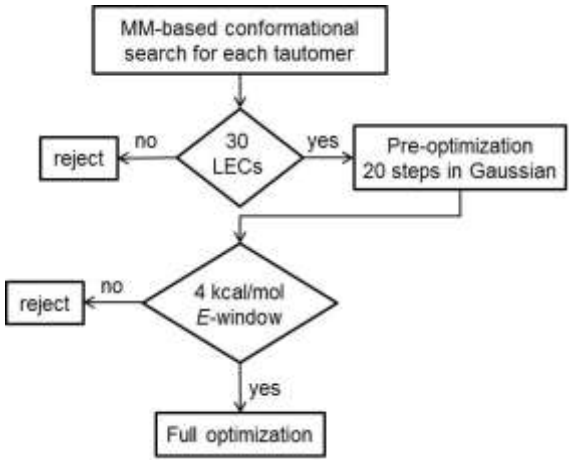
The added advantage of using a competition reaction is that a significant cancellation of different errors inherent in solvation model and electronic structure method used to optimise molecules should take place. Also, one expects that structural similarity of reactants and products, as is the case here, should result in errors minimization (cancellation). To this effect and knowing that computational optimisation of structures with multiple charges is still a challenge when accuracy goes, one should also consider the selection of reference molecules in terms of resultant placement of charges among reactants and products – see PART 3 in the SI for details.

4. RESULTS AND DISCUSSION

4.1. Pre-optimisation protocol

Firstly, it is important to realize how enormous computational task this kind of study requires when all conformers were to be optimised with frequency calculations; retaining 30 MM-selected lowest energy conformers of each tautomer results in 210 structures for each ligand, $\mathrm{L}^{(1)}$ and $\mathrm{L}^{(2)}$. When structures with explicit water molecules are also considered, as is the case here, then the starting minimum number of conformers one must consider is 840.

Because we wanted to develop a feasible protocol, we decided to seek alternative avenues. To this effect, we took advantage of having a large data bank from previous work [40] where hundreds of 2,2,2-tet conformers were fully optimised in Gaussian. A thorough inspection of the optimisation profiles generated for all tautomers of protonated forms of 2,2,2-tet revealed that in order to predict 'safely' the set of lowest energy conformers needed for the purpose of this study it would be sufficient to implement a pre-optimisation operation which involves terminating the optimisation process after 20 steps – for details see PART 4 in the SI. We decided to implement this finding in the optimisation of 3,2,3-tet conformers in both solvation models and the protocol implemented is shown in Scheme 2.

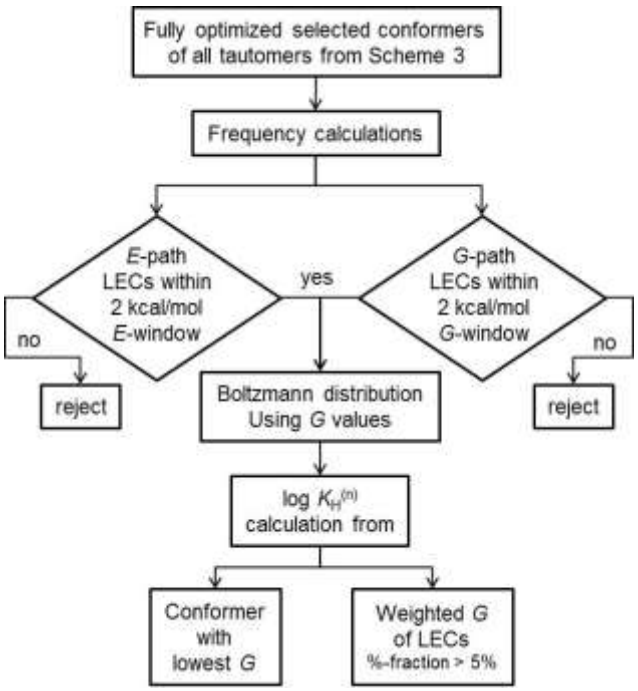


Scheme 2 Protocol used to select structures for full energy optimisation.

4.2 General purpose protocol

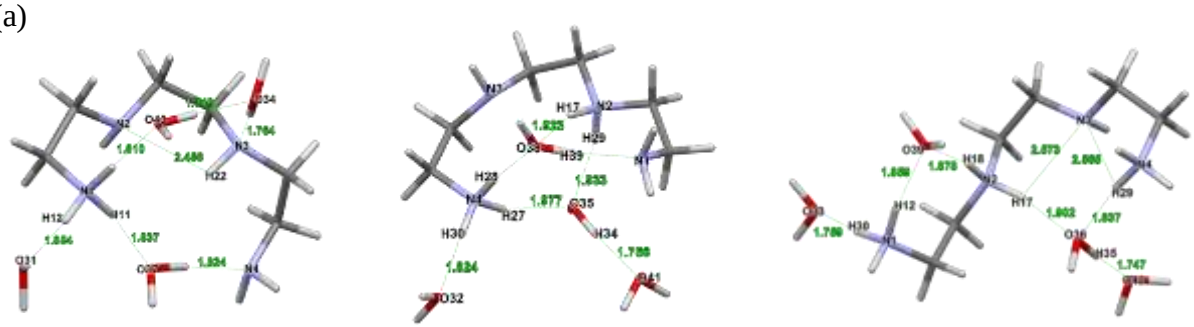
Implementation of the pre-optimisation step allowed us to preliminarily reject about 60-70% of conformers. The remaining conformers had to be fully energy optimised and subjected to frequency calculations as *G* values are required to predict protonation constants. Knowing that frequency calculations are extremely time consuming and this is particularly true when explicit water molecules are included, we decided to explore additional two selection paths with a hope that maybe it would be possible to reduce the number of necessary conformers even further, hence reduction in computational time should result too. The general purpose protocol developed in this work is shown in Scheme 3. It incorporates a step-wise elimination of 'redundant' conformers and specific strategies tested in computing protonation constants.

Examples of 2,2,2-tet and 3,2,3-tet LECs for all protonated forms of each tautomer selected after full optimisation in the continuum and discrete-continuum solvation model are shown in PART 5 (as Figures S4–S13) and PART 6 (as Figures S14–S23), respectively, of the SI. The lowest energy HL, H_2L and H_3L conformers discovered for each ligand in DCSM are shown in Figure 2. To ensure easy identification and differentiation between conformers of the various



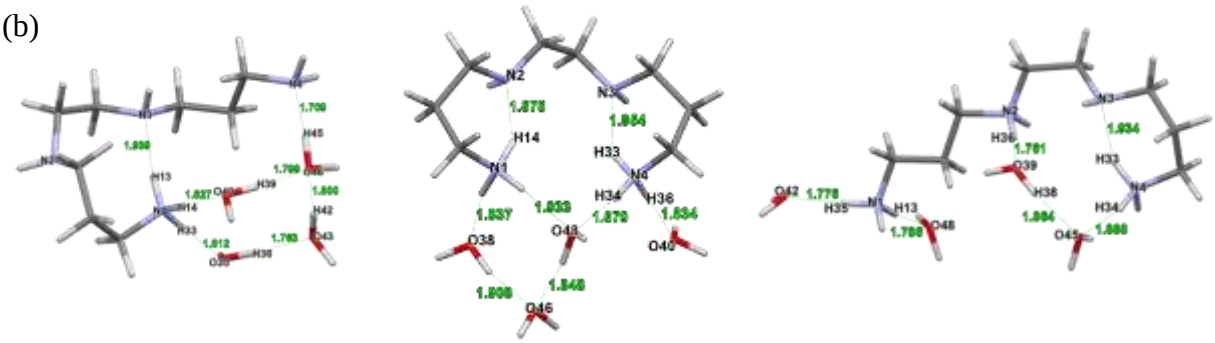
Scheme 3 General purpose approach used in testing different methodologies in search of time (cost) most-effective protocol for computational determination of protonation constants.

(a)



C_p01 $C_{ps}02$ $C_{tp}10$

(b)



C_p26 $C_{pp}18$ $C_{tp}01$

Figure 2 Lowest energy HL, H_2L and H_3L conformers with explicit water molecules of 2,2,2-tet in part (a) and 3,2,3-tet in part (b) found from the *E*-path shown in Scheme 3.

protonated (H_nL^{n+}) forms of 2,2,2-tet and 3,2,3-tet, we have consistently labelled them as:

- C_Ln for structures of the free ligand (L),

- $C_p n$ and $C_s n$ for the primary (HL_p) and secondary (HL_s) forms of monoprotonated structures where the primary and secondary N-atoms are protonated, respectively,
- $C_{ps} n$ and $C_{pp} n$, for the diprotonated structures, where $C_{ps} n$ is used to denote conformers of the tautomeric form where one primary and one secondary nitrogen atoms are protonated (H_2L_{ps}) whereas $C_{pp} n$ is used for a structure in which the two terminal nitrogen atoms are protonated (H_2L_{pp}),
- $C_{tp} n$ for structures of the triply protonated form in which both primary nitrogen atoms and one secondary nitrogen atom are protonated; in the case of ligands studied here, there is only one stable tautomer according to physical charge separation requirements, and
- $C_{fp} n$ is used to denote structures of the fully protonated form.

It is easy to establish, using Boltzmann distribution, that conformers with energies greater than 3 kcal/mol (relative to the lowest energy conformer when using either *E*- or *G*-path) contribute insignificantly to the total population, typically below 0.1 %-fraction. Furthermore, when the weighted energy of conformers was used to compute the overall *G* value of all selected conformers with %-fraction either above 1 or 5%, it became clear that incorporation of conformers characterized by 1 < %-fraction of the total population < 5 had no significant impact on the computed $\log K_H^{(n)}$ values. Analysis of Boltzmann distributions obtained for all protonated forms of both ligands (when applicable, the combined tautomers were used to generate a population of conformers, *e.g.*, HL_p plus HL_s) revealed that selecting conformers with %-fraction > 5% always resulted in the *E*- or *G*-window < 2 kcal/mol within which LECs were found.

For illustration purposes, Table 1 shows five lowest in energy conformers (with explicit water molecules) selected by the electronic and Gibbs free energy based paths for 2,2,2-tet and 3,2,3-tet; relevant data for the implicit solvation model are included in Table S1 of the SI. The *E*-path shown in Scheme 3 was implemented to test whether the selection of LECs within 2 kcal/mol *E*-window would retain conformers which, after frequency calculations, would give *G* values suitable for protonation constant calculations in terms of quality (accuracy) of computed values. This approach was taken because it might result in smaller number of conformers subjected to frequency calculations. Note, that in the case of the *G*-path shown in Scheme 3, frequency calculations were performed for all conformers falling within 4 kcal/mol electronic energy window of pre-optimisation.

Table 1. Five lowest energy conformers of all H_nL with explicit water molecules, using either *E* or *G* values, for: part (a) - 2,2,2-tet and part (b) - 3,2,3-tet.[a]

(a)

	L			HL			H_2L			H_3L			H_4L	
Conf	ΔE	%	Conf	ΔE	%	Conf	ΔE	%	Conf	ΔE	%	Conf	ΔE	%
C_L03	0.00	58.4	C_p01	0.00	69.5	$C_{ps}02$	0.00	91.7	$C_{tp}10$	0.00	52.4	$C_{fp}01$	0.00	76.5
C_L01	0.66	19.1	C_s07	0.49	30.5	$C_{ps}03$	1.44	8.0	$C_{tp}02$	0.08	45.8	$C_{fp}02$	0.88	17.3
C_L05	1.00	10.9	–	–	–	$C_{ps}12$	4.13	0.1	$C_{tp}06$	2.40	0.9	$C_{fp}04$	2.09	2.2
C_L04	1.55	4.3	–	–	–	$C_{pp}09$	4.24	0.1	$C_{tp}01$	2.94	0.4	$C_{fp}06$	2.10	2.2
C_L06	1.59	4.0	–	–	–	$C_{pp}04$	4.37	0.1	$C_{tp}11$	3.14	0.3	$C_{fp}07$	2.24	1.7
Conf	ΔG	%	Conf	ΔG	%	Conf	ΔG	%	Conf	ΔG	%	Conf	ΔG	%
C_L09	0.00	39.7	C_p01	0.00	88.7	$C_{pp}08$	0.00	84.1	$C_{tp}02$	0.00	44.1	$C_{fp}01$	0.00	89.0
C_L03	0.03	37.9	C_s07	1.22	11.3	$C_{ps}02$	1.38	8.1	$C_{tp}06$	0.44	21.1	$C_{fp}02$	1.34	9.3
C_L01	0.48	17.7	–	–	–	$C_{pp}06$	1.65	5.1	$C_{tp}01$	0.60	16.1	$C_{fp}06$	2.56	1.2
C_L05	1.39	3.8	–	–	–	$C_{ps}03$	2.44	1.4	$C_{tp}10$	0.79	11.7	$C_{fp}04$	3.31	0.3
C_L06	2.36	0.7	–	–	–	$C_{pp}07$	2.52	1.2	$C_{tp}09$	1.78	2.2	$C_{fp}07$	3.61	0.2

(b)

	L			HL			H_2L			H_3L			H_4L	
Conf	ΔE	%	Conf	ΔE	%	Conf	ΔE	%	Conf	ΔE	%	Conf	ΔE	%
C_L21	0.00	96.3	C_p26	0.00	62.5	$C_{pp}18$	0.00	92.0	$C_{tp}01$	0.00	64.2	$C_{fp}01$	0.00	35.4
C_L 19	2.45	1.5	C_p24	0.38	33.1	$C_{ps}06$	2.05	2.9	$C_{tp}25$	0.57	24.3	$C_{fp}08$	0.59	13.01
C_L 01	2.54	1.3	C_p19	2.06	1.9	$C_{pp}19$	2.15	2.4	$C_{tp}02$	1.26	7.6	$C_{fp}09$	0.97	6.9
C_L 09	3.30	0.4	C_p12	2.61	0.8	$C_{ps}21$	2.64	1.1	$C_{tp}10$	2.33	1.2	$C_{fp}04$	1.03	6.2
C_L 11	3.68	0.2	C_p07	2.85	0.5	$C_{ps}09$	2.83	0.8	$C_{tp}15$	2.58	0.8	$C_{fp}25$	1.06	6.0
Conf	ΔG	%	Conf	ΔG	%	Conf	ΔG	%	Conf	ΔG	%	Conf	ΔG	%
C_L19	0.00	46.1	C_p07	0.00	64.3	$C_{ps}13$	0.00	59.8	$C_{tp}21$	0.00	42.5	$C_{fp}06$	0.00	28.8
C_L09	0.31	27.3	C_p24	0.96	12.7	$C_{ps}10$	0.29	36.9	$C_{tp}23$	0.07	37.4	$C_{fp}04$	0.04	27.0
C_L01	0.61	16.3	C_s07	0.99	12.0	$C_{ps}28$	2.06	1.9	$C_{tp}04$	0.97	8.3	$C_{fp}01$	0.23	19.7
C_L21	1.50	3.7	C_p27	1.42	5.8	$C_{ps}01$	2.50	0.9	$C_{tp}09$	1.23	5.3	$C_{fp}05$	0.89	6.3
C_L07	1.76	2.4	C_p26	1.82	3.0	$C_{ps}06$	3.23	0.3	$C_{tp}01$	1.28	4.9	$C_{fp}25$	0.94	5.9

[a]Conf stands for conformer; ΔE (ΔG) was calculated relative to the lowest *E* (*G*) energy conformer; % is the %-fraction of the total population from Boltzmann distribution.

4.3. Computed protonation constants.

We tested numerous competition reaction types, such as shown in Scheme 1 and Scheme S1 in PART 3 of the SI, but whenever the reference molecule $L^{(2)}$ had (i) more than one proton relative to molecule under investigation (*e.g.*, $L^{(1)} + H_2L^{(2)}$ or $HL^{(1)} + H_3L^{(2)}$), (ii) the same number of protons (*e.g.*, $HL^{(1)} + HL^{(2)}$) or (iii) smaller number of protons (*e.g.*, $H_3L^{(1)} + H_2L^{(2)}$) results obtained were of poor quality - some examples are provided in Table S2 of the SI. This is in full agreement with previous reports [12,13]. Therefore, we would only be discussing results obtained from competition reactions where two competing for a proton ligands are involved in the protonation reaction of the same order, $H_{n-1}L^{(1)} + H_nL^{(2)} = H_nL^{(1)} + H_{n-1}L^{(2)}$, where n = 1,2,…,NPS and NPS stands for the number of protonation steps a ligand can be involved in, here four. The computed protonation constants obtained in different solvation models are presented in Table 2 where either a single conformer with the lowest *G* value (shown under column heading '*G* of LEC') or the weighted *G* values of selected conformers with the %-fraction > 5 in *G* (under column heading 'Weighted *G*') were used.

To assess quality of computed protonation constants one must consider two important aspects, namely (i) the error in computed protonation constant relative to the relevant experimental $\log K_H^{(n)}$ values (9.75, 9.07, 6.58 and 3.27 for the consecutive, from first to forth, stepwise protonation constant[14] of 2,2,2-tet; for the reference molecule, 3,2,3-tet, $\log K_H^{(n)}$ values of 10.53, 9.77, 8.30 and 5.59 for the first to fourth stepwise protonation constant [14] were used) and (ii) theoretically predicted sequence in values of protonation constants. The second criterion is also of an utmost importance because the experimental first and second protonation constants of 2,2,2-tet differ only by less than 0.7 log unit (a typical feature among polyamines; note also that 0.76 log unit difference is observed for 3,2,3-tet) which, in principle, can be seen as hardly achievable when typical accuracy obtained from computational work reported to date is considered.

The analysis of the data in Table 2 demonstrates that, indeed, it is possible to predict all stepwise protonation constants of 2,2,2-tet in correct sequence and with errors smaller than 1 log unit but only when structures with explicit water molecules were used and *E*-path was followed. Interestingly and importantly, results obtained from a single and weighted *G* values (data under the '*G* of LEC' and 'Weighted *G*' headings in *E*-path) are comparable as they differ by about ±0.1 log unit (see $\Delta_1 - \Delta_2$ values in Table 2). The first protonation constant can be seen as of analytical quality as it differs from the experimental $\log K_H^{(1)}$ value by –0.01 and 0.08 log unit when a single or weighted *G* value was used, respectively, whereas the second

and third protonation constants we regard as excellent prediction as they reproduced experimental values just to within –0.3±0.1 log units.

Table 2. Computed from *E*- and *G*-paths protonation constants, as $\log K_H^{(n)}$, for 2,2,2-tet using data from a discrete-continuum solvation model (DCSM) in part (a) and continuum solvation model (CSM) in part (b).[a]

(a)	***E*-path**				
DCSM	***G* of LEC**		**Weighted *G***		$\Delta_1 - \Delta_2$
Reaction	$\log K_H^{(n)}$	Δ_1	$\log K_H^{(n)}$	Δ_2	
$L^{(1)} + HL^{(2)} = HL^{(1)} + L^{(2)}$	9.74	–0.01	9.83	0.08	–0.09
$HL^{(1)} + H_2L^{(2)} = H_2L^{(1)} + HL^{(2)}$	8.87	–0.20	8.75	–0.32	0.13
$H_2L^{(1)} + H_3L^{(2)} = H_3L^{(1)} + H_2L^{(2)}$	6.12	–0.46	6.19	–0.39	–0.07
$H_3L^{(1)} + H_4L^{(2)} = H_4L^{(1)} + H_3L^{(2)}$	2.41	–0.86	2.50	–0.77	–0.09
	***G*-path**				
	***G* of LEC**		**Weighted *G***		
$L^{(1)} + HL^{(2)} = HL^{(1)} + L^{(2)}$	10.14	0.39	10.21	0.46	–0.07
$HL^{(1)} + H_2L^{(2)} = H_2L^{(1)} + HL^{(2)}$	7.37	–1.70	7.15	–1.92	0.22
$H_2L^{(1)} + H_3L^{(2)} = H_3L^{(1)} + H_2L^{(2)}$	7.40	0.82	7.42	0.84	–0.02
$H_3L^{(1)} + H_4L^{(2)} = H_4L^{(1)} + H_3L^{(2)}$	3.34	0.07	3.44	0.17	–0.09

(b)	***E*-path**				
CSM	***G* of LEC**		**Weighted *G***		$\Delta_1 - \Delta_2$
Reaction	$\log K_H^{(n)}$	Δ_1	$\log K_H^{(n)}$	Δ_2	
$L^{(1)} + HL^{(2)} = HL^{(1)} + L^{(2)}$	8.95	–0.80	8.88	–0.87	0.07
$HL^{(1)} + H_2L^{(2)} = H_2L^{(1)} + HL^{(2)}$	7.71	–1.36	7.80	–1.27	–0.09
$H_2L^{(1)} + H_3L^{(2)} = H_3L^{(1)} + H_2L^{(2)}$	6.34	–0.24	6.34	–0.24	0.00
$H_3L^{(1)} + H_4L^{(2)} = H_4L^{(1)} + H_3L^{(2)}$	–0.96	–4.23	–1.13	–4.40	0.17
	***G*-path**				
	***G* of LEC**		**Weighted *G***		
$L^{(1)} + HL^{(2)} = HL^{(1)} + L^{(2)}$	8.95	–0.80	8.85	–0.90	0.10
$HL^{(1)} + H_2L^{(2)} = H_2L^{(1)} + HL^{(2)}$	7.71	–1.36	7.72	–1.35	–0.01
$H_2L^{(1)} + H_3L^{(2)} = H_3L^{(1)} + H_2L^{(2)}$	6.34	–0.24	6.45	–0.13	–0.11
$H_3L^{(1)} + H_4L^{(2)} = H_4L^{(1)} + H_3L^{(2)}$	–0.96	–4.23	–1.23	–4.50	0.27

[a]Weighted *G* values were obtained using each conformers fraction of the total population (from Boltzmann distribution) as a weight for their *G* contribution ($w \times G$); Δ_n = computed – experimental $\log K_H^{(n)}$.

Furthermore, one observes a unidirectional error obtained for the second, third and fourth protonation constants (computed values are consistently smaller relative to experimental values) and the departure increases with the increase in the protonation constant number. This, most likely, might be attributed to somewhat poorer performance of energy optimisation

in case of highly charged molecules; the larger the positive charge on molecules involved in the CRn, the larger difference between experimental and computed values.

To this effect, it has been noted previously that aliphatic polyamines are difficult to model using most quantum chemical solvation models [41]; hence, we consider results reported here as highly satisfactory and significant improvement relative to data reported for amines previously. It is also possible to assume that closer to experimental values third and fourth protonation constants could be obtained by placing larger number of explicit water molecules to 'better' disperse charges on the macro-molecular assembly (*e.g.*, $H_nL + 8H_2O$) immersed in a simulated water environment.

Our focus now is on *G*-path for data obtained with explicit water molecules – part (a) in Table 2. Except for the second protonation constant ($\log K_H^{(2)}$ was underestimated by about 1.7–1.9 log units) results obtained could be seen as satisfactory because they fall within or below the typically reported error ranges when TCs are used for neutral or singly charged molecules [8–10]. Unfortunately, the overall quality of data obtained from the *G*-path in DCSM must be seen as unacceptable. This is because the experimental sequence of protonation constants, $\log K_H^{(n)} > \log K_H^{(n+1)}$, is not reproduced. To illustrate this, performance of different methodologies tested in this work is depicted in Figure 3 as differences between successive protonation constants, $\Delta \log K_H^{(n,n+1)} = \log K_H^{(n)} - \log K_H^{(n+1)}$, where such values obtained for experimental data are also included. Clearly, most accurate protonation constants were computed from *E*-path in DCSM – see top left graph in Figure 3.

To gain some insight on the origin of the observed disparity in accuracy between *E*- and *G*-paths, we compared structures of relevant conformers; their *E* and/or *G* values were used to select conformers for computing the second protonation constant. Figure 4 shows the lowest energy conformers of diprotonated 2,2,2-tet and 3,2,3-tet obtained from the *G*-path whereas those for *E*-path are shown in Figure 2 (additional structures are shown in Figures S4–S23 in PARTS 5 and 6 of the SI). Structural comparison revealed that conformers selected from *E*-path have a compact structure with water molecules in the first solvation shell being arranged such that (i) polyamines form a ring closed by water molecules and (ii) each protonated site is involved in interactions with several water molecules. In contrast, the *G*-path produced conformers with extended configurations of polyamines with explicit water molecules (i) distributed unevenly between two terminal functional groups and (ii) not interacting with all protonated sites. Hence, these structures tend to have increased entropic contributions to their free energy compared to those selected from the *E*-path. Their increased entropic correction

is due to arrangements of explicit water molecules which may artificially result in a greater number of

(a)

(b)

(c)

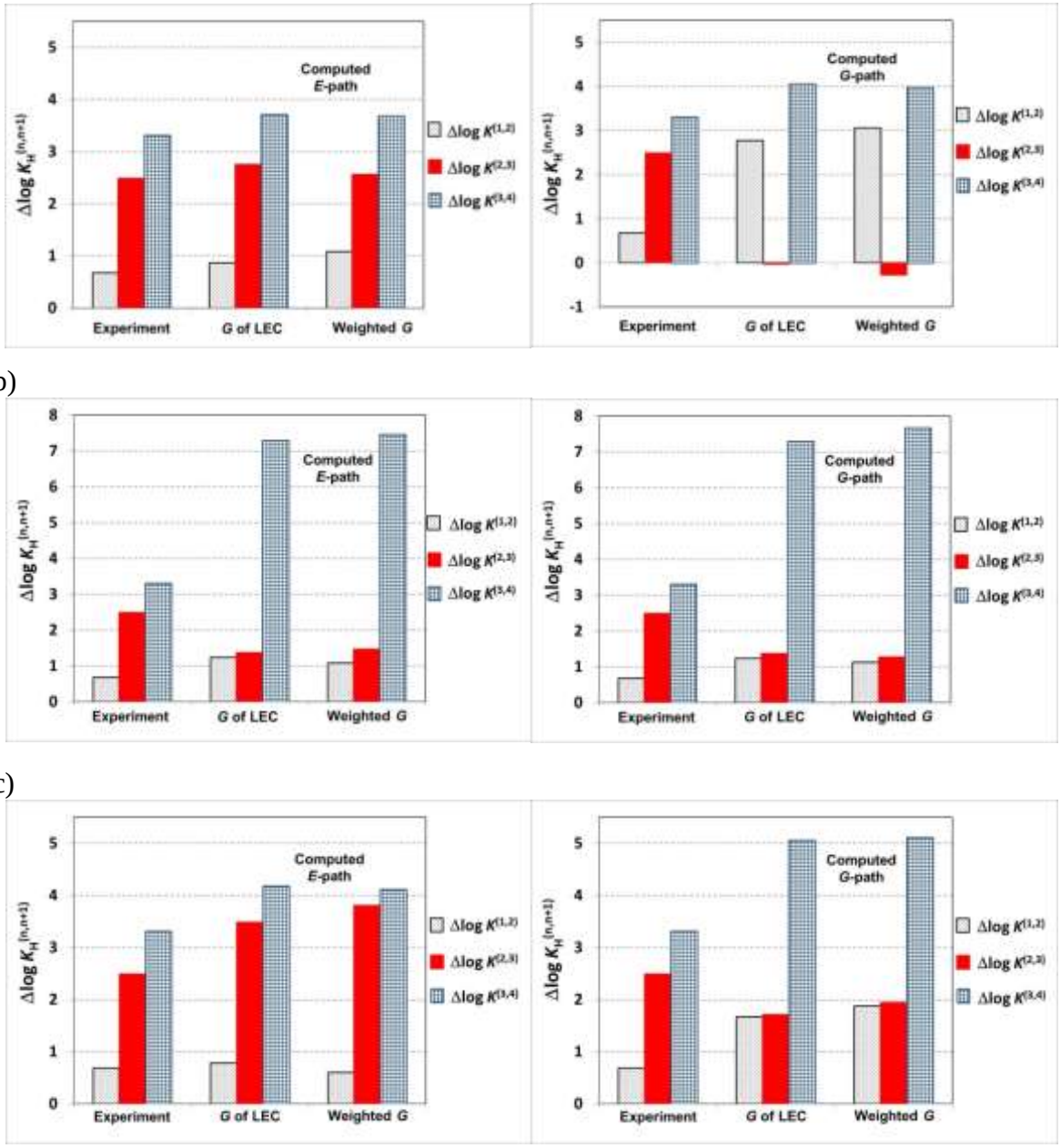


Figure 3 Graphical presentation of differences between successive stepwise protonation constants, $\Delta \log K_{\mathrm{H}}^{(\mathrm{n},\mathrm{n+1})} = \log K_{\mathrm{H}}^{(\mathrm{n})} - \log K_{\mathrm{H}}^{(\mathrm{n+1})}$, for experimental and computed data in: part (a) – DCSM at the B3LYP level, part (b) – CSM at B3LYP level, and part (c) – DCSM with dispersion corrected B97D level of theory, all with the 6-311++G(d,p) functional.

low frequency (*i.e.* soft) vibrational modes. Consequently, these low frequency vibration modes contribute significantly to increased thermal entropy and lower ZPVE contributions to the Gibbs free energy of a molecular system [51–53]. Also, the inability of the RRHO model to correctly evaluate vibrational frequencies, especially for such low vibrational modes, may compound this problem since thermal corrections to electronic energies depend on computed

vibrational frequencies [51]. To correct for this effects, one would have to introduce the anharmonic correction, specifically for those identified at low frequency modes (*i.e.* using the so-called quasi-harmonic model) and this is not a trivial task especially when the DCSM is utilized to describe solvent environment.

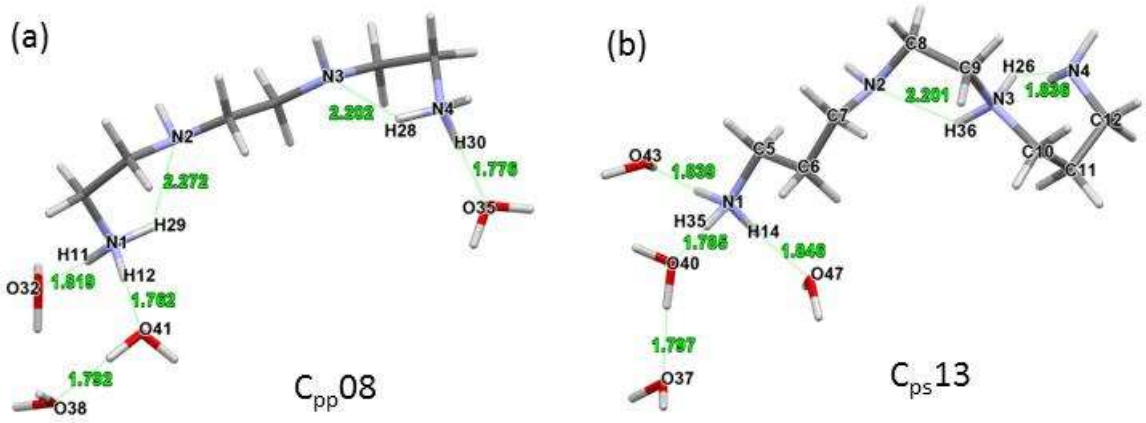


Figure 4 LECs selected from the *G*-path for H_2L of (a) 2,2,2-tet and (b) 3,2,3-tet.

Furthermore, other factors such as the coupling of rotational and vibrational modes in solution, particularly when discrete solvent molecules are included, may also affect the accuracy of computed thermal corrections [9]. The overall effect of all these factors is such that, in most cases, the LEC selected by the *G*-path dominates the conformer population even though it has a higher electronic energy (*i.e.* it is less stable) than the one selected from the *E*-path. A similar discrepancy between Gibbs free energy and electronic energy based selection of LECs has been reported by Salehzadeh et al [16] in their study of micro protonation constants of spermine.

Considering results obtained in computationally least expensive medium, CSM, data in Table 2 and Figure 3(b) shows that although the sequence in protonation constants has been reproduced correctly by both, *E*- and *G*-paths, the results obtained for (i) the fourth protonation constant which was underestimated by more than four log units and (ii) the difference between the second and third protonation constants, which is much too small, must be seen as unacceptable. Moreover, the overall mean absolute deviation (1.72) of predicted stepwise protonation constants in CSM is over four times larger than that obtained in DCSM (0.37). This is not entirely surprising as (i) proper modelling of the solvation environment determines to a large extent the accuracy of computed protonation constants in general [9,41–44] and (ii) reasonable computational evaluation of the solution free energy for highly charged ionic species, such as H_3L and H_4L forms of aliphatic polyamines, is usually prone to

large errors unless explicit water molecules are incorporated to describe the first solvation layer [9,41].

Interestingly, we noted that in the case of the *G*-path implemented in CSM, the second protonation constant was correctly predicted to be larger than the third one whereas the same protocol failed in this respect when implemented in DCSM. In search for possible origin of this observation we examined relevant conformers which were selected from these two solvation models. Considering DCSM, we found that, in line with experimental observations [5,7], the LECs of 2,2,2-tet were mainly those of HL_p and H_2L_{pp} tautomers of the HL and H_2L forms, respectively. In contrast, the LECs of 3,2,3-tet with largest %-fraction of the total population were those of HL_p and H_2L_{ps} tautomers. Therefore, differences in charge distribution on conformers of H_2L_{pp} and H_2L_{ps} tautomers selected for 2,2,2-tet and 3,2,3-tet, respectively, in combination with uneven water molecules' distribution might be responsible for inaccuracy of the second protonation constant of 2,2,2-tet when *G*-path was followed in DCSM. This correlates well with previous studies where it has been pointed out that similarity of charge and its distribution between a reference molecule and the molecule of interest appears to be of utmost importance in accurate prediction of protonation constants using the CRn methodology [12,13].

It is also important to note that with CSM (part (b) in Table 2), there is no apparent difference in predicted protonation constants using either *E*- or *G*-paths. This is due to the fact that the computed electronic and Gibbs free energies of the LECs followed exactly the same trends in relative values; hence, the selected sets of LECs with %-fraction > 5% from *E*- and *G*-paths were very much the same (a feature which is not observed in DCSM). Clearly, the absence of explicit water molecules eliminated all the above mentioned complications and uncertainties in computed *G* values.

Finally, we also tested whether accuracy of predicted protonation constants could be improved by accounting for dispersion interactions as their importance in obtaining accurate thermochemical parameters has been emphasized recently [54]. To accomplish this, we re-optimised all conformers with %-fractions > 5 found in DCSM at the B97D and B3LYP-gD3 levels of theory, both with 6-311++G(d,p) functional. For both levels of theory protonation sequence was predicted correctly from both, *E*- and *G*-paths (Table S3 in the SI) but overall results obtained at B97D are much better than those at B3LYP-gD3 – see Figure S24 in the SI; hence, we will focus on the former. In general, one could consider B97D-predicted protonation constants as reasonable as, on average, the departure in absolute terms from experimental $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values for all protonation constants from *E*- and *G*-paths combined was

0.9 ± 0.5 log units with the largest deviations found for the first protonation constant which was overestimated by about 1.6 log units. However, it has been pointed out [55] that in certain instances, addition of empirical dispersion correction accounts properly for short-range (intramolecular) but fails for long-range (intermolecular) dispersion effects. This results in imbalance between intra- and intermolecular dispersion effects on electronic structure which might be responsible for larger errors in computed protonation constants when compared with dispersion-uncorrected B3LYP functional. In addition, accuracy in $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values obtained using the B3LYP functional might be also due to hidden error cancellations [56,57], a unique situation for 'electronically simple' molecules (such as aliphatic polyamines) and, as such, our results do not preclude the use of dispersion corrected functionals when carrying out this kind of investigation on other molecules. However, one must also realize that full-scale comparative studies on entire sets of all conformers found from MM-search using several functionals would be even more time demanding and because results obtained here are very satisfactory, one would have to justify if it is really worthwhile to strive for a small, a fraction of a log unit, just possible but not guaranteed improvement for the third and fourth protonation constant.

4.4. Testing reliability of pre-optimisation protocol.

Even though the developed general purpose protocol provided theoretically predicted protonation constants (i) of outstanding quality relative to typically reported data in the computational field, (ii) appears to work well for molecules with multiple positive charges and (iii) can be seen as reliable to provide a valuable insight on relevant properties of 2,2,2-tet for a solution chemist, we decided to test it further. Clearly, prior to recommending any protocol as of general purpose, it is important to find out whether some LECs were missed and, if this was the case, what impact on quality of computed $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values that would have.

To this effect, we have focused on structures with explicit water molecules as the best and reliable results were obtained only in DCSM and decided to fully energy optimise the 'redundant' conformers which were rejected after the pre-optimisation. Firstly, we wanted to find out whether (i) a new and the lowest in electronic energy conformer could be discovered and secondly (ii) new conformers would have to be included in the LECs sets, within 2 kcal/mol window, which had to be used in computing $\log K_{\mathrm{H}}^{(\mathrm{n})}$ in case of 'Weighted *G*' strategy. Data in Table S4 in the SI shows that:

a) In all cases the lowest electronic energy conformer has been identified from the pre-optimisation protocol. This is gratifying finding because, as pointed out above (Table 2) it is sufficient to use the single *G* value from *E*-path to obtain excellent prediction in the $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values (recall that they hardly differ from those obtained using computationally more expensive weighted *G* values of the selected LECs).

b) Only in one case we found an additional conformer, that of H_2L_{ps}, which was within the 2 kcal/mol window of conformers to be selected for computing protonation constants. It is important to stress here that this conformer was not the lowest in electronic energy; hence, it could only influence results obtained from *E*-path involving weighted *G*.

The newly discovered H_2L_{ps} conformer of 2,2,2-tet has changed the number of LECs from two (from pre-optimisation) to three (after full optimisation of all conformers) – see Figure 5. These three conformers were combined with four H_2L_{pp} LECs of 2,2,2-tet (this set has not changed after full optimisation) and those with %-fraction above 5% (from Boltzmann distribution done on seven combined conformers) were used to compute protonation constants. In other words, the protocol developed here and described in details in proceeding sections was fully followed and we found that this conformer was predicted to contribute 6% to the total population when free energies of seven LECs were used.

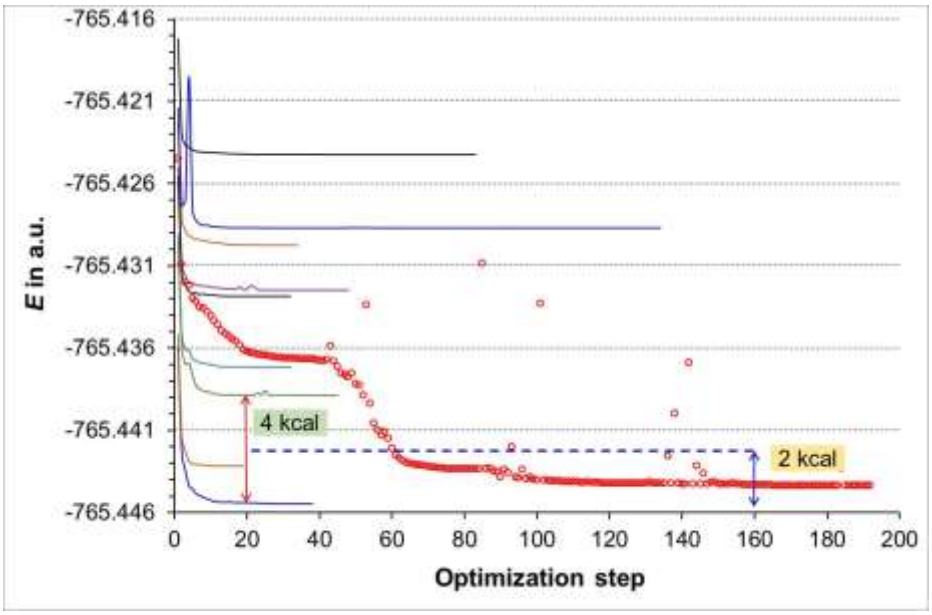


Figure 5 Optimisation profile for conformers of 2,2,2-tet in DCSM showing the 4 kcal/mol *E*-window at twentieth step used to select structures for full optimisation when the pre-optimisation protocol was implemented and 2 kcal/mol *E*-window to select conformers required to compute protonation constants.

One must note that two protonation constants, for H_2L and H_3L forms of 2,2,2-tet, had to be re-computed because $H_2L^{(1)}$ is involved in two protonation reactions, $HL^{(1)} + H_2L^{(2)} = H_2L^{(1)} +$

$HL^{(2)}$ (for $\log K_H^{(2)}$) and $H_2L^{(1)} + H_3L^{(2)} = H_3L^{(1)} + H_2L^{(2)}$ (for $\log K_H^{(3)}$). The values of the re-calculated protonation constants changed by ±0.07 log unit; the second decreased by ~0.08 log unit whereas the third protonation constant increased by ~0.06 log unit. Clearly, this had no effect on the overall quality of $\log K_H^{(n)}$ values as well as the sequence of stepwise protonation constants.

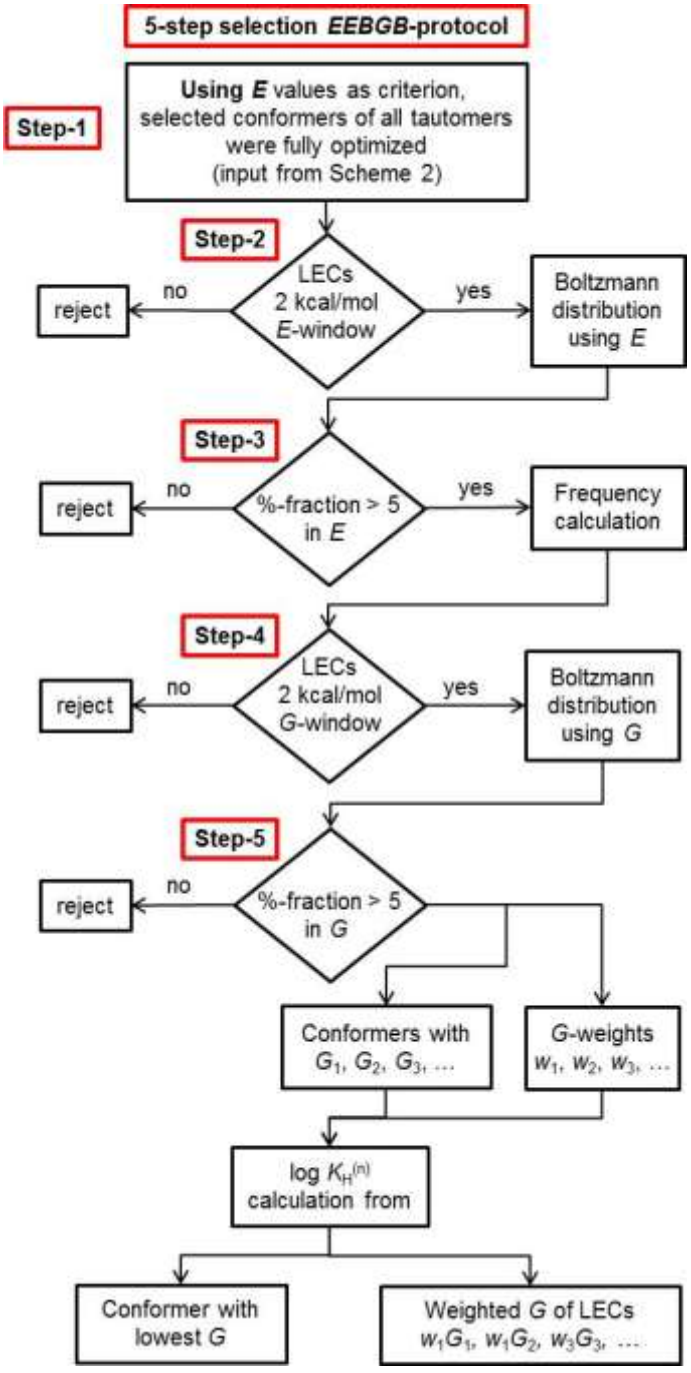
Just for completeness, we have also found one new conformer, this time for HL_p of 2,2,2-tet, when full optimisation data in CSM was analysed and it is seen in Figure S3(b) in PART 4 of the SI as empty circles. In this instance, it had no influence on predicted protonation constants as its energy was at the border line of the 2 kcal/mol *E*-window.

4.5. Recommended protocols for protonation constants calculations

Data in Table S4 in the SI gave us confidence to reanalyse the entire set of optimised structures in DCSM as we wanted to come up with the final protocol(s) which would generate excellent results with minimum computational time. Hence, our focus was on fully optimised conformers which were selected from Scheme 2 and fine-tuning of steps implemented in *E*-path shown in Scheme 3. Analysis in variation in electronic energies and their influence on selected conformers lead to the final protocol shown in Scheme 4. We call it a 5-step selection *EEBGB*-protocol because it incorporates two steps involving *E*-based selection (one in the pre-optimisation operation and the second after full optimisation), Boltzmann distribution using *E* values to select conformers for frequency calculations, followed by *G*-based selection of conformers within 2 kcal/mol window of the free energy, and the final step from which, based on Boltzmann distribution using *G*-values of retained conformers, only those with %-fraction > 5 were retained. As an example, we will illustrate the performance of the *EEBGB*-protocol, in terms of reduction of the number of conformers after each step, using data obtained for the H_4L of 3,2,3-tet:

- **Step-1 (*E*-based selection)**: out of 30 MM-identified LECs, 28 were within 4 kcal/mol *E*-window after the pre-optimisation operation and they were fully optimised.
- **Step-2 (*E*-based selection)**: out of 28 fully optimised conformers, 14 were selected which were within 2 kcal/mol *E*-window. Note that in case of HL (HL_p and HL_s) and H_2L (H_2L_{pp} and H_2L_{ps}) tautomers are combined after full optimisation and only conformers within the 2 kcal/mol *E*-window are selected.
- **Step-3 (*B*-based selection)**: using *E* values of 14 selected conformers, 7 met the criterion of Boltzmann distribution %-fraction > 5; they were submitted for frequency calculations.

- **Step-4 (*G*-based selection)**: Only conformers within 2 kcal/mol *G*-window were selected; the set of 6 structures with their G_1, G_2, G_3, …G_n values (G_1 is the lowest in the free energy) were submitted for the Boltzmann distribution calculation.
- **Step-5 (*B*-based selection)**: using G_n values of retained 6 conformers, a selection criterion of Boltzmann distribution generated %-fraction > 5 was applied to obtain the FINAL set of conformers. In this particular case all 6 were retained and their fraction contributions were used as weights w_n.



Scheme 4 Recommended and time most-effective 5-step selection *EEBGB*-protocol for protonation constants calculations of polyamines.

Finally, the selected 6 conformers were used to compute protonation constants using either the single value of *G* of the lowest in the free energy conformer or weighted *G* value obtained by pairing G_n and w_n values ($w_1 \times G_1 + w_2 \times G_2 + \ldots + w_n \times G_n$).

This protocol has decreased the initial number of H_4L conformers from 30 MM-generated in Step-1 to 7 which were submitted for frequency calculations and 6 for protonation constants calculations. The same *EEBGB*-protocol (Scheme 4) was implemented for each tautomer in the DCSM (a complete set of data for each tautomer is presented in Table S5 in the SI) and this resulted in:

a) 94% reduction in the number of conformers submitted for frequency calculations, from initial 420 MM-generated to 25 in the Step-5;
b) 75% reduction relative to the *G*-path which can be seen as a 3-step selection *EGB*-protocol – see Scheme S2 in the SI where 118 conformers were submitted for frequency calculations;
c) An additional 30% reduction of frequency calculations relative to the *E*-path shown in Scheme 3, which is a 4-step selection *EEGB*-protocol – see Scheme S3 in the SI. Importantly, this had no detrimental effect on the computed protonation constants at all - see Table 3.

We have also tested a protocol where the only selection criterion was variation in electronic energy - it can be seen as a 2-step selection *EE*-protocol shown in Scheme S4 in the SI. Interestingly, the computed protonation constants, as stepwise $\log K_H^{(n)}$ values, resulted in 10.39, 8.24, 5.55, and 3.15 (note that they follow protonation sequence correctly) with differences from experimental values of 0.64, –0.83, –1.03 and –0.12, respectively (0.7±0.3 log units for absolute differences from experimental values). From this follows that to compute preliminary but still reasonable estimates of stepwise protonation constants (they are at least as good, if not better, when commonly reported from thermodynamic cycles) it is sufficient to find the lowest in electronic energy conformer among possible tautomers and use their *G*-values as components in each stepwise competition reaction, $H_{n-1}L^{(1)} + H_nL^{(2)} = H_nL^{(1)} + H_{n-1}L^{(2)}$. To appreciate simplicity of the latter *EE*-protocol, one can write a general expression for the free energy change of competition reaction applicable to each *n*th protonation step

$$\Delta G^{n}_{CRn} = {}^{*}G(H_nL_t^{(1)}) + {}^{*}G(H_{n-1}L_t^{(2)}) - {}^{*}G(H_{n-1}L_t^{(1)}) - {}^{*}G(H_nL_t^{(2)}) \qquad (9)$$

where *G stands for the free energy of the lowest electronic energy conformer found among tautomers of each protonated form of two polyamines, the one under investigation ($L^{(1)}$) and that used as a reference molecule ($L^{(2)}$). In contrast, when one would need to use weighted G values of LECs found from Scheme 4 then the following expression for $\Delta G^{n}_{\mathrm{CRn}}$ applies

$$\Delta G^{n}_{\mathrm{CRn}} = \sum_{k=1}^{t}\sum_{x=1}^{\mathrm{LEC}} w_x G_x(\mathrm{H_n L^{(1)}_t}) + \sum_{m=1}^{t}\sum_{y=1}^{\mathrm{LEC}} w_y G_y(\mathrm{H_{n-1} L^{(2)}_t}) - \sum_{n=1}^{t}\sum_{z=1}^{\mathrm{LEC}} w_z G_z(\mathrm{H_{n\text{-}1} L^{(1)}_t}) - \sum_{o=1}^{t}\sum_{s=1}^{\mathrm{LEC}} w_s G_s(\mathrm{H_n L^{(2)}_t}) \quad (10)$$

where symbols are as described for expressions 3 and 4. Note that expression 10 is equally applicable to *E*- and *G*-paths in Scheme 3 as well as the refined *EGB*-, *EEGB* and *EEBGB*-protocols with the only, but significant, difference in the decreasing number of LECs obtained from the final third, fourth and fifth selection step, respectively.

Table 3. PART (a) Comparison of theoretically computed four stepwise protonation constants using the recommended and time most-efficient 5-step selection *EEBGB*-protocol and, for comparison shown in brackets, second time-efficient 4-step selection *EEGB*-protocol. PART (b) Averaged values from two methods (*G* of LEC and weighted *G*) of *EEGB*- and *EEBGB*-protocols.[a]

PART (a)

Step-wise protonation constant	**5-step selection *EEBGB*-protocol (4-step selection *EEGB*-protocol)**			
	***G* of LEC**	**Δ**	**Weighted *G***	**Δ**
$\log K^{(1)}_{\mathrm{H}}$	9.76 (9.74)	0.01 (–0.01)	9.94 (9.83)	0.19 (0.08)
$\log K^{(2)}_{\mathrm{H}}$	8.87 (8.87)	-0.20 (–0.20)	8.75 (8.75)	-0.32 (–0.32)
$\log K^{(3)}_{\mathrm{H}}$	6.12 (6.12)	-0.46 (–0.46)	6.19 (6.19)	-0.39 (–0.39)
$\log K^{(4)}_{\mathrm{H}}$	2.44 (2.41)	-0.83 (–0.86)	2.65 (2.50)	-0.62 (–0.77)
	Average|Δ|	0.37 (0.38)	Average|Δ|	0.38 (0.39)

PART (b)

Step-wise protonation constant	**Average from 2 methods of *EEGB*- and *EEBGB*-protocols**	
	Value	**Δ**
$\log K^{(1)}_{\mathrm{H}}$	9.82	0.07
$\log K^{(2)}_{\mathrm{H}}$	8.81	-0.26
$\log K^{(3)}_{\mathrm{H}}$	6.16	-0.42
$\log K^{(4)}_{\mathrm{H}}$	2.50	-0.77
Average|Δ|		0.38
Standard deviation|Δ|		0.30

[a] Δ = computed – experimental $\log K^{(n)}_{\mathrm{H}}$ value.

5. CONCLUSIONS

This work has demonstrated that it is possible as well as time-wise and computationally feasible to theoretically predict stepwise protonation constants, as $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values, of polyamines with the largest error smaller than 1 log unit, relative to the experimentally determined values from glass electrode potentiometry (GEP) which is most accurate among all analytical techniques in the field. In this particular case, where 2,2,2-tet was investigated, the predicted first protonation constant can be seen as of GEP-analytical quality as it differs from the experimental $\log K_{\mathrm{H}}^{(1)}$ value by less than ±0.1 log unit whereas the second and third might be seen as of NMR-analytical quality because they were predicted to within 0.2–0.4 log units of the GEP experimental values. Deviation from experimental values was systematic and unidirectional when going from the second $\log K_{\mathrm{H}}^{(2)}$ to the fourth $\log K_{\mathrm{H}}^{(4)}$ value (they all were underestimated) and the largest deviation, of about –0.8 log unit, was observed for the $\log K_{\mathrm{H}}^{(4)}$ value. It is important to stress that these results (which we see as of excellent overall quality) were obtained even though aliphatic linear polyamines are characterised by (i) numerous tautomers, (ii) almost an infinite number of possible conformers for each tautomer and (iii) very small, often well below 1 log unit, differences between consecutive protonation constants. Regarding the latter point, protocols developed here were also able to predict the values in correct order, in each case $\log K_{\mathrm{H}}^{(\mathrm{n})} > \log K_{\mathrm{H}}^{(\mathrm{n+1})}$ was reproduced as observed from experimental data.

Considering the quality of computed protonation constants we attribute this to successful implementation of the competition reaction (CRn) based methodology which requires (i) a polyamine under investigation, here $\mathrm{L}^{(1)}$ = 2,2,2-tet, and reference molecule, here $\mathrm{L}^{(2)}$ = 3,2,3-tet, to be structurally similar, (ii) correct selection of lowest energy conformers of all possible $\mathrm{H_nL}^{(1)}$ and $\mathrm{H_nL}^{(2)}$ tautomeric forms and (iii) balanced charge distribution between reactants and products, which in this case translates to $\mathrm{H_{n-1}L}^{(1)}$ and $\mathrm{H_nL}^{(2)}$ to be involved in a stepwise CRn, $\mathrm{H_{n-1}L}^{(1)} + \mathrm{H_nL}^{(2)} = \mathrm{H_nL}^{(1)} + \mathrm{H_{n-1}L}^{(2)}$, used to compute the $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values. Furthermore, this work has shown that it is not only sufficient to select lowest in electronic energy conformers (their Gibbs free energy values, *G*, are used in computing protonation constants) but it has resulted in higher quality of the $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values relative to *G*-based selection of LECs.

Regarding highly improved time and computational feasibility of theoretically predicting stepwise protonation constants, this has been achieved by implementing thoroughly

investigated selection protocols developed in this work. The proposed *EEBGB*-protocol (*E*, *B* and *G* stand for electronic-energy-, Boltzmann-distribution- and Gibbs-free-energy-based stepwise selection of conformers – see Scheme 4) resulted in the 94% reduction of conformers submitted for frequency calculations from which four protonation constants were calculated, from initial 420 conformers selected from MM-based conformational search, to 25 in the final Step-5 of this protocol. Further reduction in time has been achieved by selecting conformers from an accelerated 'optimisation' operation, *i.e.*, instead of fully energy optimise all 420 MM-selected conformers, they were subjected to pre-optimisation involving only first 20 optimisation steps in Gaussian. Two important comments are in order here: (i) although we have verified validity of the accelerated 'optimisation' protocol by full optimisation of all, 420 2,2,2-tet and 3,2,3-tet structures, there is no guarantee that for larger polyamines (like penta- or hexamines) initial 20 optimisation steps will work perfectly well (one would have to consider either increasing the number of initial steps or perform full optimisation) and (ii) the pre-optimisation step will only influence time required for the first selection step in the developed *EEBGB*-protocol; the overall efficiency in the reduction of the number of conformers subjected to the frequency calculation remains intact. The reduced number of time-demanding frequency calculations is beneficial because, as this work shown, involving explicit water molecules significantly improves predictions in protonation constants.

Let us now comment on the systematic departure of computed $\log K_{\mathrm{H}}^{(\mathrm{n})}$ values from experimental ones. We attribute this to intrinsic errors in computed energies when charges on molecules increase. The possible solution is to implement a stepwise increase in the number of explicit water molecules to dissipate the charge throughout the macromolecular assembly, ($L + nH_2O$) from four H_2O molecules for the singly protonated tetramines (this resulted here in excellent prediction of $\log K_{\mathrm{H}}^{(1)}$ value) to, *e.g.*, seven H_2O molecules when H_3L and H_4L are involved (to compute the $\log K_{\mathrm{H}}^{(4)}$ value). It is also reasonable to assume that in case of polyamines with a larger number of protonation sites one should also need to increase the number of explicit water molecules.

Finally, it would be of fundamental importance to investigate an impact the functionals, such as B97D or the latest B3LYP-gD3, can make on the quality of computed protonation constants. To achieve that one would have to use them from the very beginning of the proposed protocol, namely all conformers selected from MM-based search would have to be (pre)optimised using dispersion-included functional. In our opinion, however, regardless of all the above comments related to feature studies, the proposed *EEBGB*-protocol can be

successfully used and we are also of an opinion that its applicability is not restricted to polyamines.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Batista de Carvalho LAE, Marques MPM, Tomkinson J (2006) Transverse Acoustic modes of biogenic and α,ω-polyamines: a study by inelastic neutron scattering and Raman spectroscopies coupled to DFT calculations. J Phys Chem A 110:12947–12954

[2] Agostinelli E, Marques MPM, Calheiros R, Gil FPSC, Tempera G, Viceconte N, Battaglia V, Grancara S, Toninello A (2010) Polyamines: fundamental characters in chemistry and biology. Amino Acids 38:393–403.

[3] Marques MPM, Batista de Carvalho LAE (2007) Vibrational spectroscopy studies on linear polyamines. Biochem Soc Trans 35:374–380.

[4] Cooper GJS (2011) Therapeutic potential of copper chelation with triethylenetetramine in managing diabetes mellitus and Alzheimer's disease. Drugs 71:1281–1320.

[5] Borkovec M, Cakara D, Koper GJM (2012) Resolution of Microscopic Protonation Enthalpies of Polyprotic molecules by means of cluster expansions. J Phys Chem B 116:4300–4309.

[6] Cukrowski I, Matta CF (2011) Protonation sequence of linear aliphatic polyamines from intramolecular atomic energies and charges. Comput Theoret Chem 966:213–219.

[7] Hague DN, Moreton AD (1994) Protonation Sequence of Linear Aliphatic Polyamines by ^{13}C NMR Spectroscopy. J Chem Soc Perkin Trans 2:265–270.

[8] Casasnovas R, Ortega-Castro J, Frau J, Donoso J, Munoz F (2014) Theoretical pK_a calculations with continuum model solvents, alternative protocols to thermodynamic cycles. Int. J. Quantum Chem. 114:1350–1363.

[9] Ho J (2014) Predicting pK_a in implicit solvents: current status and future directions. Aust. J. Chem. 67:1441–1460.

[10] Ho J, Coote ML (2011**)** First-principles prediction of acidities in the gas and solution phase. Comput. Mol. Sci. 1:649–660, doi:10.1002/WCMSW.43.

[11] Namazian M, Halvani S (2006) Calculations of pK_a values of carboxylic acids in aqueous solution using density functional. theory. J.Chem. Thermodyn. 38:1495–1502.

[12] Govender KK, Cukrowski I (2009) Density functional theory in prediction of four stepwise protonation constants for Nitrilotripropanoic acid (NTPA). J. Phys. Chem. A. 113:3639–3647.

[13] Govender KK, 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. 114:1868–1878.

[14] NIST, Standard Reference Database 46. NIST Critically Selected Stability Constants of Metal complexes Database, Version 8.0, Data collected and selected by R.M. Smith and A.E. Martell, U.S. Department of Commerce, National Institute of Standards and Technology,

[15] The IUPAC Stability Constants Database, http://www.iupac.org distributed and maintained by Academic Software, Sourby Old Farm, Timble, Otley, Yorks, LS21 2PW, U.K. (http://www.acadsoft.co.uk/scdbase/).

[16] Salezadeh S, Gholiee Y, Bayat M (2011) Prediction of microscopic protonation constants of polybasic molecules via computational methods: a complete microequilibrium analysis of spermine. Int. J. Quantum Chem. 111:3608–3615.

[17] Saracino GAA, Improta R, Barone V (2003) Absolute pK_a determination for carboxylic acids using density functional theory and the polarizable continuum model. Chem. Phys. Lett. 373:411–415.

[18] Schuurmann G, Cossi M, Barone V, Tomasi J (1998) Prediction of the pK_a of carboxylic acids using the ab initio continuum-solvation model PCM-UAHF. J. Phys. Chem. A. 102:6706–6712.

[19] Namazian M, Heidary H (2003) Ab initio calculations of p*K*a values of some organic acids in aqueous solution. J. Mol. Struct. (THEOCHEM) 620:257–263.

[20] Namazian M, Halvani S, Noorbala MR (2004) Density functional theory response to the calculations of pK_a values of some carboxylic acids in aqueous solution. J. Mol. Struct. (THEOCHEM) 711:13–18.

[21] Namazian M, Kalantary-Fotooh F, Noorbala MR, Searles DJ, Coote ML (2006) Møller–Plesset perturbation theory calculations of the pK_a values for a range of carboxylic acids. J. Mol. Struct. (THEOCHEM) 758:275–278.

[22] Namazian M, Zakery M, Noorbala MR, Coote ML (2008) Accurate calculation of the pK_a of trifluoroacetic acid using high-level Ab initio calculations. Chem. Phys. Lett. 451:163–168.

[23] Charif IE, Mekelleche SM, Villemin D, Mora-Diez N (2007) Correlation of aqueous pK_a values of carbon acids with theoretical descriptors: a DFT study. J. Mol. Struct. (THEOCHEM) 818:1–6.

[24] Liptak MD, Shields GC (2001) Experimentation with different thermodynamic cycles used for pK_a calculations on carboxylic acids using complete basis set and Gaussian-n models combined with CPCM continuum solvation methods. Int. J. Quantum Chem. 85:727–741.

[25] Liptak MD, Shields GC (2001) Accurate pK_a calculations for carboxylic acids using complete basis set and Gaussian-n models combined with CPCM continuum solvation methods. J. Am. Chem. Soc. 123:7314–7319.

[26] Silva CO, Silva EC, Nascimento MAC (2000) Ab initio calculations of absolute pK_a values in aqueous solution II. Aliphatic alcohols,thiols, and halogenated carboxylic acids. J. Phys. Chem. A. 104:2402–2409.

[27] Chipman DM (2002) Computation of pK_a from dielectric continuum theory. J. Phys. Chem. A. 106:7413–7422.

[28] Sastre S, Casasnovas R, Munoz F, Frau J. (2013) Isodesmic reaction for pK_a calculations of common organic molecules. Theor Chem Acc 132:1310

[29] Riojas AG, Wilson AK (2014) Solv-ccA:Implicit solvation and the correlation consistent composite approach for the determination of pK_a. J. Chem. Theory Comput. 10:1500–1510.

[30] Pliego JR, Riveros JM (2002) Theoretical calculation of pK_a using the cluster−continuum model. J. Phys. Chem. A. 106:7434–7439.

[31] Klicic JJ, Friesner RA, Liu S, Guida WC (2002) Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods. J. Phys. Chem. A. 106:1327–1335.

[32] Adam KR (2002) New density functional and atoms in molecules method of computing relative pK_a values in solution. J. Phys. Chem. A. 106:11963–11972.

[33] Wiberg KB, Clifford S, Jorgensen WL, Frisch MJ Origin of the inversion of the acidity order for haloacetic acids on going from the gas phase to solution (2000) J. Phys. Chem. A. 104:7625–7628.

[34] Sang-Aroon W, Ruangpornvisuti V (2008), Determination of aqueous acid dissociation constants of aspartic acid using PCM/DFT method. Int. J. Quantum Chem. 108:1181–1188.

[35] Ho J, Klamt A, Coote ML (2010) Comment on the correct use of continuum solvent models. J. Phys. Chem. A. 114:13442–13444.

[36] Ho J (2015) Are thermodynamic cycles necessary for continuum solvent calculation of pK_as and reduction potentials? Phys. Chem. Chem. Phys. 17:2859–2868.

[37] Sutton CR, Franks GV, da Silva G (2012) First principles p*K*a calculations on carboxylic acids using the SMD solvation model: effect of thermodynamic cycle, model chemistry, and explicit solvent molecules. J. Phys. Chem. B. 116:11999–12006.

[38] Afaneh AT, Schreckenbach G, Wang F (2014) Theoretical study of the formation of mercury (Hg^{2+}) complexes in solution using an explicit solvation shell in implicit solvent calculations. J. Phys. Chem. B. 118:11271–11283.

[39] Bryantsev VS, Diallo MS, Goddard III WA (2007) pK_a calculations of aliphatic amines, diamines, and aminoamides via density functional theory with a Poisson−Boltzmann continuum solvent model. J. Phys. Chem. A. 111:4422–4430.

[40] Adeyinka AS, Cukrowski I (2015) Structural-topological preferences and protonation sequence of aliphatic polyamines: a theoretical case study of tetramine trien. J. Mol Model. 21:162–180.

[41] Eckert F, Diedenhofen M, Klamt A (2010) Towards a first principles prediction of pK_a: COSMO-RS and the cluster-continuum approach. Mol. Phys. 108:229–241.

[42] Marenich AV, Ding W, Cramer CJ, Truhlar DG (2012) Resolution of a challenge for solvation modeling: calculation of dicarboxylic acid dissociation constants using mixed discrete–continuum solvation models. J. Phys. Chem. Lett. 3:1437–1442.

[43] Klamt A, Eckert F, Diedenhofen M, Beck ME (2003) First principles calculations of aqueous pK_a values for organic and inorganic acids using COSMO−RS reveal an inconsistency in the slope of the pK_a scale. J. Phys. Chem. A. 107:9380–9386.

[44] Kelly CP, Cramer CJ, Truhlar DG (2006) Adding explicit solvent molecules to continuum solvent calculations for the calculation of aqueous acid dissociation constants. J. Phys. Chem. A. 110:2493–2499.

[45] Mennucci B (2010) Continuum solvation models: what else can we learn from them? J. Phys. Chem. Lett. 1:1666–1674.

[46] Tomasi J, Persico M (1994) Molecular interactions in solution: An overview of methods based on continuous distributions of the solvent. Chem. Rev. 7:2027–2094.

[47] Spartan'10, version 1.1.0 (2010) Wavefunction, Inc., 18401 Von Karmen Ave., Suite 370, Irvine, CA92612, USA.

[48] Frisch MJ, Trucks GW, Schlegel HB et al (2009) Gaussian 09, Revision D.1, Gaussian, Inc., Wallingford CT.

[49] A. E. Frisch (1998) Gaussian 09 User's Reference, Gaussian Inc., Pittsburgh, PA,

[50] Cukrowski I, Govender KK (2010) A density functional theory- and atoms in molecules-based study of NiNTA and NiNTPA complexes toward physical properties controlling their stability. A new method of computing a formation constant. Inorg. Chem. 49:6931–6941.

[51] Temelso B, Shields GC (2011) The role of anharmonicity in hydrogen-bonded systems: The case of water clusters. J. Chem. Theory Comput. 7:2804–2817.

[52] Njegic B, Gordon MS (2006) Exploring the effect of anharmonicity of molecular vibrations on thermodynamic properties, J. Chem. Phys. 125:224102,1-12.

[53] Jinich A, Rappoport D, Dunn I, Sanchez-Lengeling B, Olivares-Amaya R, Noor E, Bar Even A, Aspuru-Guzik A (2014) Quantum chemical approach to estimating the thermodynamics of metabolic reactions. Sci. Rep. 4:7022, 1-6 DOI: 10.1038/srep07022.

[54] Grimme S (2011) Density functional theory with London dispersion corrections. Comput. Mol. Sci. 1:211–228, DOI i: 10.1002/WCMSW.30.

[55] Steinmann SN, Csonka G, Cominboeuf C (2009) Unified inter- and intramolecular dispersion correction formula for generalized gradient approximation density functional theory. J. Chem. Theory Comput. 5:2950–2958.

[56] Kozuch S, Bachrach SM, Martin JML (2014) Conformational equilibria in Butane-1,4-diol: A benchmark of a prototypical system with strong intramolecular H‑ bonds. J. Phys. Chem. A. 118(1):293–303.

[57] Boese AD (2015) Density functional theory and hydrogen bonds: Are we there yet? ChemPysChem. 16:978–985.