How Structural Details Influence the Result of pK_A Calculations in Proteins

Dissertation zur Erlangung des akademischen Grades des

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie

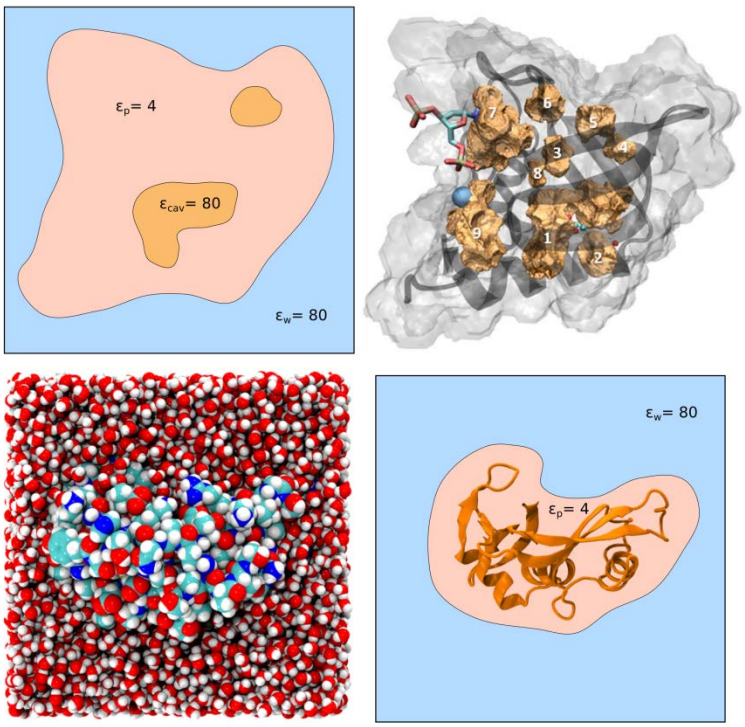
der Freien Universität Berlin

vorgelegt von

Tim Meyer

aus Ludwigsburg

April 2015



Die vorliegende Arbeit wurde unter Anleitung von Prof. Dr. E. W. Knapp im Zeitraum 01.04.2010-01.06.2015 am Institut für Chemie/Kristallographie der Freien Universität Berlin im Fachbereich Biologie, Chemie und Pharmazie durchgeführt.

1. Gutachter: Prof. Dr. Ernst-Walter Knapp, Freie Universität Berlin

2. Gutachterin: Prof. Dr. Maria Andrea Mroginski, Technische Universität Berlin

Disputation am 12.06.2015

Table of Content

1. **Publications**

The work presented in this thesis has been partially published in the following articles:

Meyer, T.; Knapp, E. W.
pK_A Values in Proteins Determined by Electrostatics Applied to Molecular Dynamics Trajectories
J. Chem. Theory Comput. **2015**, 11 (6), pp 2827–2840

Meyer, T.; Kieseritzky, G.; Knapp, E. W.
Electrostatic pKa Computations in Proteins: Role of Internal Cavities
Proteins: Struct., Funct., Bioinf. **2011,** 79, 3320-3332.

Mair, C. M.; Meyer, T.; Schneider, K.; Huang, Q.; Veit, M.; Herrmann, A.,
A Histidine Residue of the Influenza Virus Hemagglutinin Controls the pH Dependence of the Conformational Change Mediating Membrane Fusion.
J. Virol. **2014,** 88, 13189-13200.

During the time period of this thesis, I also conducted and participated in further research projects that lead to the following publications:

- Meyer, T.; Knapp, E. W.
  Database of Protein Complexes with Multivalent Binding Ability: Bival-Bind
  Proteins: Struct., Funct., Bioinf. **2014**, 82, 744-751.
- Woelke, A. L.; Wagner, A.; Galstyan, G.; Meyer, T.; Knapp, E. W.
  Proton Transfer in the K-Channel Analog of B-Type Cytochrome c Oxidase from Thermus thermophilus
  Biophys. J. **2014**, 107, 2177-2184.
- Woelke, A. L.; Kuehne, C.; Meyer, T.; Galstyan, G.; Dernedde, J.; Knapp, E. W.
  Understanding Selectin Counter-Receptor Binding from Electrostatic Energy Computations and Experimental Binding Studies
  J. Phys. Chem. B **2013**, 117, 16443-16454.
- Woelke, A. L.; Galstyan, G.; Galstyan, A.; Meyer, T.; Heberle, J.; Knapp, E. W.
  Exploring the Possible Role of Glu286 in CcO by Electrostatic Energy Computations Combined with Molecular Dynamics
  J. Phys. Chem. B **2013**, 117, 12432-12441.
- Klippel, S.; Wieczorek, M.; Schuemann, M.; Krause, E.; Marg, B.; Seidel, T.; Meyer, T.; Knapp, E.-W.; Freund, C.
  Multivalent Binding of Formin-Binding Protein 21 (FBP21)-Tandem-WW Domains Fosters Protein Recognition in the Pre-Spliceosome
  J. Biol. Chem. **2011**, 286, 38478-38487.

2. **Introduction**

The proton concentration in solution, described by the pH value, is a well controlled property in biological systems. Of the 20 standard amino acids, six are pH dependent and adjust their protonation in the specific environments. When these so called titratable residues are part of a folded protein, their protonation does not only depend on the pH in the solvent anymore, but also on details of the protein structure around them. The most dominant factors determining the protonation states of amino acids are electrostatic interactions with other amino acids in the protein in the form of salt-bridges and hydrogen bonds. Hydrogen bonds and protonation states of titratable residues on the other hand depend on each other and are crucial for the structural integrity and function of proteins.[1-5]

Titratable residues can also be seen as pH dependent switch devices, which can change protein structure and function. An illustrative example for this relationship is the protein hemagglutinin, an envelope protein of the influenza virus. This protein reacts to a drop in the pH in the late endosome in that the protonation of histidine residues triggers a large conformational change of the protein that leads ultimately to the entry of the virus into the host cell.[6, 7] In chapter 6 of this thesis, the ability of the hemagglutinin protein to 'sense' the pH is discussed in further detail. Especially, a new structural mechanism is suggested that allows a single histidine residue to modulate the pH sensitivity of the whole protein. The relationship between protein structure and pH also works in the opposite direction, as shown for example with the protein cytochrome c oxidase. This protein changes the pH in the cell by pumping protons across a membrane, driven by chemical and structural changes in the protein.[8]

To get insight into the function of many proteins it is crucial to understand the pH dependent protonation of titratable residues. The major part of this thesis, presented in chapters 3-5, is therefore dedicated to the development of computational methods that allow a more accurate computation of pK_A values of amino acids in proteins. The pK_A values of titratable residues are experimentally well known, in case they are individually solvated in water. When they become part of a protein structure their pK_A can be shifted, due to interactions with the protein. As a result of the high complexity and diversity of protein structures, the theoretical computation of this shift is a challenging task. Various approaches had been developed in the past to give insight into the behavior of titratable residue, all comprising different strength and weakness.[9] Some are very fast in terms of computational cost, e.g. those based on an empirical prediction scheme[10], but lack detailed insight into the structural consequences of protonation. Other methods can give this insight like e.g. the constant-pH molecular dynamics approaches[11, 12], but are demanding in terms CPU time. The starting point of the work presented in this thesis was the approach implemented in the software Karlsberg+[13], that employs automatic modeling procedures and electrostatic energy calculations based on solving the Poisson Boltzmann equation to predict pK_A values. The computational cost of this approach is moderate compared

to other methods, it is easy to use and to a certain degree it provides insight into the structural consequences of changes in the protonation pattern.

In this work two novel extensions of Karlsberg+ are presented - each one addressing a weakness of the Karlsberg+ approach - with the aim to improve the accuracy of pK_A calculations. In chapter 3 a new algorithm that locates and models cavities in proteins more faithfully is introduced and a procedure to account for these cavities in electrostatic energy calculations is suggested. In chapter 4 a new procedure that employs molecular dynamic simulations to account for conformation changes of proteins due to changes of the protonation pattern is discussed. Finally, in chapter 5 it is demonstrated that an even more accurate pK_A computation can be achieved by combining the two new procedures. Both methods have been tested with extensive benchmark calculations.

2.1. Calculating pK_A values

For many applications it is important to know the most likely protonation state of titratable residues and the energy necessary for changing it. This can be described by the pK_A value of the molecule. In this work only pK_A values of amino acids are discussed, nevertheless the equations below are valid for any kind of titratable molecule.

pK_A values

In short, the pK_A is the pH value at which the probability of a molecule to be protonated is 50%. It can be derived for a generic acid HA that dissociates to the deprotonated acid A^- and a proton H^+. The chemical reaction of this protonation-deprotonation process can be written formally as

$$HA \rightleftarrows A^- + H^+ . \tag{1}$$

The dissociation constant K_A for this reaction can be expressed as a quotient of the equilibrium concentrations of right and left side of the reaction equation:

$$K_A = \frac{[A^-][H^+]}{[HA]}. \tag{2}$$

With the definition $pK_A = -\log_{10}(K_B)$ and the pH being the logarithm of proton concentration $[H^+]$, the Henderson-Hasselbalch equation can be formulated:

$$pK_A = pH - \log\left(\frac{[A^-]}{[HA]}\right). \tag{3}$$

It is often more useful to express the equation in terms of protonation probabilities ρ_{prot}, leading to the expression

$$pK_A = pH - \log\left(\frac{1-\rho_{prot}}{\rho_{prot}}\right) \tag{4}$$

and, if solved for ρ_{prot}, to

$$\rho_{prot} = \frac{e^{-\ln(10)[\mathrm{pH-pK_A}]}}{1+e^{-\ln(10)[\mathrm{pH-pK_A}]}} \tag{5}$$

From this expression one can derive the Gibbs reaction free energy for the deprotonation of the acid

$$\Delta G(\mathrm{HA} \rightarrow \mathrm{A}) = -\ln(10) \cdot \mathrm{RT} \cdot [\mathrm{pH} - \mathrm{pK_A}] \tag{6}$$

with R being the universal gas constant and T the absolute temperature.

pK_A values in proteins

The above equations are valid for a titratable molecule solvated individually in water. If the molecule is part of a polypeptide chain (like an amino acid) or bound to a protein structure (e.g. a cofactor or a ligand) the situation becomes more complex and the reaction free energy can be written as

$$\Delta G(\mathrm{HA} \rightarrow \mathrm{A}) = -\ln(10) \cdot \mathrm{RT} \cdot [\mathrm{pH} - \mathrm{pK_A}] + \Delta G_p, \tag{7}$$

where ΔG_p is the change in protonation free energy due to the presence of the protein or other molecules, that partially or completely replace the water environment around the residue. This includes interactions with the non titratable residues in the protein as well as with other titratable residues. The protonation state of the other titratable residues may change with pH and as a consequence, even the whole protein structure may change. Therefore, ΔG_p is a function depending on the protonation states of all other titratable residues described by the protonation vector $\mathbf{p}$(pH). In a protein it is therefore insufficient to simply look at an individual titratable residue. Instead the free energy for $\Delta G_p(\mathbf{p}_{ref} \rightarrow \mathbf{p}_i, \mathrm{pH})$ for changing the protonation of all titratable residues at a given pH has to be considered. Here, $\mathbf{p}_{ref}$ is a reference and $\mathbf{p}_i$ any other protonation vector. $\mathbf{p}_{ref}$ is an arbitrary choice, which, in the past[13], was often chosen so that all titratable residues in the protein are in a state with neutral total charge.

The function $\Delta G_p(\mathbf{p}_{ref} \rightarrow \mathbf{p}_i, \mathrm{pH})$ offers a possibility to compare the energies of different protonation vectors and therefore to evaluate the probability of each protonation state at a given pH. Details on how ΔG_p can be obtained are discussed in chapter 2.2 "*Electrostatic energy calculations*". The probability $\rho(\mathbf{p}_i)$ of a protonation vector $\mathbf{p}_i$ at a given pH is

$$\rho(\mathbf{p}_i, \mathrm{pH}) = \frac{e^{-\frac{\Delta G_p(\mathbf{p}_i, \mathrm{pH})}{k_B T}}}{\sum_j^{N_p} e^{-\frac{\Delta G_p(\mathbf{p}_j, \mathrm{pH})}{k_B T}}}, \tag{8}$$

where k_B is the Boltzmann factor, T the absolute temperature and N_p the total number of possible protonation states. Even for a small protein involving for instance 50 titratable residues, the number of possible protonation state is too large to evaluate $\rho(\mathbf{p}_i)$ directly from

equation (8). For a protein with N titratable residues, of which each has just two states, N_p becomes 2^N. Instead, probabilities are evaluated approximately using a Metropolis-Monte-Carlo algorithm implemented in the software Karlsberg[14]. Repeating this procedure for each pH within the pH interval of interest can be regarded as a 'virtual titration'. From this titration one can easily obtain the pH dependent protonation probability $p_{prot}(n, pH)$ of a specific residue n. In general, p_{prot} can be an arbitrary complex function of pH, if there is a strong interaction network with other titratable residues. For most residues embedded in a protein, however, p_{prot} can still be approximated with a Henderson-Hasselbalch equation (see equation (4)). Therefore, it is still possible to define a pK_A ($pK_{A,protein}$) of the residue in the protein. The effect of the protein environment on the titration behavior of residues is then simply described by shifts of the $pK_{A,protein}$ values compared to the residues pK_A in solution.

2.2. **Electrostatic energy calculations**

Electrostatic energies describe the interactions in between a distribution of charges. The charges can be distributed continuously over a volume or they can be discrete point charges. Electrostatic energies are only a part of all the energies that have to be considered to completely understand the electrostatic properties of a molecule. What makes them especially interesting is their long range nature. In the simple case of two point charges (q_1 and q_2) in vacuum, the force acting between them can be expressed with Coulombs law

$$\vec{F} = \frac{1}{4\pi\varepsilon_0} \cdot \frac{q_1 q_2}{|\vec{r}_{1,2}|^2} \cdot \frac{\vec{r}_{1,2}}{|\vec{r}_{1,2}|}, \tag{9}$$

whereby ε_0 is the vacuum permittivity and $\vec{r}_{1,2}$ the vector connecting the position of the two charges. The electrostatic potential of a charge q_1 is

$$\phi(\vec{r}) = \frac{1}{4\pi\varepsilon_0} \cdot \frac{q_1}{|\vec{r}|}. \tag{10}$$

The electrostatic inter action energy of two charges at a distance $|\vec{r}_{1,2}|$ can be expressed as

$$E_{1,2} = q_2 \cdot \phi_1(\vec{r}_{1,2}) \tag{11}$$

The equation (11) shows that the energy decays comparably slow with being proportional to 1/r. The vdW interaction, in contrast that is usually modeled with a Lennard-Jones potential, decays with $1/r^6$. For atoms with a distance of just a few Ångströms between each other, the electrostatic energy becomes the dominating type of interaction. In the context of pK_A calculations the electrostatic energies are used to describe the interaction between different residues and partially to account for conformational changes within a residue. The latter point is discussed in more detail in chapter 4. All other energy differences associated with the protonation of a residue are described by its pK_A and equation (6). These energies concern intra-molecular effects, like the formation of new bonds and redistribution of the atomic partial

charges, as well as the interaction of the molecule with the proton concentration (the pH value) in the solvent.

In this work atomic charges are taken from the CHARMM22[15] molecular mechanic force field (see chapter 2.3). Each atom is represented as a point charge and in each protonation state the titratable residue has an individual set of charges. Electrostatic energies are obtained by solving the Poisson-Boltzmann equation, as described in the section 'The Poisson-Boltzmann Equation' (PBE) below. The framework of PBE offers a more general description than equation (10). It allows the use of an inhomogeneous dielectric medium with a spatially varying dielectric constant $\varepsilon(\mathbf{r})$ to account for effects that influence the charge-charge interactions, but are not explicitly included in the model. Inside the protein volume this is e.g. the change of the dipole moment of a residue as a reaction to the electric field. Outside of the protein a dielectric constant of 80 is commonly used to describe the influence of the water molecules that solvate the protein, without the need for any explicit charges. Additionally, the PBE can account for salt concentration in the solution around the molecule or protein. After discussing the PBE in the following chapter, detailed information on how it is used for predicting pK_A shifts is given in the section 'Protonation Energies'.

The Poisson-Boltzmann Equation

The electrostatic theory is based on a fundamental law, namely Gauß´ law. It is one of the four Maxwell equations of classical electromagnetism and relates net electric flux through a closed surface with net electric charge (q) enclosed within that surface

$$\oint \vec{E} \cdot d\vec{A} = \frac{q}{\varepsilon_0} \tag{12}$$

with ε_0 being the vacuum permittivity. This equation can be converted into a volume integral by applying the divergence theorem

$$\int \vec{\nabla} \vec{E} \cdot dV = \frac{q}{\varepsilon_0}. \tag{13}$$

By expressing the total charge in terms of a charge density (ρ) integrated over the volume it becomes

$$\int \vec{\nabla} \vec{E} \cdot dV = \int \frac{\rho}{\varepsilon_0} \cdot dV. \tag{14}$$

Differentiating both sides of the equation with respect to the volume results in the differential equation

$$\vec{\nabla} \vec{E} = \frac{\rho}{\varepsilon_0}. \tag{15}$$

Since the electrostatic field is conservative it can be expressed in terms of the negative gradient of a potential Φ to obtain the Poisson equation of electrostatics

$$\vec{\nabla}\vec{\nabla}\phi(\vec{r}) = -\frac{\rho(\vec{r})}{\varepsilon_0}. \tag{16}$$

The equation is only valid for vacuum. In other dielectric media the electrostatic field generated by the charge distribution can induce dipoles, reducing the effective electrostatic field that is actually transmitted through space. This can e.g. be caused by small solvent molecules that change their orientation. To account for this effect, the electrostatic displacement field $\vec{D} = \varepsilon \vec{E}$ is inserted in equation (14) and the Poisson equation becomes

$$\vec{\nabla}\,\varepsilon(\vec{r})\,\vec{\nabla}\phi(\vec{r}) = -\frac{\rho(\vec{r})}{\varepsilon_0}. \tag{17}$$

The scalar function $\varepsilon(\vec{r})$ is named dielectric constant. Its choice is crucial for all electrostatic energy calculations. It usually has two functions. The first one is to account for the effect of the solvent around the molecule, without the need for explicit charges. Water, that is assumed to fill the volume around a protein, can be described by a dielectric constant of $\varepsilon_w = 80$. The way the volume of a protein is defined is discussed in the next section 'The protein surface'. The second function of the dielectric constant is to account for polarizability effects in that part of the medium that does contain explicit charges. The correct choice of the dielectric constant inside the protein is an ongoing discussion. It has been argued[16] that a range from $\varepsilon_p = 2$ to 20 would be reasonable. For unpolar organic molecules the dielectric constant usually tales a value of 2 and accounts for the reaction of the electrons to the electrostatic field. This electronic polarizability is also present in proteins, but additionally many amino acids are polar. They carry a permanent dipole caused by the charges of their side chains, which can adjust their orientation according to the electrostatic field. Polarity depends on the type of amino acid and flexibility in changing its side chain orientation which depends on details of the protein structure around the molecule. Therefore, it has been suggested that the actual dielectric constant should be inhomogeneous throughout the protein[17], starting from 4 in the hydrophobic core to 20 at the more polar regions. In this work the lower boundary of this range ($\varepsilon_p = 4$) is used for the protein volume in all calculations.

The Poisson equation as it is given in equation (17) would be sufficient for electrostatic energy calculations if the protein would be solvated in pure water. Usually this is not the case as the surrounding water does contain a certain amount of salt. To account for these ions, the Poisson equation can be modified with the help of the Debye-Hückel theory. Thereby the charge distribution in equation (17) is extended to account for the mobile ions dissolved in water. The result is the so called Poisson-Boltzmann Equation (PBE), a non-linear differential equation:

$$\vec{\nabla}\,\varepsilon(\vec{r})\,\vec{\nabla}\phi(\vec{r}) = -4\pi\left[\rho(\vec{r}) + \kappa^2 \frac{kT}{e_c}\upsilon(\vec{r})\sinh\left(\frac{e_c\phi(\vec{r})}{k_B T}\right)\right] \tag{18}$$

with κ being the inverse Debye length

$$\kappa = \sqrt{\frac{8\pi N_A e_c^2 I_s}{k_B T}}, \tag{19}$$

e_c the elementary charge, N_A the Avogadro's number, T the absolute temperature and k_B the Boltzmann constant. The parameter

$$I_S = \frac{1}{2}\sum_i c_i \cdot z_i^2 \tag{20}$$

considers the ionic strength that depends on concentration c_i and charge z_i (in units of the elementary charge) the ion of type i in the solvent. The volume exclusion function $\nu(\vec{r})$ defines the volume accessible for the ions by having a value of zero in the protein volume and one everywhere else.

The PBE can be simplified to a linearized form by approximating the sinus hyperbolicus function with the linear term of its Tailor expansion. With $\sinh(x) = x + x^3/3! + x^5/5! + \ldots$ one obtains the linearized PBE (LPBE):

$$\vec{\nabla}\varepsilon(\vec{r})\vec{\nabla}\phi(\vec{r}) + \kappa^2 \upsilon(\vec{r})\phi(\vec{r}) = -4\pi\rho(\vec{r}) \tag{21}$$

This equation has the advantage that the resulting electrostatic potentials and energies are additive. That means that the total energy can be split in several partial interaction energies. In case of the simple example of three charges (A, B, C), the interactions energies could be calculated pairwise (A ↔B, A↔C, B↔C) by just considering the charges of the corresponding pair. The total electrostatic energy can then be obtained by summing up all three interaction energies. This is an important property required for the framework used to obtain pK_A values as described in section 'Protonation Energies'.

The protein surface

Definition of the protein volume is crucial for two parameters in the PBE, the dielectric constant ε and the volume exclusion function for ions. The protein volume is defined here as the volume confined by the so called solvent exclusion surface (SES). To obtain the SES, each protein atom is represented by a sphere with van-der-Waals (vdW) radius. The SES is then the area that can be reached by the surface of a probe sphere whose radius reflects the size of solvent molecules, without overlapping with the vdW spheres of any nearby atom as illustrated in Figure 1. The vdW radii are taken from the CHARMM22[15] force field. The volume exclusion function that defines the volume accessible for salt ions is defined in a similar way. The only difference is the use of a larger probe radius of 2 Å instead of 1.4 Å used for the dielectric constant. If not specified otherwise the SES algorithm implemented in the software APBS[18] is used here.

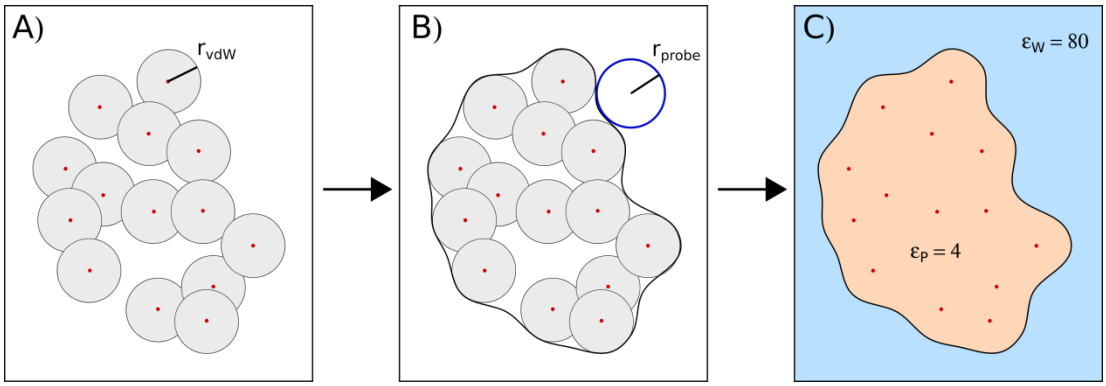


Figure 1: Illustration of the procedure to generate the solvent excluded surface of a molecule need for the spatial dependency of the dielectric constant $\varepsilon_{min}(\mathbf{r})$. The red dots mark the centre of the atoms where their charges are located.
A) The volume of each atom is represented by a vdW sphere.
B) The solvent exclusion surface (SES) is defined as the area reachable by a sphere with a radius of r_{probe} = 1.4 Å, that rolls over the protein surface.
C) The volume contained in the SES is defined as protein volume, everything else as solvent.

Definition of the protein volume is a crucial factor for the electrostatic energy calculations. In this work two variants of the procedure described above have been explored. In chapter 3 ('*Electrostatic pKA Computations in Proteins: the Role of Internal Cavities*') an alternative method has been developed and tested to account for the problem that small internal cavities are sometimes considered to be part of the protein volume, even though they are filled with one or more water molecules in the corresponding crystal structure. In chapter 4 an alternative set of vdW radii is used to define the size of the sphere drawn around each atom (length of r_{vdW} in Figure 1A) and the consequences for pK_A calculations are described.

Numerical Solution of the Linear Poisson Boltzmann Equation

The solution to the linearized Poisson-Boltzmann (LPBE) equation as defined in equation (21) cannot be obtained analytically. Instead, a numerical solution is obtained using the finite-difference (FD) method implemented in the software APBS[18]. The FD method uses simple cubic grids to represent the charge distribution $\rho(\mathbf{r})$, the spatial distribution of the dielectric constant $\varepsilon_{min}(\mathbf{r})$ (see Figure 1C) and the volume exclusion function $\nu(\vec{r})$. The LPBE is then solved on these grids. The accuracy of the result depends on the resolution of the grids, which should therefore be as fine as possible. The limit for the size of the grids is given by the memory available on the computer running the calculation. For average sized proteins a good trade-off between accuracy and memory requirement is offered by grid constants of about 0.3-0.25 Å as used in this work, usually requiring about 1 to 4 GB of memory.

The LPBE is a second order partial differential equation. Solving it poses a boundary value problem. Hence, for obtaining the solution on every grid point, the electrostatic potential at the boundaries (the edges of the grid) has to be specified. There are two approaches to address this problem. The first one is the usage of a grid large enough so that an electrostatic potential

vanishes at the boundary surface. This requires a very large grid. A grid of moderate size can be used with the second approach, the Debye-Hückel approximation. Hereby the protein is assumed to be a single point charge, whose value corresponds to the sum of all individual atomic charges. The boundary potential can then be approximated analytically. The demand for a large grid contradicts the requirement for an accurate solution of the LPBE, that is a grid being as fine resolved as possible in the volume of the protein. This problem is solved by using the so called focussing method. Thereby a low resolution grid that covers a large volume is used initially and the electrostatic potential obtained by solving the LPBE on this grid is then used as the boundary condition for a second calculation on a finer grid that covers just the volume occupied by the atomic charges. Optionally, one or more additional calculations on grids with intermediate sizes can be performed in between to ensure that the leap in resolution in between two calculations remain moderate.

The numerical solution (Φ_{calc}) of the LPBE obtained with the FD method has an important limitation. It is not possible to obtain the absolute electrostatic potential (Φ_{el}), instead there is always an additive error (Φ_{err}) resulting in

$$\phi_{calc}(\mathbf{r}) = \phi_{el}(\mathbf{r}) + \phi_{err}(\mathbf{r}) \quad (22)$$

The additive error is the so called grid artefact that occurs due to the use of point charges on discrete grids. As a consequence the calculated electrostatic potential $\Phi_{calc}(\mathbf{r})$ generated by the charges $q_i(\mathbf{r})$ can only be used in two specific ways. The first one is to evaluate interaction energies with charges that are not part of the charges [$q_i(\mathbf{r})$] that generate the electrostatic potential and whose position is not too close to any of these charges. The minimum distance between two not covalently bound atoms is about 1.8 Å, which is sufficient to fulfil this criterion. The second option is to evaluate energy differences. Examples are solvation energies, that is the energy required to bring the protein from vacuum into an aqueous environment. For this propose the electrostatic potential of two calculations are subtracted, one with and another without the solvent being present. These calculations differ only in the value of the dielectric constant used in the volume of the solvent (usually 80 and 4). The formalism used to obtain protonation energies as described in the next section was chosen so that these two options can be applied.

Protonation energies

To obtain the pK_A of a residue in a protein, the free energy difference for changing the protonation state of the residue has to be calculated as stated in equation (7). As discussed in section 2.1 one of the difficulties of pK_A calculations is that titratable residues can have strong mutual electrostatic interactions. That means that the free energy for the protonation of a residue may depend on the current protonation state of one or more of its neighboring residues. Therefore, the protonation energy can usually not be specified for an individual residue, but instead must be evaluated for all titratable residues simulatneously. This results in the free

energy difference $\Delta G_p(\mathbf{p}_{ref} \rightarrow \mathbf{p}_k, pH)$, with **p** being a vector (named protonation vector in the following) that specifies the protonation state of all titratable residues. Protonation states as they are defined here do not necessarily differ in the amount of protons being bound to the molecule. A histidine for example has three protonation states, one with the residue being protonated and two being deprotonated. $\mathbf{p}_{ref}$ is the reference protonation state, whose energy is defined to be zero. Its choice is arbitrary and has been chosen in the past[13] to be a state of neutral charge (e.g. the deprotonated state for Lys and the protonated state for Asp). Due to the large number of possible protonation vectors it is not possible to obtain ΔG_p for each vector with an individual electrostatic energy calculation. Instead ΔG_p is formulated as an energy function composed of elements that can be calculated individually.[19-21] The expression for this energy function $\Delta G_p(\mathbf{p}_{ref} \rightarrow \mathbf{p}_k, pH)$ will be derived hereinafter.

In a first step the protonation energy of each titratable residue is calculated individually. This individual protonation energy $\Delta G_{p,single}$ is the energy to change the protonation state of a specific titratable residue while all other titratable residues remain in their reference protonation state. To obtain $\Delta G_{p,single}$ the thermodynamic cycle shown in Figure 2 is used. In short the procedure is the following. The specific residue being in its reference protonation state is placed from protein environment (step A in Figure 2) into aqueous solution (step B). In the aqueous solution the protonation state is changed (step B to C). Afterwards it is placed back into protein environment (step D). The energy to switch the protonation state of a titratable residue in aqueous solution can be obtained from the experimental pK_A with equation (6). The energy ΔG to bring the residue from water into the protein is named desolvation energy and is obtained by electrostatic energy calculations.

Two calculations are required for obtaining ΔG, one with protein environment (*prot*) and another with aqueous environment (*solv*), generating the two electrostatic potentials $\Phi_{solv}(q_m)$ and $\Phi_{prot}(q_m)$. In both calculations the same grid geometry and the same set of charges q_m is used. In Karlsberg$^+$ these are those charges of the residue that do change their value in between different protonation states. For the new method, as introduced in chapter 4 (*KB2$^+$MD*), this set contains all charges of the residue. The two calculations differ in the spatial dependency of the dielectric constant $\varepsilon(\mathbf{r})$. For Φ_{prot} the whole protein volume is filled with a low dielectric constant of $\varepsilon_p = 4$ (beige area in Figure 2) while for Φ_{solv} it is only the volume of the residue itself. Both volumes are defined with the SES procedure discussed above. Water is considered to have a dielectric constant of $\varepsilon_w = 80$ (blue area in Figure 2).

The desolvation energy for residue i in protonation state s is now split into two parts:

$$\Delta G_{i,s} = \Delta G_{Born}^{i,s} + \Delta G_{back}^{i,s} \tag{23}$$

The first term ΔG_{Born} is the so called Born energy, the self-energy of the N_q charges q_m that have been used to generate Φ_{solv} and Φ_{prot} and is given by

$$\Delta G_{Born} = \frac{1}{2}\sum_{m}^{N_q} q_m \left[\phi_{prot}(\mathbf{r}_m) - \phi_{solv}(\mathbf{r}_m) \right]. \tag{24}$$

The coordinates of the charges q_m are specified by the vectors $\mathbf{r}_m$. Since the same charges are used to obtain the energy and to generate the electrostatic potentials, the factor ½ is required to avoid double counting. The second term in equation (23) (ΔG_{back}) is the so called background energy, the interaction energy between the charges q_m of a specific titratable residue and the remaining charges in the system. The background charges differ for the two calculations. For Φ_{solv} they contain those charges ($q_{solv,n}$) of the residue that are not already contained in q_m. For Φ_{prot} the background charges contain all charges ($q_{prot,n}$) in the protein that are not already contained in q_m, whereby all other titratable residue are set to their reference protonation state. In Karlsberg+ the definition of the group of atoms that belong to a residue is slightly extended for the solvent calculation Φ_{solv} and does also include the backbone C=O and N-H group of the neighboring residues.

With N_x being the number and $\mathbf{r}_x$ the coordinates of the background atoms in the environment (solv or prot), the background energy can be evaluated as follows

$$\Delta G_{back} = \sum_{n}^{N_{prot}} q_{prot,n} \cdot \phi_{prot}(\mathbf{r}_n) - \sum_{n}^{N_{solv}} q_{solv,n} \cdot \phi_{solv}(\mathbf{r}_n) \,. \tag{25}$$

For both energies (ΔG_{Born} and ΔG_{back}) the grid artifact contained in the electrostatic potentials vanishes. For ΔG_{Born} it cancels in equation (24), because $\Phi_{solv}(q_m)$ and $\Phi_{prot}(q_m)$ contain exactly the same grid artifact. For ΔG_{solv} the grid artifact does not occur, since both potentials are evaluated at the coordinates of $q_{prot,n}$ and $q_{solv,n}$ that do not overlap with the coordinates of the charges q_m used to generate the electrostatic potentials.

Finally the relative desolvation energy for a titratable residue i is defined as

$$\Delta\Delta G_{desolv,i}(s_i) = \Delta G(s_i) - \Delta G(s_{ref,i}), \tag{26}$$

comparing the desolvation energies of the residue being in protonation state s_i or in its reference protonation state $s_{ref,i}$.

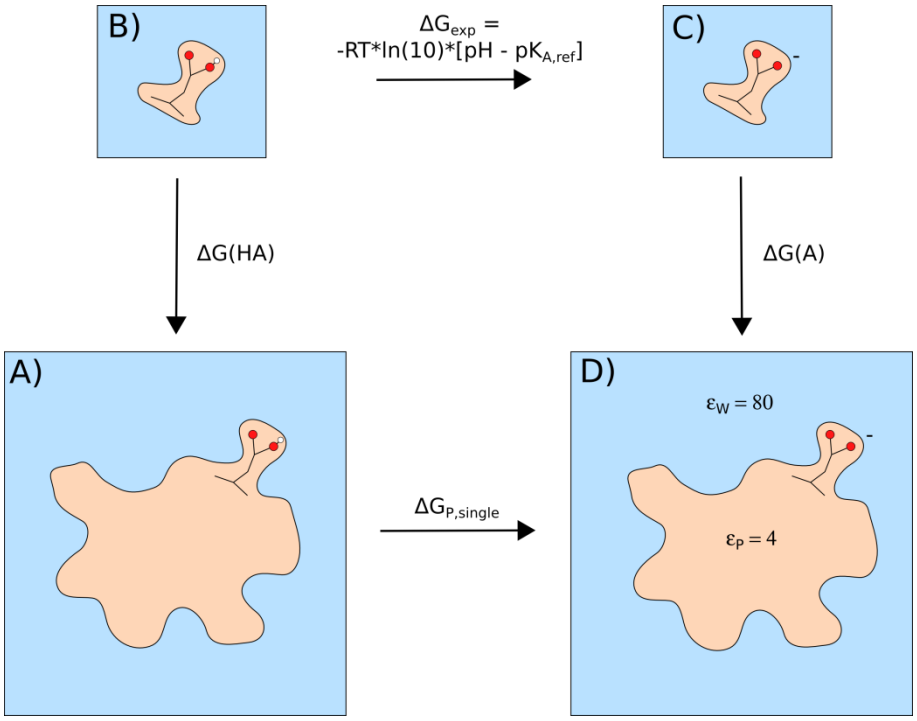


Figure 2: Illustration of the thermodynamic cycle used to calculate the protonation energy $\Delta G_{p,single}$ for the example of the deprotonation of an aspartate residue. The beige area represents the protein and is filled with a dielectric medium of $\varepsilon_p = 4$, while the area accessible to the solvent is colored blue and filled a dielectric medium of $\varepsilon_w = 80$. In frame A) and B) the Asp residue is deprotonated, in C) and D) it is deprotonated. The energies ΔG(HA) and ΔG(A) are obtained by electrostatic energy calculations, while ΔG_{exp} is the experimental free energy for the protonation of an Asp residue in aqueous solution. Details are discussed in the text.

The next step is to add the missing pairwise interaction energies between all titratable residues. For each pair of residues (described by the indices i and j) this energy $W_{i,j}(s_i, s_j)$ depends on the protonation states s_i and s_j of both residues. The desolvation energy term does already include the pairwise interaction energies between residue i in state s_i and all other titratable residues being in their corresponding reference protonation states. What is missing is the change in the pairwise electrostatic interaction energy if both residues are in one of their non-reference states.

The interactions $W_{i,j}(s_i, s_j)$ can be calculated without performing any further electrostatic energy calculation. Instead the electrostatic potential Φ_{prot} can be reused

$$W_{i,j}(s_i,s_j) = \sum_{m}^{N_i}\left[q_{m,i}(s_i) - q_{m,i}\left(s_i^{ref}\right)\right]\cdot\left[\phi_{prot,j}(s_j,\mathbf{r}_{m,i}) - \phi_{prot,j}\left(s_j^{ref},\mathbf{r}_{m,i}\right)\right]. \qquad (27)$$

Here $q_{m,i}(s_i)$ is one of the N_i charges of residue i in state s_i at position $\mathbf{r}_{m,i}$. $\Phi_{prot,j}(s_j, \mathbf{r}_{m,i})$ is the electrostatic potential generated by residue j in state s_j in the protein environment. The charges $q_{m,i}$ are those used to generate the electrostatic potentials as discussed above. The symbol s_{res}^{ref} refers to the reference state of residue *res*. The diagonal elements of $W_{i,j}(s_i, s_j)$ are zero since the self energy is already contained in the desolvation term, eq. (23). Furthermore all entries in eq. (27) with $s_i = s_i^{ref}$ or $s_j = s_j^{ref}$ vanish.

The complete protonation energy ΔG_p for transition from the protonation patter $\mathbf{p}_{ref}$ to $\mathbf{p}_k$ can now be obtained by inserting the $\Delta\Delta G_{desolv}$ and W into equation (7) and extending it to a sum over all N titratable residues in the protein:

$$\begin{aligned}\Delta G_p(\mathbf{p}_{ref} \rightarrow \mathbf{p}_k, pH) = &\sum_{i=0}^{N}\left(x_i^k - x_i^{ref}\right)\cdot RT\ln(10)\cdot\left(pH - pK_{A,ref}^i\right) \\ &+ \sum_{i=0}^{N}\Delta\Delta G_{desolv,i}(s_i) \\ &+ \sum_{i}^{N}\sum_{j,i\neq j}^{N} W_{i,j}(s_i, s_j)\end{aligned} \tag{28}$$

As discussed above the protonation vector $\mathbf{p}_k$ determines the protonation state $s_i = \mathbf{p}_k(i)$ of the titratable residue i. The expression $\left(x_i^k - x_i^{ref}\right) \in \{-1,0,1\}$ is introduced to ensure that the experimental free energy for protonation is only used if the protonation state is actually changed and to give the correct sign for the experimental free energy. The first symbol x_i^k is defined to be 0 or 1 if the state s_i of residue i refers to the deprotonated or protonated state. The second symbol x_i^{ref} is 0 or 1 if the residue i is an acid or a base. The protonation energy ΔG_p can now be inserted into equation (8) to perform a complete titration of the protein in order to obtain the final pK_A values.

For the sake of clarity equation (28) does not cover the situation that a titratable residue has a protonation state with the same protonation but not the same experimental pK_A as the reference state. This is for example the case for a histidine that has two deprotonated states with pK_A values of 6.6 and 7.0 of which one is defined as its reference state. For the transition of a residue into a state that is a tautomer of the reference state, the first term in equation (28) vanishes. If there is an experimentally measured energy available for this transition (e.g. 0.4 pH units for histidine), it simply has to be added to the protonation energy ΔG_p.

The protonation energy defined by equation (28) is sufficient if only a single structure is available. Since a change in protonation of individual residues may have structural consequences, it may be necessary to include the possibility of a conformational change to the titration. This can be achieved by providing a set of protein structures, each representing the protein at a certain pH interval. These can be crystal structures or modeled structures like e.g. the pH adapted conformations of Karlsberg+ discussed in chapter 2.4. For each structure c in the set, the desolvation $\Delta\Delta G_{desolv,i}^c$ and interaction $W_{i,j}^c$ energies have to be calculated individually. The results are then combined for the final titration by extending equation (28) to:

$$\begin{aligned}\Delta G_p(\mathbf{p}_{ref} \rightarrow \mathbf{p}_k, c, pH) = &\sum_{i=0}^{N}\left(x_i^k - x_i^{ref}\right)\cdot RT\ln(10)\cdot\left(pH - pK_{A,ref}^i\right)\\ &+\sum_{i=0}^{N}\Delta\Delta G_{desolv,i}^c(s_i)\\ &+\sum_{i}^{N}\sum_{j,i\neq j}^{N}W_{i,j}^c(s_i,s_j)\\ &+\Delta G_{conf}^c\end{aligned} \quad (29)$$

The additional term ΔG_{conf}^c is the conformational energy of structure c with all titratable residues being set to their corresponding reference protonation states. A procedure to obtain conformational energies is discussed in the next section *"Conformational Energies"*. The conformation c becomes another parameter of the function ΔG_p in addition to the protonation vector **p** and the pH.

Historically[22, 21, 13] equations (28) and (29) are formulated using an expression called *intrinsic pK$_A$* $\left(pK_A^{int}\right)$. The *intrinsic pK$_A$* is defined as the sum of experimental pK_A and desolvation energy:

$$pK_{A,i}^{int} = pK_{A,ref}^i + \frac{\Delta G_{solv,i}(s_i)}{RT\ln(10)} \quad (30)$$

The procedure described so far including the definition of the reference state follows the traditional approach[21] as it is implemented in the Software Karlsberg+[13]. The formalism has one huge disadvantage, which is that the calculated terms for the relative desolvation energy and the interaction matrix are not directly interpretable. The desolvation energy of a titratable residue contains the pairwise interaction energies with all other residues being in their reference states, while in the interaction matrix these terms are lacking. As a consequence it is not possible to obtain the detailed interaction network, e.g. to answer the question, why the calculated pK_A of a specific residue shifts strongly. With the procedure named *KB2*MD* that is introduced in chapter 4, the reference state is chosen in a different way that makes all terms directly interpretable. Details of the procedure are discussed in that chapter.

Conformational energies

In the previous section it has been discussed how the LPBE can be used to obtain protonation energies. Another common application for the LPBE is the calculation of conformational energies used to compare two or more conformations of a molecule or protein. An important assumption is here, that the electrostatic energy difference is the most important factor for the comparison of the structures. For large molecules like proteins, this only holds if the conformational changes are not too large. Otherwise entropic effects not covered by the Poisson-Boltzmann framework may become important. With small changes, like the rearrangement of side chains, it can be assumed that the entropies of the involved conformations are nearly the same such that the influence from entropy can be neglected.

The conformational energy cycle used to calculate the electrostatic free energy difference ΔG_{conf} between two conformations is illustrated in Figure 3. Due to the presence of the grid artifact in electrostatic potentials obtained by numerically solving the LPBE, ΔG_{conf} cannot be calculated directly. Instead the protein being in one conformation (A and B in Figure 3) is first desolvated into an homogeneous dielectric environment ($\Delta G_{PB,1}$). In the homogeneous dielectric environment (B and C, corresponds to vacuum) the protein conformation is changed ($\Delta G_{Coulomb}$). Finally the protein in its new conformation is solvated again in water (C and D, $\Delta G_{PB,2}$). The electrostatic free energy for the transition from conformation one to conformation two is then given by

$$\Delta G_{conf}\ (1 \rightarrow 2) = \Delta G_{Coulomb} + (\Delta G_{PB,2} - \Delta G_{PB,1})\ . \tag{31}$$

Often there are more than two structures that are to be compared. It is therefore useful to define an absolute electrostatic energy $G_{conf}(c)$ for each conformation c, so that $\Delta G_{conf}(n \rightarrow m) = G_{conf}(m) - G_{conf}(n)$. With $\Delta G_{Coulomb} = G_{Coulomb,m} - G_{Coulomb,n}$ this energy can be expressed as

$$G_{conf}(c) = \Delta G_{PB,c} + G_{Coulomb,c} \tag{32}$$

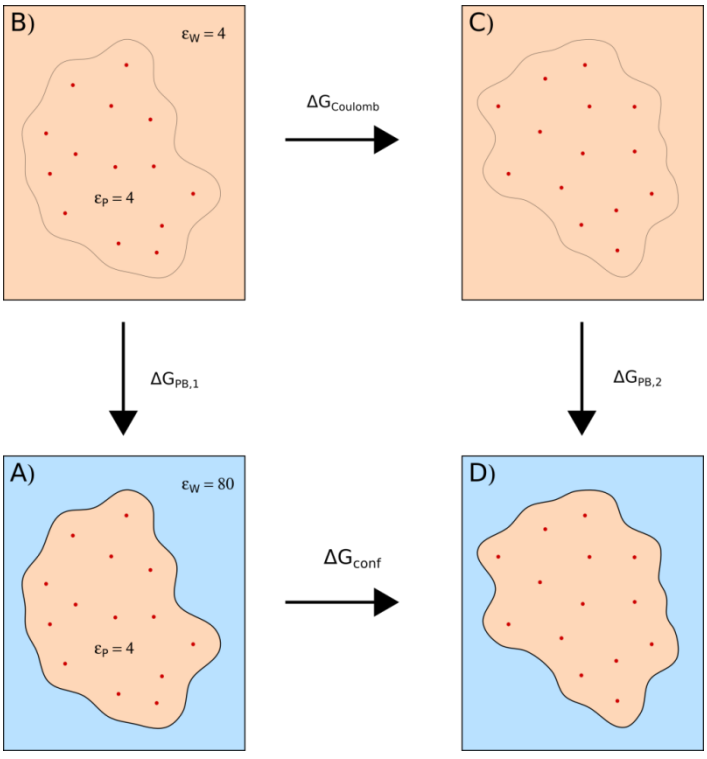


Figure 3: The thermodynamic cycle used to calculate the electrostatic conformational energy difference (ΔG_{conf}) for a protein in two different conformations. The first step (A to B) is the transition of the protein from an inhomogeneous into a homogeneous dielectric environment (desolvation). In the second step (B to C) the protein conformation is changed. The third step (C to D) is the transition of the protein in its new conformation back into the inhomogeneous environment. The free energy differences $\Delta G_{PB,1}$ and $\Delta G_{PB,2}$ are obtained with electrostatic energy calculations, while for $\Delta G_{Coulomb}$ the energy difference can be calculated analytically.

The free electrostatic energy $\Delta G_{PB,c}$ is calculated by solving the LPBE for the protein in conformation c for both the inhomogeneous ($\Phi_{prot,c}$) and the homogeneous environment ($\Phi_{vac,c}$), e.g. for frame A and B in Figure 3, and is then obtained by

$$\Delta G_{PB,c} = \frac{1}{2}\sum_{i}^{N} q_i \cdot \left[\phi_{prot,c}(\mathbf{r}_{c,i}) - \phi_{vac,c}(\mathbf{r}_{c,i})\right]. \tag{33}$$

Here q_i is the charge of one of the N atoms of the protein being located at the coordinates $\mathbf{r}_{c,i}$. The factor ½ is required to avoid double counting, since the charges q_i have also been used to generate the electrostatic potentials Φ. The grid artifact cancels, since $\Phi_{prot,c}$ and $\Phi_{vac,c}$ contain exactly the same additive errors if the same grid geometry is used.

The term $G_{Coulomb,c}$ is the self interaction energy of all charges in a homogeneous dielectric environment and can be easily obtained analytically with equation (10). The way the self-interaction energy is calculated here differs from how it is calculated in the CHARMM[15] force field. In CHARMM all electrostatic interactions between atoms that are connected through less than three covalent bonds are neglected, since their interaction energy is completely defined by bonded parameters (see section *The CHARMM force field*). In the past[13] these interactions have, nevertheless, been included in Karlsberg+ calculations. For the procedure named KB2+MD that is introduced in chapter 4 it was found that excluding these interactions significantly improved the results.

2.3. **Molecular dynamic simulations**

The basis of many theoretical studies of proteins and other bio-molecules are structures that have been experimentally resolved via x-ray crystallography or NMR (Nuclear Magnetic Resonance) spectroscopy and deposited in the Protein Data Bank[23, 24] (PDB). These structures can give a detailed insight into the properties and function of proteins. They can also directly be used as the starting point for pK_A computations. Nevertheless, these structures also have their limitations. To represent a protein just by a single structure is often insufficient. A protein may be mostly rigid, but can still have some very flexible regions like a couple of residues forming a short loop. NMR structures are advantageous in this respect since they provide an ensemble of structures, but they have the problem of a significant higher uncertainty in atomic coordinates compared to x-ray structures. Furthermore, every structure has been resolved under certain conditions e.g. at pH 7 or with a cofactor bound, that may not necessarily be the condition that is of interest. The focus in this work is set on conformational changes of a protein structure due to the change of the protonation state of titratable residues in the protein.

A common and well established technique to gain insight into the conformational variability of a protein as well as to study its dynamical behavior is a molecular dynamic (MD) simulation. For MD simulations a molecule is described by a set of properties, defined in the so called molecular mechanic (MM) force field. This description is similar to the one used for the electrostatic energy

calculations, but is more detailed, since it also accounts for forces emerging from van-der-Waals interactions and from covalent bond geometries. The force field that is used in this work is CHARMM22[15] and is discussed in the next section '*The CHARMM force field*'. With this more detailed description the dynamic behavior of a molecule can be simulated. In short, the simulation technique can be summarized as follows: The sum of all forces give rise to individual force vectors acting on each atom. Following Newton's law of motion these forces alter the atoms velocity vectors that finally lead to a displacement of the atoms. This procedure is then repeated in discrete set of small time steps.

An important aspect of the used CHARMM22[15] MM force field is, that the topology of the protein cannot change by itself in the course of the simulation. That means that all covalent bonds remain as initially defined. As a consequence it is e.g. not possible to simply define the pH of the solvent. Instead a choice has to be made for the protonation state of each titratable residue. This is a consequence of the force field's simplified description of the protein. The advantage of this simplified description is the opportunity to study dynamics on comparably long time scales of nano- to micro-seconds. The MD simulations in this work have been performed on GPUs (graphical compute units) using the software NAMD[25], with the fastest GPUs being 'NVIDIA GTX Titian' cards. With this setup is was possible to simulate a small protein in a water box (about 20 thousand atoms in total) with a speed of about 40 ns/day.

An MM force field offers an excellent compromise between computational cost and accuracy for the problem addressed in this work. As discussed in chapter 4, it has been used to model the conformational consequences of proteins to changes of the protonation state of their titratable residues.

The CHARMM force field

The CHARMM force field[26, 15] can be separated into two major categories: bonded and non-bonded forces. In general each atom is represented by a point. All forces are acting in between these points. The non-bonded forces are the charge-charge interactions and van-der-Waals interactions. For the charge-charge interactions a point charge is assigned to each atom, located in its center. The van-der-Waals interactions are modeled by a Lenard-Jones potential.

The bonded interactions account for covalent bonds in between atoms and consist of four terms. The first term accounts for the bonds in between pairs of atoms (F_{bond}), the second one for angles defined by groups of three atoms (F_{angle}), the third and fourth term describe forces acting on dihedrals (F_{dihe}) and improper dihedrals (F_{impr}) for groups of four atoms. The total force (F_{total}) acting on an atom n is given by the sum of all these forces:

$$\begin{aligned} F_{total,n} &= F_{bonded} + F_{non-bonded} \\ &= F_{elec} + F_{vdW} + F_{bond} + F_{angle} + F_{dihe} + F_{impr} \end{aligned} \tag{34}$$

The fastest movements in this system are the oscillations of hydrogen atoms bound to heavy atoms (e.g. oxygen or nitrogen). A hydrogen atom is 16 times lighter than an oxygen atom. To resolve the hydrogen movements properly Newton's law of motion has to be evaluated in steps of 1 fs for the forces from equation (34). Their amplitudes of the bond fluctuations are comparably low due to the strong bond in between a hydrogen and its heavy atoms. Therefore these oscillations are usually not of interest and are neglected by modeling this covalent bond as being completely stiff. This allows the usage of a larger time step of 2 fs for the simulation, reducing the computational cost of the simulation by a factor of two.

Simulation conditions

The environment that a protein is exposed to in a living organism can be very complex. It is a mixture of water, ions, ligands, lipids and other proteins in various compositions. Due to the limitations in the available computation resources, the protein has to be simulated in a very simplified environment. Besides water this usually includes ions, lipids forming a membrane as well as ligands and other proteins as far as interactions with the studied protein are of interest. The proteins that have been studied in this work were all water solvable which obviates the need for a lipid membrane. In preliminary tests ions had been used to ensure a certain ion concentration, but the impact on the results turned out to be negligible. Therefore the environment of the protein is here composed solely of water molecules.

Each protein has been solvated in a cubic box of water molecules (see Figure 4). To avoid having a solid wall at the borders of the water box, periodic boundary conditions have been applied. If an atom leaves the box on one side, it enters the box on the opposing side. If the box size is chosen large enough the resulting environment is approximately that of a water bath that extends infinitely in all directions. Water molecules are represented by the widely used TIP3 model[27]. In this water model the bond angle is fixed. Together with the above mentioned approximation of fixed bond length between hydrogen and heavy atoms, the only degrees of freedom of a TIP3 water are translation and rotation of the whole molecule. This simplified model is additionally increasing simulation speed since most of the simulated atoms do actually belong to water (e.g. 92% for the simulation of Ribonuclease H shown in Figure 4).

All MD simulations have been performed under isothermal-isobaric conditions (NPT ensemble) using the software NAMD[25]. Temperature is kept constant by adjusting the velocities of the atoms using langevin-dynamics. Pressure is controlled by adjusting the size of the water box.

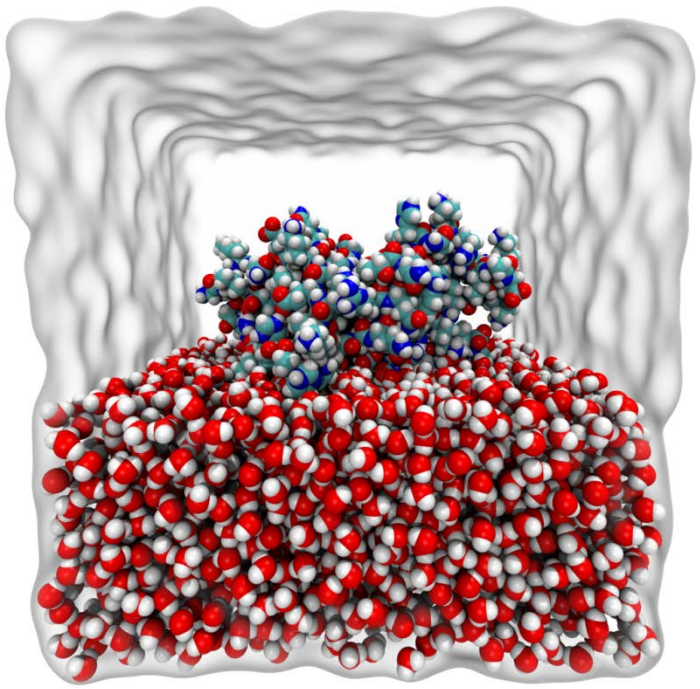


Figure 4: Illustration of the protein Ribonuclease H[28] solvated in water, right before the start of the MD simulation. The water molecules in the upper half of the box have been removed to make the protein visible. The radii of the atoms are changed for illustrative purpose. Especially polar hydrogen are much smaller (~0.2 Å) in the CHARMM22 force field. The gray area marks the surface of the water molecules. The actual box edges, as defined by periodical boundary conditions, are completely flat. The image has been rendered using the software VMD[29].

2.4. pK_A calculation in proteins with pH adapted conformations: Karlsberg+

The software Karlsberg+ has been developed by Gernot Kieseritky[13, 30] to provide an easy to use tool to compute pK_A values as well as redox potentials on the basis of protein crystal structures. In this chapter firstly the motivation for the development of Karlsberg+ will be discussed, followed by a brief introduction into the philosophy it uses to address the challenges of pK_A computation in proteins.

The computation of pK_A values generally starts with a crystal structure, usually downloaded from the Protein Data Bank[23, 24] (PDB). The most accurate structures deposited there have been resolved with the method of x-ray crystallography. This method provides the coordinates of heavy atoms e.g. oxygen and carbon atoms. But, in most crystal structures hydrogen atoms are partially or completely absent. These missing atoms have to be modeled prior to any calculation, which is usually done with the software CHARMM[26]. A protein structure prepared in this way can then be used to calculate the energies required for titrating the residues in a protein as described in chapter 2.2.

The protocol described above is the most simple way to obtain pK_A values with electrostatic energy calculations. But, as it has been shown in the past (sc_{ph7} procedure in Kieseritzky et al.[13]), the accuracy of the computed pK_A values can be poor. The Karlsberg+ protocol uses a single structure to describe the protein at the whole pH range of interest. As it has already been discussed in the introduction, for most proteins this assumption does not hold. The analysis of the computed pK_A values for a large benchmark set containing 185 experimentally measured pK_A values suggested that the most common structural changes occurring as a result of the change in pH can be summarized in two classes.[13] The first class is the rearrangement of the hydrogen bond network as a result of the protonation or deprotonation of individual residues. Hydrogens are very mobile depending on the constrains given by their chemical bond and can quickly adjust to any change in their environment. The second class of structural changes is the opening of salt bridges. Salt bridges are strong hydrogen bonds between two residues of opposing charge. This occurs e.g. with the protonation of acids (Asp and Glu) at a pH below 4 or the deprotonation of bases (Lys) above pH 10. For almost all proteins in the PDB, the available crystal structures are resolved at a pH close to 7. That means that for all other pH values the software for computing pK_A values has to take care of the correct modeling of conformational changes that may occur.

The concept of pH adapted conformations

Karlsberg+ addresses the challenge of structural variation with the concept of *pH adapted conformations* (PACs). A PAC is a protein structure that has been modeled automatically to represent the protein at a certain pH. The pH itself is not an external parameter that can simply

be set, instead it is represented indirectly by its effect on the protein, i.e. the protonation pattern applied to its titratable residues. As a consequence the protonation pattern for a certain pH has to be known in order to model the corresponding protein conformation. The correct structure itself, however, is needed for an accurate computation of the protonation pattern. This contradiction is solved by a self-consistent iterative procedure, illustrated in Figure 5. In an alternating sequence the protonation is computed as discussed in chapter 2.2 (step A) and, using the resulting protonation pattern, the protein structure is refined through modeling (step B). The development of an automatic modeling protocol that is applicable to a broad range of protein structures is a challenging task. Karlsberg+ offers the choice of two protocols, each addressing one of the two classes of common structural changes discussed in the introduction. The implementation of both protocols make use of the software CHARMM[26].

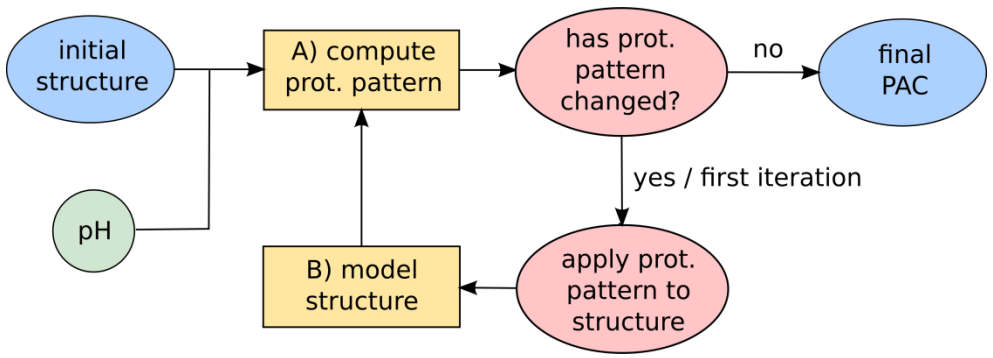


Figure 5: Schematic representation of the concept of the self-consistent cycle used to create an individual pH adapted conformation (PAC). An initial structure (e.g. the crystal structure) and the choice of a pH (usually pH -10, 7 or 20) are the input for the cycle. Step A, the computation of the protonation pattern, follows the scheme discussed in chapter 2.2. The modeling is done by the software CHARMM[26] following the chosen protocol.

The first modeling protocol (named H_{min} here) is simply an energy minimization of all hydrogen atoms of the protein. It is useful for PACs created in the range of pH 4 to 10. The second protocol (SB_{open}) perturbs all salt bridges of the protein, except for those that are completely buried inside the protein. It consists of three steps. In the first step, the side chain dihedral angle of any titratable residue that participates in a salt bridge is randomized. Since there is a high chance to produce unphysical conformations, 30 independent structures are created in this way. Subsequently, all 30 structures are energy minimized, where all atoms, except for those that are either part of a salt bridge residue or a hydrogen, are kept fixed. Finally, the structure with the lowest total energy (calculated in CHARMM, including all bonded and non-bonded force field terms) is selected. The SB_{open} protocol is only useful for a pH value were either all acids (or all bases) that are part of salt bridge are protonated (or deprotonated). Therefore, the SB_{open} PACs are generated at a pH of -10 and 20. All energy minimizations are performed in a homogeneous dielectric medium of $\varepsilon = 1$. A more detailed description of the protocol can be found in the publication of Karlsberg+[13].

The standard procedure of Karlsberg+ proposes the generation of 11 PACs for the pK_A calculation for a protein. One PAC at pH 7 with the H_{min} modeling protocol and five PACs for both, pH -10 and 20 using the SB_{open} modeling protocol. The five SB_{open} PACs created for the same pH differ in their random seed used for the side chain randomization. All 11 PACs are then combined according to equation (29). The final pK_A values are obtained by a titration of the protein i.e. by evaluating the Boltzmann sum (equation (8)) for each pH in the range of -10 to 20 in steps of 0.5 pH units.

The software Karlsberg+

The source code of Karlsberg+ is not publicly available, but pK_A calculations can be performed via a web interface on:

http://agknapp.chemie.fu-berlin.de/karlsberg

The software Karlsberg+ itself is written in Pearl. It deploys several external programs for the computational demanding tasks:

- Protonation energies are computed with the program TAPBS. It is a modified version of APBS, specifically designed to perform this task efficiently. This software is freely available on the website of Karlsberg+.
- The titration, i.e. the evaluation of the Boltzmann sum (equation (8)) for a range of pH values, is performed with a Metropolis Monte Carlo algorithm by the software Karlsberg[14]. This software is freely available on the website of Karlsberg+.
- Conformational energies are calculated with the software APBS[18].
- Protein modeling is in done in CHARMM[26] using the CHARMM22 force field[15].

3. Electrostatic pK_A Computations in Proteins: the Role of Internal Cavities

The project presented in this chapter has been published in the time of the dissertation:

Meyer, T.; Kieseritzky, G.; Knapp, E. W., Electrostatic pK_A Computations in Proteins: Role of Internal Cavities. *Proteins: Struct., Funct., Bioinf.* **2011,** 79, 3320-3332
DOI: 10.1002/prot.23092

All calculations and their analysis have been performed by myself and I also wrote the implementation of the algorithm described in the paper. Gernot Kieseritzky provided valuable advices for the implementation and background information concerning the software TAPBS and Karlsberg+. Prof. Ernst Walter Knapp was the supervisor of the project. The following text is a revised presentation of the project; no results have been added or changed. The pK_A values calculated for the non-mutated residues in the SNase Δ+PHS variant are not discussed here. These residues are instead studied in chapter 4. All figures except for Figure 7, and part of the text have been taken from the publication. © Proteins 2008; 71:1335–1348. Wiley-Liss, Inc.

3.1. Introduction

In 2009 a new benchmark set for pK_A prediction became available (www.pkacoop.org). The benchmark set consisted of a large number of titratable residues in staphylococcal nuclease (SNase) proteins, a Ca^{2+} dependent extracellular enzyme of 149 residues with a thymidine binding site.[31-38] Most of the measured pK_A values for the benchmark set were initially hold back to allow a bind prediction. The scientific community was invited to submit their predictions in advance to a workshop on "Protein Electrostatics" in Telluride in summer 2009. The software Karlsberg+ also participated in the competition. Surprisingly the predicted pK_A values showed very poor agreement with the experiments, yielding an $RMSD_{pKa}$ of 8.7 pH units. All prediction programs participating struggled with the new benchmark set, revealing their weaknesses and the need for further improvements.[39]

One of the specialties of the SNase protein is that the crystal structures of many variants contain water molecules in deep crevasses and internal cavities. Some of them are in close proximity to the titratable residues whose pK_A values should be predicted. Karlsberg+ employs the standard SES-algorithm described in the section *"The protein surface"* in chapter 2.2 to define the protein surface. It turned out that the SES-algorithm failed to detect many of the internal cavities that are filled by waters. In this project I developed a new algorithm that locates cavities more reliably. I further tested the influence of water-filled cavities on pK_A calculations with Karlsberg+. To achieve this, the spatial dependency of the dielectric constant in the electrostatic energy calculations was corrected using the information on cavity volumes found by the new algorithm.

3.2. **Materials and Methods**

The SNase benchmark set and its challenges

The SNase benchmark set was created by the García-Moreno group by replacing a single uncharged residue in the mostly hydrophobic interior of one SNase protein with either the amino acid Asp, Glu, Lys or Arg using site-directed mutagenesis. Subsequently, the pK_A of the introduced residue was measured. This yielded experimental data of highly shifted pK_A values. In total they introduced about 100 charged residues at 25 internal locations.[37, 40-42] For many of the variants they also resolved the corresponding crystal structures. The availability of a crystal structure was the central criteria for the choice of variants used to benchmark the method presented here. The 11 locations of the 20 variants considered in this project[32-36, 38, 40, 41, 43] are shown in Figure 6. The introduction of a residue that is charged under physiological conditions in the hydrophobic interior of a protein is usually energetically unfavorable. To avoid destabilization all mutation were inserted into two particular stable SNase variants, namely PHS and Δ+PHS.[44, 35] For some SNase variants experimental studies found hints for local rearrangements or even global unfolding.[34, 40, 45] Nevertheless for most variants the secondary structure seemed to remain intact, even on ionization of the mutated site.

Even though the overall fold of the protein remains the same, it is very likely that the local environment of the protein reacts to the protonation or deprotonation of the introduced residues in some way. This could explain the very poor agreement of the pK_A values predicted with Karlsberg$^+$ and other software that do not take these structural changes into account. Two possibilities for the structural reaction seem to be the most obvious explanations. The first one is a local rearrangement of side chains or flexible loop regions. The second one, that water enters the protein interior and thereby partially solvates the charged residue. The first possibility reveals one weakness of Karlsberg$^+$, its dependency of a representative crystal structure. To a certain degree the program can model structural changes that accompany a change in protonation, but as discussed in chapter 2.4 its modeling capability is limited to the relaxation of the hydrogen bond network and the perturbation of salt bridges. This limitation of Karlsberg$^+$ was the motivation for the second project of this thesis discussed in section 4. This project addresses the latter possibility that water enters the interior of the protein.

Besides its high tolerance against acidification, there is another specialty of the SNase protein. That is the existence of a large internal cavity, which is empty in all wild type crystal structures. No protein in the benchmark set used to optimize Karlsberg$^+$ had a cavity of comparable size.[13] The first question that arises if a crystal structure with an internal cavity is analyzed is to what degree the cavity is actually filled with water molecules. In crystal structures of sufficient high resolution oxygen atom coordinates of ordered water can be determined. These crystal waters are usually found at the surface of the protein, but sometimes they are located in crevasses or even in closed cavities. While the presence of crystal waters in a structure is a strong indication

that this water is also present in the native structure, the opposite is not necessarily true. Water needs to remain in a sufficiently well defined position to be resolvable by x-ray crystallography. If a single water or cluster of water molecules experiences many ways of saturating their hydrogen bonds or they can adopt only a few of all possible hydrogen bonds, the water oxygen positions may be disordered and therefore not be visible in the crystal structure regardless of the resolution.[46, 47, 34, 48] Nevertheless, it is possible that a cavity, also having a sufficient size to contain one or more waters, is partially or completely empty.[49] The large SNase cavity mentioned above is very hydrophobic in the wild-type (WT). It is located roughly in the volume confined by the residues 25, 62, 66 and 92 in Figure 6. It is not a surprise that all WT crystal structures lack water in this cavity. With the introduction of polar residues in the proximity of this region the hydrophobicity of the cavity is changed significantly and buried water molecules have been found at 8 positions in this cavity in different variants.[50] Altogether this lead to two suggestion: (I) in some variants there may be more waters in the crystal structure then visible and (II) water may partially or even completely flood the central cavity as a consequence of the mutated residue becoming charged.

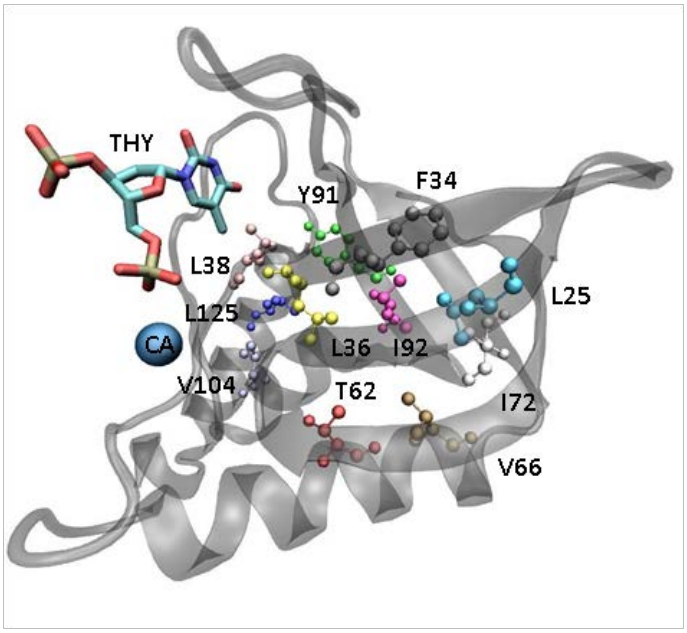


Figure 6: Crystal structure of the SNase variant D+PHS (PDB code 3BDC)[51] generated with VMD[29]. The protein backbone trace is displayed in gray. The blue sphere is the Ca^{2+} ion, the inhibitor thymidine-3',5'-diphosphate (THY) is displayed in the upper left part as stick model. The eleven residues, which are subject to one of the 20 considered point mutations, are shown as ball and stick models.

Cavities and electrostatic energy calculations

There are two strategies to include internal water molecules in a protein for electrostatic energy calculations. The waters can either be considered explicitly by including their atomic partial charges in the calculations or implicitly by filling the volume occupied by the water molecules with a dielectric medium whose dielectric constant is $\varepsilon_w = 80$, like it is done for bulk water. The first method has the drawback that the result of a calculation is very sensitive to both the exact amount and positioning of the internal waters. As discussed above information about these parameters can often not be obtained reliably from a protein crystal structure. The second strategy, the use of an implicit description, has the advantage that precise coordinates are not required. The latter approach is therefore the one pursued in this work.

The SES-algorithm discussed in chapter 2.2 is appropriate to define the protein volume for convex and weakly concave protein surfaces. It also detects crevasses and cavities inside the protein if they are large enough. However, it may underestimate the volume of narrow cavities, and tube-like cavities with diameters of 2.8 Å or just below may not be detected at all. To define the boundaries of protein cavities more reliably a new algorithm was developed, named cavity-algorithm in the following. This algorithm defines cavity volumes completely on the basis of geometric criteria and will be discussed in detail in the next section "*Modeling protein cavities with the cavity-algorithm*". It should be noted here, that in the publication for this work the term SASA (Solvent Accessible Surface Area) is used instead of SES (Solvent Exclusion Surface). Both terms refer exactly to the same concept.

For an electrostatic energy calculation the spatial dependence of the dielectric constant is represented on a discrete grid (ε-grid). The ε-grid has the same resolution as the grid that is actually used to solve the Poisson-Boltzmann equation. To obtain a complete description of the spatial dependence of the dielectric constant, first the conventional SES-algorithm is applied and the result is mapped on the ε-grid. In a second step all ε-grid points that were found to belong to a cavity according to the cavity-algorithm are set to $\varepsilon_{cav} = 80$. The resulting spatial dependency of the dielectric constant is illustrated in Figure 7. Note that this procedure would keep protein cavities found by the conventional protein surface algorithm before, although it is highly unlikely that the SES-algorithm finds a protein cavity not found by the new cavity-algorithm. Both, the cavity algorithm and the transfer of the cavities onto the ε-grid, have been implemented into the software TAPBS. Therefore, the new procedure can be directly applied with Karlsberg$^+$ to investigate the influence of cavities on pK_A calculations.

All results presented in this chapter are obtained with the sc_{pH7} protocol[13] of Karlsberg$^+$. This protocol includes only one pH adapted conformation (PAC) optimized at pH 7. The reason for this choice was that none of the point mutated residues forms a salt bridge. Many SNase variants were crystallized with thymidine-3',5'-diphosphate as an inhibitor and a bound calcium ion. The inhibitor was absent in all pK_A titration experiments and the same is assumed for the calcium ion

here. Therefore, inhibitor and ion were removed from the crystal structure together with all crystal water prior to a pK_A calculation.

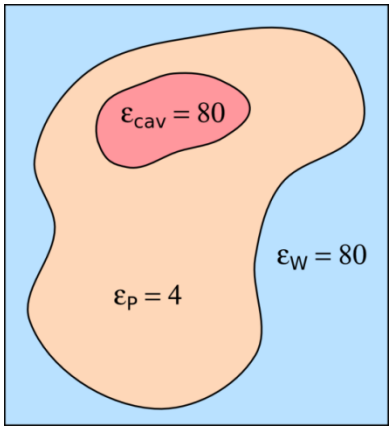


Figure 7: Illustration of the special dependency of the dielectric constant for a protein that contains a cavity. The brown and blue volume is defined by the SES-algorithm, the red volume is the cavity volume found by the cavity-algorithm and filled with a dielectric medium of $\varepsilon_{cav} = 80$. All cavities found by the cavity-algorithm are considered to be filled with water.

Modeling protein cavities with the cavity-algorithm

The new procedure to define protein cavities more reliably consists of three steps. All steps operate on a high resolution grid (grid constant $a = 0.1$ Å) that embeds the protein. Figure 8 illustrates this three step procedure. The cavity-algorithm uses libraries from APBS version 0.4[18].

(1) In the first step all grid points which are in the vdW-volume of protein atoms are marked as belonging to the "protein".

(2) For the remaining (non-protein) grid points a "ray"-algorithm is applied to explore their neighborhood, especially to find out whether these grid points are part of the bulk water volume or should be considered to be potentially part of a cavity. The approach is similar to the so called "finger"-algorithm used in other studies.[52, 53] A cavity, as defined here, must not necessarily be completely isolated from the bulk water but can also be "open". The exact point of transition between bulk water and an open cavity (deep crevasses) cannot be defined unambiguously. Anyway, in most of these regions both the SES- and the cavity-algorithm lead to the same results about the volume accessible for water. The ray-algorithm uses thirteen grid based rays. All thirteen rays cross in the grid point to be characterized (reference grid point). Their length is only limited by the dimensions of the grid. Orientations of the thirteen rays are defined with respect to a dice with the reference grid point in its center and an edge length twice the grid constant a: three rays connect opposite face center grid points, four rays are along the spatial

diagonals and six rays connect the centers of diagonally opposite edges. Each half-ray corresponds to a finger which is declared to be filled if it contains at least one grid point belonging to the protein. For a given reference grid point one counts the number of filled fingers, n_{filled}. Grid points not belonging to the protein with $n_{\text{filled}} < 21$ are considered as being part of bulk water, while grid points with $n_{\text{filled}} \geq 21$ potentially belong to deep crevasses or cavities inside the protein. The latter are named gap points in the following. Gap points potentially belong to a cavity but may alternatively also belong to the protein.

(3) In the third step these gap points are further analyzed and classified to belong to protein or cavities using the sphere-algorithm. The purpose of this algorithm is to smooth cavity walls and avoid considering small interstitial volumes between vdW spheres of protein atoms as part of the cavity. For this purpose, the center of a sphere is placed at each grid point P_i whose radius 1.4 Å corresponds to the size of a single water molecule. The number of gap points $N_{\text{sphere}}(i)$ inside the sphere is then counted. With grid constant a this sphere contains not more than $N_{\text{sphere}}(\text{max}) = (4\,\pi)\,/3\,(1.4/a)^3$ grid points. To explore whether the neighborhood of the gap point P_i belongs to a cavity whose diameter is large enough to host at least one water molecule, $N_{\text{sphere}}(i)$ is compared with $N_{\text{sphere}}(\text{max})$. If the inequality

$$N_{\text{sphere}}(\text{i}) > N_{\text{sphere}}(\text{max}) \times \text{c} \qquad (35)$$

is valid all gap points inside this sphere remain gap points but obtain the additional label "cavity". In this equation c is an adjustable parameter ($c \in [0.0, 1.0]$) named cavity parameter. If the inequality is violated no action is applied. After this sphere has been placed at all gap points, the gap points which did not obtain the label "cavity" are considered to belong to the protein, while the other gap points belong to cavities. The cavity parameter $c = 1.0$ essentially corresponds to the conventional method determining the electrostatic boundary between protein and solvent as used by Karlsberg+ and many other programs. Nevertheless, discrepancies could arise since the conventional method of defining the protein surface uses a vdW-surface based grid and maps the results directly to the low resolution ε-grid, while the sphere-algorithm applied in the present study uses a volume based high resolution grid. Thus, with finite grid resolutions, protein cavities with a diameter of just about 2.8 Å may be detected with one method but not with the other. On the other hand, the conventional algorithm as well as the new one (with large enough cavity parameter c) are constructed such that they do not detect cavities far too small to host at least one water molecule, which in bulk water at room temperature and ambient pressure occupies the average volume of 30 Å^3 .

To host water molecules a protein cavity should have a diameter larger than 2.8 Å. Hence, a cavity parameter of $c = 1.0$ should be sufficient. But, to account for the finite grid resolution (grid constant $a = 0.1$ Å) and to introduce an additional tolerance for tightly packed buried water molecules, a cavity parameter c smaller than unity is used. Figure 9A shows protein cavities found by the cavity-algorithm with cavity parameter $c = 0.7$ in the crystal structure of the

PHS/V66E SNase variant (PDB code: 1U9R). In Figure 9B the volume shape of the cavities found in SNase using the new cavity-algorithm with *c* = 0.9 and *c* = 0.6 are compared.

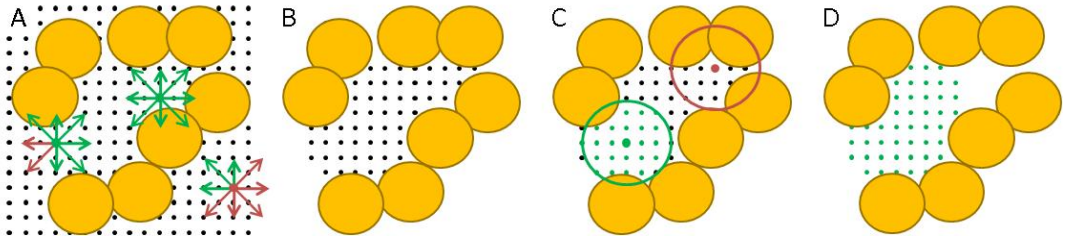


Figure 8: Two-dimensional schematic illustration of how protein cavities are found using the new cavity-algorithm. Yellow circles symbolize the vdW spheres of protein atoms, the dots belong to a high resolution spatial grid.
A: The grid points cover the space not occupied by the protein defined the vdW-volume of the protein atoms. The systems of rays centered at three different grid points denote the application of the ray-algorithm schematically. The ray-algorithm is applied to all grid points to locate empty volumes inside the protein. Green rays meet protein atoms, red rays point to bulk water only. **B:** If for an individual grid point a certain number of its rays hit protein atoms, they are kept as potential cavity points (the so called-gap points). All other grid points are deleted. **C:** These gap points are filtered with the sphere-algorithm finally leading to the cavity points displayed in part **D.** The sphere-algorithm places spheres of radius 1.4 Å on each gap point. If the number of gap points the sphere contains is larger than $c \cdot N_{max}$, where $c \in [0.0, 1.0]$ is the cavity parameter (see eq. (35)) and N_{max} being the maximum possible number of grid points, all gap points inside the sphere are considered to belong to a protein cavity. The green sphere in part C denotes a case where this criteria is fulfilled, while this is not the case for the red sphere.

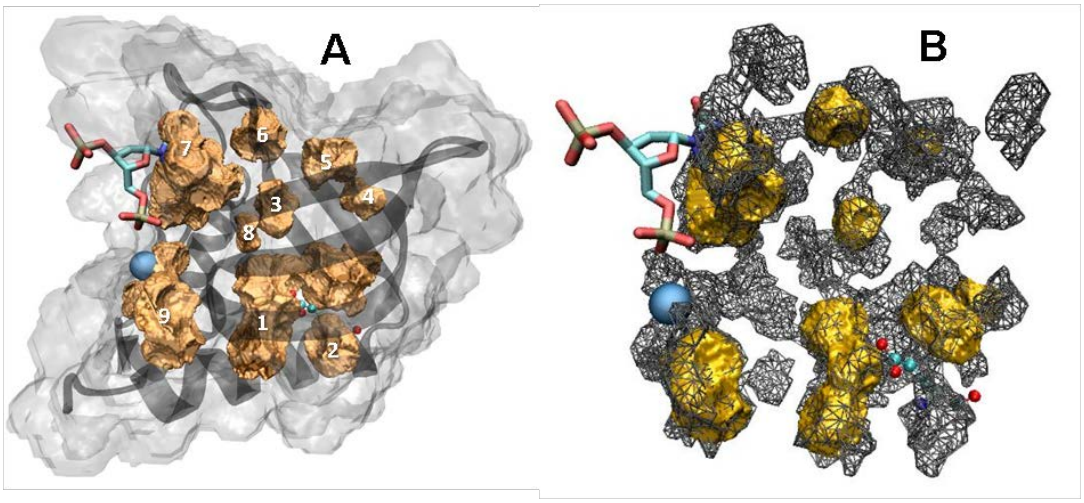


Figure 9: Protein cavities found in the SNase variant V66E/PHS (PDB code 1U9R)[32] using the cavity-algorithm. The SNase structure is displayed in the same orientation as in Figure 6. As orientation marks are introduced Ca^{2+} (blue sphere on the left side), thymidine (stick model above Ca^{2+}) and the point mutated residue Glu66 (ball and stick model in the lower right). Glu66 is in contact with the cluster of the three closely related cavities 1-3, which form the large central cavity. All structures are generated with VMD[29]. **Part A:** The transparent light gray surface shows the protein surface defined by the SES-algorithm, total protein volume $V_{protein}$ = 19027 Å^3. The dark gray rubber band is the trace of the protein backbone. The solid brown surfaces mark the protein cavities (total cavity volume V_{cavity} = 849 Å^3) found with the cavity parameter c = 0.7. The white numbers label the cavities whose volumes are listed in Table 1. **Part B:** Protein cavities found with cavity parameter c = 0.9 (yellow surfaces, total cavity volume V_{cavity} = 419 Å^3) and c = 0.6 (gray surfaces, total cavity volume V_{cavity} = 1140 Å^3).

3.3. **Results**

In the following section two aspects of the cavity-algorithm are examined. First, the ability of the cavity-algorithm to reliably find cavities and the influence of the cavity parameter on the generated cavity volumes is studied. Second, the influence of integrating cavity volumes into the electrostatic energy calculations on the accuracy of pK_A calculations with Karlsberg$^+$ is examined.

Definition of cavities

The ability of the cavity-algorithm to locate cavities as well as the effect of different choices for the cavity parameter are examined based on the example of the V66E/PHS SNase variant. The crystal structure of this variant has the PDB code 1U9R. Shape and location of the found cavities for three different cavity parameters (c = 0.9, 0.7 and 0.6) are displayed in Figure 9. All cavities found in the structure for c = 0.7 are listed in Table 1. Note that only 9 of all 15 cavities that were obtained at c = 0.7 and listed in Table 1 are visible in Figure 9A.

An interesting example for the necessacity of a more reliable method to determine cavities is the case of the thymidine binding pocket. The V66E/PHS variant crystal structure contains a thymidine-3',5'-diphosphate which definitely should create a pocket in SNase if removed due to its size and slim shape. But its removal does not lead to the appearance of a cavity neither by using the SES-algorithm nor by using the new cavity-algorithm with c = 1.0.

A detailed analysis of this situation revealed that a large part of thymidine is solvent exposed (see Figure 9A), while the buried part is in a very narrow pocket. In addition there are five crystal waters in the thymidine neighborhood forming a narrow tube. Removing thymidine and the crystal waters the SES- as well as the cavity-algorithm with c = 1.0 generates only a shallow pocket on the protein surface. This pocket can be observed visually, but is too shallow to be classified as cavity by the cavity-algorithm, while the water tube and the buried parts of thymidine are not detected at all. However, reducing the cavity parameter to c = 0.9 a cavity with a volume of $V_{cavity7}(0.9) = 153$ Å^3 appears (see cavity No 7 in Figure 9B). This cavity does include the volume from all five buried waters.

When the cavity parameter is reduced to c = 0.9, 9 cavities are found that are large enough to host at least one water molecule. The total volume of these cavities is $V_{cavity}(0.9) = 419$ Å^3 as compared to the total volume of the protein $V_{protein} = 19027$ Å^3 including these cavities. Decreasing the cavity parameter further to c = 0.7 the total cavity volume increases to $V_{cavity}(0.7) = 849$ Å^3 and the number of cavities to 15. From these 15 cavities (listed in Table 1) seven are closed (no access to bulk water) and eight are open (with access to bulk water). The largest cavity No 7 ($V_{cavity7}(0.7) = 214$ Å^3) is related to the binding pocket of thymidine. In the volume of the second largest cavity No 1 ($V_{cavity1}(0.7) = 166$ Å^3) crystal waters were observed. With a cavity parameter of c = 0.6 the total cavity volume increases to $V_{cavity}(0.6) = 1140$ Å^3. But, now the cavity number diminishes slightly to 13, since new cavities appear at c = 0.6, while

cavity	volume (Å^3)	access
1	165.7	open
2	27.3	open
3	26.8	closed
4	13.4	closed
5	58.4	open
6	39.3	closed
7	214.0	open
8	24.5	closed
9	88.4	open
10	14.4	open
11	16.3	closed
12	54.6	closed
13	49.6	open
14	26.9	closed
15	29.6	open

Table 1: Cavities and their volumes in the crystal Structure with PDB id 1U9R obtained with the new cavity-algorithm with a cavity parameter of *c* = 0.7. Cavity numbers refer to Figure 9A. A cluster of the cavities 1 to 3 located in close proximity to each other form the main cavity of SNase. The last column indicates whether the cavity has access to bulk water (open) or not (closed).

cavities already existing at c = 0.7 merge at c = 0.6. With a cavity parameter of c = 0.5 the total cavity volume increases further to $V_{cavity}(0.5) = 1430$ Å^3, which may overestimate number and size of cavities.

Finally, the accuracy of the cavity-algorithm was probed by computing the cavity volumes for different orientations of the same structure. Generating rotated structures of the V66E/PHS variant for eight arbitrary orientations, a very small average volume variation of 0.35% is obtained for closed cavities, while the volume variation is much larger for open cavities varying by 3.6%. The large uncertainty in the volume of open cavities is due to the finite orientational resolution of the ray-algorithm. This resolution is critical for the definition of the transition region from bulk water to cavity volume and could be increased by using more rays. But since usually the volumes generated by both, the SES- and the cavity-algorithm, do overlap in this region anyway, the total protein volume is not influenced by this uncertainty.

pK_A calculation

The pK_A values for the 20 SNase variants computed with the new version of Karlsberg$^+$, named *Karlsberg$^+$(cav)*, that includes the cavity-algorithm are shown in Table 2. The table also contains the experimentally measured pK_A values as well as the total cavity volumes. Due to subtle changes in Karlsberg$^+$ related to the random number seed used to generate the amino acid side chain conformations of the PACs, the pK_A values computed for the Telluride meeting in 2009 differ in some details from the pK_A values obtained now with the actual Karlsberg$^+$ version. However, the overall root mean square deviation ($RMSD_{pKa}$) between computed and measured pK_A values remained nearly the same with being 8.7 and 8.8 pH units for the previous and actual Karlsberg$^+$ version respectively. Using a cavity parameter of c = 1.0, the PACs of all 20 considered SNase variants did not show any further cavities, besides the open cavities already found by the SES-algorithm. Therefore the pK_A values obtained with this parameter are the same as those obtained with the unmodified Karlsberg$^+$ version.

The computed pK_A values for cavity parameters in the range of c = 1.0 to 0.6 are listed in Table 2. A more detailed dependence of the 20 computed pK_A values on the cavity parameter is shown in Figure 10. The $RMSD_{pKa}(c)$ of the 20 pK_A values for the original Karlsberg$^+$ is $RMSD_{pKa}(1.0) = 8.8$ and diminishes with decreasing cavity parameter c to $RMSD_{pKa}(0.9) = 6.8$, further to $RMSD_{pKa}(0.7) = 4.7$ and finally to $RMSD_{pKa}(0.6) = 3.6$ with decreasing cavity parameter c. The $RMSD_{pKa}$s for c > 0.8 do not include the V104E variant, since calculations returned only a limit for

the pK_A value (>20.0). The RMSD for the NULL hypothesis that assumes that all pK_A values in the protein are the same as in solution yields $RMSD_{pKa,NULL} = 3.8$.

It can be expected that the influence of the cavity-algorithm on the results strongly depends on the structural details around the residue whose pK_A is calculated. The locations of 20 point mutations that are considered here are scattered over the whole interior of the SNase protein. Four positions are especially interesting since they are located in direct proximity of the main cavity (see Figure 9). In Figure 10 the SNase variants are therefore split into two groups, one group (A) containing the 9 mutations at positions 25, 62, 66 and 92 and a second group (B) for all other mutations (except for Y91E). The last three rows of Table 2 show the total $RMSD_{pKa}$s of each group.

The best agreement between computed and measured pK_A values is obtained with cavity parameters of c = 0.7 and c = 0.6. For the cavity parameter of c = 0.7 the following results were obtained for the 20 pK_A values as listed in Table 2. The pK_A values of three residues (T62K, V66E, and I92E) could be reproduced with less than 0.5 pH unit deviation from the measured values. For six pK_A values (L25E, L36K, V66D, I72E, I92D and V104K) the deviation was less than 2 but larger than 0.5 pH units. In seven cases (L25K, L38E, V66K, Y91E, I92K, V104D and V104E) the result could be improved, but the deviations from the experimental pK_A values were still larger than 2 pH units. For four SNase mutants (F34K, I72E, I72K and L125K) the introduction of the cavities had only a small influence on the computed pK_A values. In one case (L38K) the introduction of cavities resulted in a worse result. In two cases listed before (L38E, V66K) a considerable overshooting was observed when decreasing the cavity parameter too much, such that the deviation to the experimental pK_A value was of opposite sign for c = 1.0 and c = 0.7 respectively.

For the mutant Y91E the SNase appeared as a dimer in the unit cell of the crystal. Here both SNase monomer structures were considered individually. With a cavity parameter of c = 1.0 the computed pK_A values are very different from each other and both deviate strongly from the measured pK_A values. Generating the cavities with c = 0.7 the deviations from the measured pK_A were reduced, while the conflict of the two computed pK_A values remained. The reason for this conflict is probably the complex environment of Glu91, which is very close to two other acidic groups (Glu75, Asp77) and one histidine (His121).

Table 2: List of pK_A values computed with a modified version of Karlsberg+ that uses the new cavity-algorithm to account for cavities. The pK_A values of 20 SNase variants have been calculated for different values of the cavity parameter c. Except for Y91E all pK_A values are also shown in Figure 10. The last three rows contain the $RMSD_{pKA}(c)$ of **A:** mutated residues located near the central cluster of cavities 1-3, **B:** residues located distant to these cavities and **C:** all residues together.

SNase variant						pK_A RMSD				
PDB id	**crystal pH[a]**	**mutation / SNase type**	**residue**	**cavity volume (Å³)**	**exp. pK_A**	**c = 1.00**	**c = 0.90**	**c = 0.80**	**c = 0.70**	**c = 0.60**
3EVQ	6	L25E/Δ+PHS	Glu25	81.3	7.5	15.4	13.6	12.2	9.2	7.3
3ERQ	8	L25K/Δ+PHS	Lys25	96.3	6.3	-4.9	-3.4	-1.7	2.7	6.5
3ITP	8	F34K/Δ+PHS	Lys34	16.1	7.1	-2.4	-1.8	-1.5	-1.3	5.2
3EJI	8	L36K/Δ+PHS	Lys36	78.1	7.2	-1.3	-0.4	4.5	8.3	9.2
3D6C	7	L38E/PHS	Glu38	66.8	7.2	12.0	6.2	5.2	3.1	1.8
2RKS	10.5	L38K/PHS	Lys38	126.0	10.4	9.0	11.6	13.1	13.7	12.9
3DMU	8	T62K/PHS	Lys62	90.9	8.1	-2.6	-2.1	3.9	7.7	8.9
2OXP	6	V66D/PHS	Asp66	161.8	8.7	14.8	10.2	8.0	6.9	6.5
1U9R	6.4	V66E/PHS	Glu66	176.3	8.5	16.7	12.2	9.6	8.7	7.2
3HZX	9	V66K/Δ+PHS	Lys66	137.1	5.76	-2.3	3.3	7.8	9.1	10.1
3ERO	9	I72E/Δ+PHS	Glu72	87.2	7.3	5.9	5.8	5.9	5.8	5.7
2RBM	8	I72K/Δ+PHS	Lys72	33.7	8.6	13.7	13.7	13.7	13.9	13.6
3D4D[b]	8	Y91E/Δ+PHS	Glu91	98.3 / 96.4	7.1	-1.3 / 19.5	-1.6 / 14.7	15.5 / 12.8	2.2 / 10.6	3.3 / 4.6
2OEO	6	I92D/Δ+PHS	Asp92	167.1	8.1	16.0	10.7	8.7	7.3	6.3
1TQO	6	I92E/Δ+PHS	Glu92	142.2	8.7	14.6	10.1	9.0	8.3	7.0
1TT2	8.1	I92K/Δ+PHS	Lys92	149.0	5.6	-5.5	-2.8	-0.5	3.1	4.2
3P75	9.5	V104D/Δ+PHS	Asp104	35.9	9.7	17.4	15.1	14.6	13.6	10.1
3H6M	8	V104E/Δ+PHS	Glu104	43.5	9.4	>20.0	>20.0	18.9	17.9	17.3
3C1F	7	V104K/Δ+PHS	Lys104	55.1	7.7	-1.1	-1.0	7.8	8.7	9.6
3C1E	9	L125K/PHS	Lys125	22.1	6.2	-9.9	-7.9	-6.7	-6.7	-3.0
pK_A RMSD (c)		**A) 9 residues near cavities 1-3 (Figure 10A)**				**8.77**	**6.12**	**4.04**	**2.04**	**1.95**
		B) 10 (except Y91E) other residues (Figure 10B)				**8.75**	**7.27**	**6.28**	**6.01**	**4.55**
		C) all 20 residues				**8.76**	**6.75**	**5.39**	**4.66**	**3.62**

[a] The pH value of the crystal as given in the corresponding PDB file.

[b] The SNase appears as dimer in the unit cell. Both monomers of the dimer are considered for the pK_A computations.

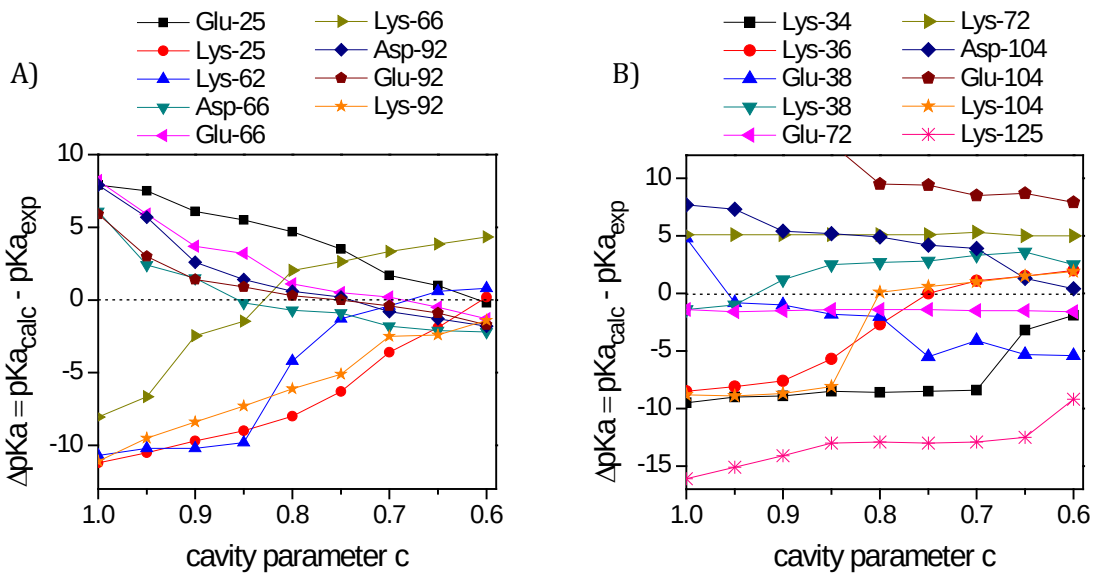


Figure 10: Computed difference between calculated and measured pK_A values ($\Delta pK_A = pK_{A,calc} - pK_{A,exp}$) for 19 SNase mutants as a function of the cavity parameter *c*. The results of the mutant Y91E where the two monomers of the SNase dimer yield conflicting pK_A values is not displayed (see discussion in text). The cavity volumes become larger with smaller *c*. A cavity parameter of $c = 0.7$ yields the best compromise between agreement with measured values with an RMSD of 4.66 pH units and a moderate cavity parameter (see discussion in text). **Part A** shows the pK_A values of nine SNase point mutations that are located near the main cluster of three cavities (1-3) (see Figure 3A). For the other ten SNase point mutations the mutated residues are located more distant from these cavities (**part B**). The exact location of the mutated residues in SNase is displayed in Figure 6. The shape and location of the cavities are displayed in Figure 9.

3.4. **Discussion**

Definition of cavities

The SNase crystal structures for both the wild type (PHS / Δ+PHS) and the variants with point mutations do contain buried water molecules at different positions, e.g. in the thymidine binding pocket or in the central cavity (composed of the cavities 1-3 in Figure 9). The conventional SES-algorithm could not detect these cavities and considered the occupied volume as being part of the protein. The new cavity-algorithm was able to detect these water molecules already at a very conservative cavity parameter of $c = 0.9$.

In the SNase crystal structures the main cluster of cavities 1-3 shows crystal water occupancies of up to five waters simultaneously at eight different positions.[50] A single water molecule occupies under normal conditions of room temperature and 1 atm pressure an average volume of 30 Å^3 in bulk water. In the crystal structure of the V66E/PHS variant the three cavities 1-3 possess a total volume of $V_{cavity1-3}(0.9) = 108$ Å^3. This volume would be enough to host 3 water molecules, it is, however, much too small to host all of the eight crystal waters that should in principle fit into these cavities. With a smaller cavity parameter of c = 0.7 the volume increases to $V_{cavity1-3}(0.7) = 220$ Å^3 and becomes large enough to host about 7 to 8 water molecules. With further decrease of the cavity parameter to c = 0.6 this volume becomes $V_{cavity1-3}(0.6) = 305$ Å^3

what may be too large. Therefore a cavity parameter in between 0.9 to 0.7 seems to be a reasonable choice.

The question whether the cavity volumes found by the cavity-algorithm are actually filled with water or not cannot be answered with the method applied here. Damjanovic et al[54]. obtained qualitative agreement between the position of internal crystal waters found in SNase using molecular dynamic simulations (MD simulations) and in the crystal structure. But it was also speculated that SNase may contain more waters then observed in the crystal structures, which are however partially disordered and therefore not detected.[34, 50]

pK_A calculation

From the 20 mutated amino acids that were considered here, nine are located in close proximity to the main cluster of cavities 1-3 and are therefore expected to be strongly influenced. These residues are inserted into the SNase protein at the positions 25, 62, 66 and 92 (see Figure 6 and Figure 9). How the pK_A values of these variants vary with the cavity parameter is shown in Figure 10; part A shows this dependence for the nine residues close to the main cluster of cavities part B the remaining ten residues. All nine computed pK_A values in part A show considerable improvement with increasing cavity volume except for the pK_A of Lys66, which exhibits overshooting. In the interval $c \in [0.6, 0.7]$ the variation of the individual nine pK_A values is small and the $RMSD_{pKa}$ decreases only slightly from $RMSD_{pKa}(0.7) = 2.04$ to $RMSD_{pKa}(0.6) = 1.95$. Regarding the dependence of the other eleven pK_A values on the cavity parameter c, no significant improvement can be observed down to c = 0.7, except for Lys36. With a further reduction of the cavity parameter to c = 0.6 a significant improvement of the computed pK_A values can be observed in three cases (Lys34, Lys36 and Asp104), which may however just be the result of an unspecific increase of the volume with high dielectric constant. This leads to the conclusion that the most reliable results are obtained with a cavity parameter of c = 0.7, where the main cluster of cavities has just the right size to host all of the eight water molecules that are found in crystal structures of SNase variants.

One could argue that the nine pK_A values of the titratable residues in part A are predicted with higher accuracy ($RMSD_{pKa,A}(0.7) = 2.04$) than the other eleven pK_A values ($RMSD_{pKa,B}(0.7) = 6.01$) because the pK_A shifts of the former are smaller and therefore easier to predict. However, the average absolute shift of the measured pK_A values are 4.1 and 3.2 pH units for the first and second group, respectively, ruling out this argument. This can also be seen by comparing the results obtained with the NULL hypothesis for the two groups, providing $RMSD_{pKa,NULL,A} = 4.2$ and $RMSD_{pKa,NULL,B} = 3.5$ respectively. The NULL hypothesis assumes that the pK_A of a residue embedded in protein is not shifted at all compared to the same residue being surrounded only by an aqueous solution.

The SNase crystal structures were all resolved at pH values between 6 and 10.5. The individual pH values are listed in Table 2. Accordingly, residue Glu72 is negatively charged, while Glu25, Lys25, Asp66, Glu66, Lys66, Asp92, Glu92, Lys92, Glu104 and Lys125 are charge neutral in the

corresponding crystal structures. For the other nine titratable residues no such clear statement can be made, since they possess pK_A values that deviate less than one pH unit from the pH value of the crystal. Whatever the protonation in the crystal is, the individual crystal structure used for a pK_A calculation can only cover one of the two protonation states a titratable residue can adapt to. Larger conformational changes due to the protonation or deprotonation of a residue can result in discrepancies of the computes pK_A values. Only the two residues Glu72 and Lys72 are solvent exposed. If a titratable residue is not solvent exposed or close to a cavity filled with water, it becomes very likely that the protein reacts with a conformational change to the charged state since the charge is not stabilized sufficiently by solvating water. This may explain that the four titratable residues which are neither solvent exposed nor close to any cavity do all yield computed pK_A values that are far from the measured value. These are the residues Lys34, Asp104, Glu104 and Lys125.

Specific aspects for individual SNase variants

In the following section some of the SNase variants whose calculated pK_A values deviate strongly from the experimental values are discussed in detail.

L38K and L38E: The two SNase variants L38K and L38E were discussed in detail elsewhere.[36, 38] The position 38 in SNase is in close proximity to the residues His121 and Glu122. Due to the mutual electrostatic interactions between these groups even small structural changes that are not considered correctly can have a significant influence on the computed pK_A values. For Lys38 it was shown that ionization is accompanied by water penetrating the protein and partially hydrating the lysine.[36]

I72K: The SNase crystal structure of the I72K point mutation was resolved at pH 8, while the measured pK_A value of Lys72 is 8.6. In this case the SNase crystal structure probably involves the protonated state of Lys72. This residue is not in contact with a large cavity. Hence, the new cavity-algorithm cannot improve the computed pK_A value. On the other hand, this residue is partially solvent exposed and its measured pK_A value is not too far from the reference value in solution. Nevertheless, its pK_A value could not be reproduced with standard Karlsberg$^+$. In the SNase crystal structure there are two backbone oxygen atoms that from hydrogen bonds with Lys72. However, one of the two H-bonds (O-N distance 2.5 Å) is too short. Calculations with a corrected crystal structure where the short H-bond is relaxed by energy minimization with CHARMM (with Lys72 protonated) yielded an increased H-bond length of 3.0 Å and a computed pK_A value of 11.1, which fits better with the measured value of 8.6 then does the computed pK_A of 13.9 obtained with the cavity-algorithm with c = 0.7. The $RMSD_{pKa}$ values given before do not include such corrections.

Y91E: The SNase crystal structure of the Y91E point mutation was resolved at a pH 8, while the measured pK_A value of Glu91 is 7.1. Hence, Glu91 is likely to be deprotonated in the SNase crystal structure. It is the only SNase crystal structure considered in this benchmark set that occurs as a dimer in the unit cell. Glu91 is located on the protein site opposite to the dimer

interface in close proximity to the residues Glu75, Asp77 and His121. The two monomers (α/β) of the dimer differ in the distance between Glu91 and Glu75 (shortest O-O distance of 3.1 Å for α and 5.0 Å for β) and His121 (4.0 Å for α and 4.9 Å for β). Due to this environment of strong electrostatic interactions the pK_A of Glu91 is very sensitive to even small structural changes. This results in an unusual titration behavior. The titration point is at very basic pH values but the residue can also become charged at acid pH values. As shown in Table 2 the latter case occurs only for the residue in chain A in the computations.

V104D and V104E: Both acids Asp104 and Glu104 are probably protonated, i.e. they are charge neutral, in the corresponding crystal structures. These two residues are neither solvent exposed nor are they close to a large cavity. Therefore, it is not surprising that their measured pK_A values exhibit the largest shifts adopting values of nearly 10. The computations overestimate these pK_A shifts dramatically. Glu104 is very close to Glu129, which is solvent exposed and therefore deprotonated, while Asp104 is far from Glu129. Nevertheless, for both acids nearly the same pK_A value has been measured. The shortest O-O distance between Glu104 and Glu129 is only 2.6 Å, which is likely too short. Geometry optimization of these two glutamates by energy minimization increased the O-O distance to 3.5 Å but did not improve the corresponding computed pK_A.

3.5. **Summary & Conclusions**

The software Karlsberg+ uses the standard SES-algorithm to define protein surfaces. The SES-algorithm is well suited for convex and weakly concave protein surfaces. However, the example of the SNase protein and its variants demonstrate that the ability of the algorithm to describe deep crevasses and cavities inside the protein is insufficient. The SNase crystal structures contain internal crystal waters and often a thymidine-3',5'-diphosphate located in a narrow binding pocket. The SES-algorithm could not find the corresponding cavities after removal of waters and thymidine. This deficiency has been identified as one reason for the large discrepancies between pK_A values computed with Karlsberg+ and the corresponding measured values for the SNase benchmark set.

To address this problem an improved algorithm has been designed named “cavity-algorithm”. The cavity-algorithm localizes cavities by identifying volumes in the protein interior and on its surface that are large enough to host one or more water molecules. The criteria applied by the algorithm are completely based on geometric considerations. No prediction about the actual water content of a cavity is made, e.g. by estimating the hydrophobicity. To evaluate the influence of the cavities on the results of pK_A calculations, all cavities are considered to be filled with water.

In particular the new algorithm is capable of characterizing the main cavity in SNase that is, in the case of the V66E/PHS variant, composed of a cluster of three cavities (cavity 1-3 in Figure 9). In crystal structures of different SNase variants crystal waters were found at eight different sites

located in the volume of these cavities[50]. With the optimized cavity-algorithm, i.e. employing a cavity parameter of c = 0.7, the total volume of these three cavities matches the volume needed to host up to eight water molecules.

The influence of the cavities on pK_A calculations has been studied for 20 point mutations of SNase. These variants were taken from the data set of the pK_A cooperative (www.pkacoop.org) and selected because both an experimental pK_A and a crystal structure were available at the time of the project. In each variant a titratable residue was introduced in the hydrophobic interior of the protein. For a first group of nine titratable residues, which are all close to the main cavity, the accuracy of the computed pK_A values did improve significantly with the cavity-algorithm compared to the standard version of Karlsberg+. With the optimal cavity parameter of c = 0.7 the $RMSD_{pKa}$ could be reduced from 8.77 to 2.04. For the second group containing the remaining eleven pK_A values the results could be improved only slightly reducing the $RMSD_{pKa}$ from 8.75 to 6.01, what is still a rather poor result.

For the second group of eleven residues with largely deviating pK_A values, there are obviously other factors that influence the pK_A values which are not yet included appropriately in the present computational approach. The main reason for such large deviations in pK_A values is probably not related to details of the computational approach like parameterization of atomic radii, charges or the protein dielectric constant, since Karlsberg+ normally obtains a much higher accuracy[13]. Besides the existence of large cavities, there is another aspect that distinguishes the SNase variants from the proteins used in previous benchmark sets. It seems very likely that for at least some of the residues the change in protonation is accompanied by significant structural changes.

A crystal structure can refer to only one of the two protonation states of a titratable residue. If the structure changes with the protonation state, the effect of the change cannot be considered appropriately if only the crystal structure is considered for the pK_A computation. Although it may be practically very challenging or even impossible, a second crystal structure obtained at an appropriate pH could solve the problem by representing the complementary protonation state. At the time of the project no such additional structure was available for any variant. Experimental evidences for such structural changes in SNase were found in some cases [34, 40, 45]. These changes presumably go along with water penetration[36, 55].

The results of this project showed that a faithful consideration of cavities in electrostatic energy calculations can significantly improve the accuracy of pK_A calculations. But it also became evident that this correction alone is not sufficient to predict the large pK_A shifts found in the SNase variants. It seems reasonable to assume that, in order to further improve the accuracy of Karlsberg+, the software has to account for the structural changes that occur if a residue switches its protonation state. This conclusion was the inspiration for the second project of this work, discussed in the next chapter.

4. Combining Molecular Dynamic Simulations and Electrostatic Energy Calculations to Improve pK_A Computation

The project presented in this chapter has been published in the time of the dissertation:

Meyer, T.; Knapp, E. W., pK_A Values in Proteins Determined by Electrostatics Applied to Molecular Dynamics Trajectories, *J. Chem. Theory Comput.*, **2015**, 11 (6), pp 2827–2840 DOI: 10.1021/acs.jctc.5b00123

The concept of the new procedure has been developed by me. I performed all calculations and analyses described in the paper. Prof. Ernst Walter Knapp was the supervisor of the project and formulated the new mathematical expressions for the electrostatic energy calculations.

The following text is a modified presentation of the project. The results have not changed, although the presentation differs in some aspects as described in the following. Additional results are presented that had been excluded from the publication for reasons of clarity. These additional results concern the two procedures named *KB2⁺MD** and *KB2⁺MD*(no<=1-4)*. They do not change the overall conclusions; instead they provide a further detailed insight into some of the factors that govern the accuracy of the pK_A calculations based on MDs. In the "*Material and Methods*" section the new mathematical expressions for the electrostatic energy calculations introduced in the publication are not discussed. Instead, the new procedures are introduced from a more technical point of view, focusing on how they are actually implemented. In this regard the text aims at a different goal in comparison to the publication, as the latter one aims at providing a general and complete mathematical framework for the new procedure. All figures and part of the text were taken from the publication. Reproduced with permission from J. Chem. Theory Comput., 2015, 11 (6), pp 2827–2840. Copyright 2015 American Chemical Society.

4.1. Introduction

The computation of pK_A values in a protein usually starts with a crystal structure of the corresponding protein. Every crystal structure has been resolved at the certain pH value and does therefore represent the protein very well at this pH. With variation of the pH the protonation state of titratable residues in the protein will change and as a result the protein conformation may change as well. These structural changes may involve subtle rearrangements in the hydrogen bond network, side chain reorientations, water penetrating the protein or even partial unfolding of the protein. To accurately predict pK_A values, a software has to take these structural reactions into account. The software Karlsberg⁺ addresses this challenge with the concept of generating so called self-consistent pH adapted conformations (PACs). Each PAC is a protein structure that has been automatically modeled to represent the protein in a certain pH

interval. The modeling procedures implemented in the current version of Karlsberg+ are quite limited and cover only the most common structural changes, i.e. the rearrangement of the hydrogen bond network and the opening of salt bridges at pH values between about 4 and 10. In this project I developed a new procedure that is still based on the idea of PACs but generalizes the modeling capability vastly by employing molecular dynamic (MD) simulations.

The idea to combine electrostatic energy computation with MD simulations for evaluating pK_A values is not new and has been discussed and applied extensively in the past in several variants.[56-67] The constant pH MD simulation technique is a special variant of MD simulation where the protonation of titratable groups can change during a run.[68-70, 12, 71-77] The electrostatic pK_A computations in proteins where reviewed recently.[78, 9] More recently Nilsson et al.[67] showed, with an approach similar to the one presented here, that structures from MD simulations can be used for electrostatic energy computations to predict pK_A values with very high accuracy. For the dimeric leucine zipper they analyzed frames from three MD trajectories that differed in the protonation pattern of titratable residues. They chose the following three protonation patterns: (1) all titratable residues (Arg, His, Lys, Asp, Glu) charged, tyrosine neutral; (2) glutamic acid and tyrosine neutral, all others charged; (3) lysine and tyrosine neutral, all others charged. For each pH value the protonation probability was averaged over all frames of an MD trajectory. These averaged results were then combined using a weighting function based on the protonation probabilities. In this work this concept is extended and an extensive benchmark computation is performed involving 194 measured pK_A values of 13 proteins.

The present study differs in several aspects from the work of Nilsson et al. described above. Electrostatic energies are employed to weight the data of different MD simulations and the protonation pattern employed for the MDs are based on preliminary pK_A values computed using the protein crystal structure. Furthermore the current study investigates how variations in the procedure influence the accuracy of the computed pK_A values. These are (i) the influence of 1-2, 1-3 and 1-4 electrostatic atom-pair interactions (ii) post processing of the MD structures by energy minimizing them with different dielectric constants and (iii) using different sets of atomic radii. Additionally, the accuracy of the pK_A computations is compared with the results obtained by the empirical prediction scheme PropKa[10, 79].

With a further modification of the procedure employed by Karlsberg+, introduced and tested in this study, the intermediate energy terms of the electrostatic energy calculations become easier to interpret, i.e. the desolvation energy of individual titratable residues, the residue-pair interaction energies and the conformational energy of the whole protein. These energy terms are no longer specified relative to a reference protonation pattern; instead absolute electrostatic energies are evaluated. It is now easily possible e.g. to determine whether interactions with the protein environment or interaction with a specific titratable residue is causing the pK_A shift of a titratable residue.

The new procedure differs in several central aspects from Karlsberg+, although it is based on the same concept of generating pH specific structures. Therefore, the procedure is tested and optimized with a benchmark set that contains standard proteins and is similar to the one that has been used in the past to benchmark Karlsberg+. The performance of the new procedure with the challenging SNase benchmark set is discussed separately in chapter 5.

4.2. **Materials and Methods**

Benchmark set of measured pK_A values in proteins

The procedures for calculating pK_A values were tested on a set of 13 proteins with known crystal structures and a sufficiently large number of measured pK_A values, containing 194 pK_A values in all. Eleven protein structures (PDB ids: 4PTI, 1PGA, 1A2P, 2LZT, 3RN3, 2RN2, 1HNG, 3ICB, 1PPF, 1ERT, 1XNB) have been taken from the set of 15 proteins used to benchmark the performance of the original version of Karlsberg+[13]. Since the new procedure discussed in the present study requires significantly more CPU time compared to Karlsberg+, proteins from the prior study with less than seven measured pK_A values were not included here. To enlarge the benchmark set, two proteins have been added: a leucine zipper (PDB id 2ZTA[80]), since it was considered in related work where pK_A values where also computed from MD simulation data;[67] and a staphylococcal nuclease (SNase) variant Δ+PHS (PDB id 3BDC[51]). The Δ+PHS variant is one of the two SNase variants (the PHS variant is the other) that are central proteins for the SNase benchmark set used in chapter 3[39], since all point mutations have been introduced into one of these two SNase variants. The majority of the measured pK_A values used in the current study were collected by Georgescu et al.[81] from literature. For xylanase the measured pK_A values are taken from Joshi et al.[82], for the leucine zipper they are provided by Matousek et al.[83] and for the SNase variant Δ+PHS by Castaneda et al.[51].

Necessary modeling steps for the benchmark set

All modeling steps were performed with the CHARMM[26] program using the CHARMM22 force field[15]. Any ligands or ions present in the crystal structures of the proteins were removed. Crystal water molecules were kept for the MD simulations, while they were removed for all electrostatic energy computations, if not stated otherwise. In case of the "third domain of the turkey ovomucoid inhibitor" (OMTKY3, PDB id 1PPF) only chain I was used, since the complete crystal structure also contains the protein PMN elastase. For staphylococcal nuclease variant Δ+PHS the first six and last eight residues are disordered in the crystal structure and therefore not resolved. These residues were not added by modeling, instead the termini of the crystal structure were neutralized with an acetylated N-terminus and a methylated C-terminus to avoid the artifactual introduction of charged residues. For the GCN4 leucine zipper one Gly was added to the N-terminus by modeling and a Glu was added to the C-terminus to reproduce the protein used to measure pK_A values[83]. The crystal structure for xylanase (PDB id: 1XNB) was modified slightly, interchanging the coordinates of the ND2 and OD1 atoms of residue Asn35, which has a

large influence on the pK_A value of the neighbor residue Glu172, as discussed later. For the protein rat T-lymphocyte adhesion glycoprotein CD2 (PDB id 1HNG) only part of the crystal structure involving residue 1 to 99 of chain A was used since that part was also used to measure the pK_A values. In case of thioredoxin (PDB id: 1ERT) the deeply buried crystal water 136 in chain A that is in contact with Asp26 was kept throughout all the calculations and all side chains of the protein were energy minimized with CHARMM[26], as it was done in the previous benchmark computation[13] to compensate for the influence of crystal contacts. Minimization was done in two steps, first in vacuum and then with an implicit water model using the GBSW[84] module in CHARMM[26].

pK_A computations with Karlsberg+

For comparison, all pK_A values in the benchmark set were also computed using the standard protocol of Karlsberg+ as described in chapter 2.4. This protocol consists in the automatic creation of eleven PACs: five for pH values of -10 and 20 respectively and one for pH 7. The results obtained here with Karlsberg+ may differ slightly from the values published earlier[13], due to minor adjustments in the protocol and the random search for salt-bridge geometries. Besides the standard protocol there is also an alternative simplified protocol available in Karlsberg+, which is called the sc_{pH7} procedure that generates only one PAC at pH 7.

Protocol of pK_A computations

General overview

The new protocol for calculating pK_A values introduced in the present project consists of five basic steps. Additionally several variants of the protocol were explored. All variants concern step 4 of the protocol and got an individual name starting with the prefix *KB2+MD*. The five basic steps common to all variants are in short:

1. Protein specific modeling of the structure as described above in section "Necessary modeling steps for the benchmark set".
2. Choosing a pool of protonation patterns.
3. Preparing and running a MD simulation for each selected protonation pattern.
4. Performing the electrostatic energy computations for each frame of the MD trajectories.
5. Averaging the results of the electrostatic energy computations within each MD trajectory and combining the averaged results to obtain titration curves and pK_A values.

Modeling and solvation with explicit water were performed with CHARMM[26]. Energy minimization and MD simulations were performed with NAMD[25] version 2.9. In the following steps 2-5 are described in detail.

2. Choosing the protonation pattern for the MD simulation

To prepare the MD simulation of a protein, for each titratable residue a complete protonation pattern has to be chosen. Such a set of protonation states is called a *'fixed protonation pattern*'

abbreviated here as '*FPP*'. Six residues are considered to be titratable: Asp, Glu, Lys, His, Cys and Tyr and the C- and N-termini (Cter and Nter). Since the residue Arg was not explicitly observed to be deprotonated in an intact native protein structure, it is not included as a titratable residue in the present study. The four selected *FPPs*, listed in Table 3, are the result of an extensive exploration of different possibilities. In the *FPP* denoted as *'pH<4*', all residues are protonated, corresponding to an extremely low pH value. Hence, Asp, Glu, Cys and Tyr are charge neutral while His and Lys are positively charged. The *FPP* denoted as '*pH5*' is obtained by determining the protonation of the individual titratable residues ('pK_A' in Table 3) with the simplified sc_{pH7} protocol of Karlsberg+, except for Lys, which is kept protonated. The *FPP* '*pH7*' is the same as '*pH5*', except that the acidic residues of type Asp and Glu are kept deprotonated. For titratable residues that had been added by the initial modeling procedures in step 1 at the N- or C-terminus of a protein, the protonation state was determined using the pK_A values of these residues in aqueous solution in comparison to the relevant pH value. In the fourth *FPP,* denoted as '*pH>10*', all titratable residues are kept deprotonated. Hence, in this case, His and Lys are charge neutral, while Asp, Glu, Cys and Tyr are negatively charged. As mentioned, before Arg was kept protonated throughout the present study; to treat it also as a titratable residue one would require an MD simulation with a fifth *FPP*, where Arg is also deprotonated.

Table 3: The *fixed protonation pattern* (*FPP*) used for the eight types of titratable residues in the four MD simulation runs (*pH<4*, *pH5*, *pH7*, *pH>10*). Arginines are always kept protonated. Nter and Cter are the N- and C-termini of the protein. Residues that are protonated or deprotonated are labeled as *'prot'* and *'deprot'*, respectively. The label 'pK_A' indicates that the protonation is set according to the results of a pK_A computation with the sc_{pH7} protocol of Karlsberg+ based on the crystal structure.

residue	*pH<4*	*pH5*	*pH7*	*pH>10*
Asp	*prot*	pK_A	*deprot*	*deprot*
Glu	*prot*	pK_A	*deprot*	*deprot*
Lys	*prot*	*prot*	pK_A	*deprot*
His	*prot*	pK_A	pK_A	*deprot*
Tyr	*prot*	pK_A	pK_A	*deprot*
Cys	*prot*	pK_A	pK_A	*deprot*
Nter	*prot*	pK_A	pK_A	*deprot*
Cter	*prot*	pK_A	pK_A	*deprot*

3. Preparing the MD simulation runs

For each *FPP* defined in the previous step, an MD simulation is performed. To prepare these MD runs, the corresponding *FPP* is modeled and the protein is packed in a box of TIP3 water molecules with periodic boundary conditions. The edge length of the cubic water box is equal to the maximum extension of the protein plus an additional 15 Å on all sides. First, only the water molecules and the hydrogen atoms of the protein were energy minimized while the remaining

atoms were kept fixed. Then a short MD simulation of 25 ps at 300 K without an explicit heating phase was performed with the same restraints, to equilibrate the water molecules and hydrogen atoms before the non-hydrogen atoms of the protein were allowed to move. The temperature was controlled using Langevin dynamics with a friction constant of 1 ps^{-1}. After these preparations, all atoms were energy-minimized and the system was heated up by velocity rescaling to 300 K within 50 ps in steps of 20 K. Finally the main MD production run of 10 ns was performed under NPT conditions using the Nosé-Hoover thermostat[85]. The first nanosecond of each MD run was used for equilibration and not considered in the analysis. Subsequently, trajectory frames at 100 ps intervals were extracted to be used for electrostatic energy computations in the next step resulting in 90 structures per MD simulation.

4. Performing the electrostatic energy computations

Each structure extracted from a MD trajectory is subsequently analyzed with electrostatic energy computations using the software APBS and TAPBS[13], a modified version of APBS[18]. The required energy terms are those necessary to evaluate equation (29), i.e. the desolvation energies ($\Delta\Delta G_{desolv}$) for each residue, the matrix W that contains all residue-residue pair interactions and the conformational energy (G_{conf}) of the protein with all titratable residues being in their corresponding reference state. All variations of the new procedures concern this step and can be categorized in three groups (*I* to *III)*. It is obligatory to make a choice about what variant from the first group should be used, while the variants of the second and third group are optional and can be applied additionally.

I. Variations of the protocol to obtain $\Delta\Delta G_{desolv}$, W and ΔG_{conf}. Besides using the standard protocol of Karlsberg+ (marked with *'**'*) as described in the sections "*Protonation energies*" and "*Conformational energies*" of chapter 2.2, two additional variants are introduced in the next section *"Protocols for electrostatic energy computations"*. One variant is named ***'no<=1-4'***, the other one is **without suffix** (*KB2+MD*).

II. The use of an alternative set of vdW radii to define the protein surface in the electrostatic energy calculations as described in the section *"Variation of the vdW radii"* named ***'Rashin'***.

III. As an optional step each structure was energy minimized with 1000 steps of conjugate gradient combined with line search minimization, using the standard energy minimization algorithm of NAMD 2.9[25] before the electrostatic energy computations were performed. This minimization was done using the complete system of water and protein as obtained from the simulation with periodic boundary conditions in a homogeneous dielectric of either $\varepsilon = 1$ or $\varepsilon = 4$. These two variants are named ***'ε_{min}=1'*** and ***'ε_{min}=4'***.

5. Averaging the results of electrostatic energy computations

For each of the four MD simulations (c) the electrostatic energy terms $\Delta\Delta G_{solv}$, W and ΔG_{conf} are averaged over all extracted structures, yielding $\langle\Delta\Delta G_{solv}\rangle^c$, $\langle W\rangle^c$ and $\langle\Delta G_{conf}\rangle^c$. These averaged electrostatic energy terms are inserted in equation (29) and titration curves are then computed as described in chapter 2.1 using the software Karlsberg[14] which applies a Monte Carlo (MC)

algorithm to determine the energetically most favorable combination of protonation states and MD for each pH. From the titration curves, the pK_A value of each residue is determined as the pH where the titratable residue is protonated with probability 0.5.

Protocols for electrostatic energy computations

The most straight forward way to obtain the required electrostatic energy terms in step 4 of the new procedure is to follow the traditional approach[21] as it is implemented in Karlsberg+ and described in detail in the sections "*Protonation energies"* and "*Conformational energies*" of chapter 2.2. This approach is named *KB2+MD** in the following. The *KB2+MD** procedure has two distinct disadvantages that are addressed by introducing the two variants of this protocol. The first problematic point is that *KB2+MD** does consider electrostatic interactions between atomic partial charges that are less than four covalent bonds away from each other. In the CHARMM force field these interactions are excluded, since classical electrostatic interactions do not apply for such atom pairs. Instead these short range intramolecular interactions are modeled by the so called "bonded" parameters discussed in chapter 2.3. These interactions are excluded in the *KB2+MD*(no<=1-4)* variant by using the software CHARMM to obtain the electrostatic self energy $G_{Coulomb}$ in equation (32) (see also Figure 3) when calculating the conformational energy ΔG_{conf}.

The other huge disadvantage of the traditional formalism is that the calculated terms for the relative electrostatic desolvation energy $\Delta\Delta G_{solv}$ and the interaction matrix W are not directly interpretable. The desolvation energy does contain the pairwise interaction energies with the reference states of all other residues, while the interaction matrix does lack these terms. As a consequence it is not possible to obtain the detailed interaction network of titratable residues that is required e.g. to answer the question why the calculated pK_A of a certain residue is strongly shifted. The issue has been solved with the procedure named *KB2+MD* (without any suffix). This variant differs considerable from the traditional approach. Its concept is described in the remainder of this section.

Implementation of the KB2+MD variant

The main difference of the *KB2+MD* variant compared to the original approach used by Karlsberg+ is the choice of the reference states for titratable residues. Before calculating the conformational energy of a structure all residues are set to their corresponding reference state. All entries in both the vector containing the desolvation energies and the interaction matrix are specified relative to the reference protonation pattern. As a consequence the protonation energy for an individual structure, as given with equation (28), becomes zero for the protonation vector $\mathbf{p_k}=\mathbf{p_{ref}}$, i.e. when all residues are in their corresponding reference state.

The reference state for a titratable residue can be chosen arbitrarily but Karlsberg+ follows two conventions: (i) The state has to be a physically meaningful state, i.e. it is one of the states the residue can actually occupy and (ii) it is a state where the residues total charge is zero (if possible). The *KB2+MD* variant introduces a new artificial state in which all individual atomic

partial charges of the residue are zero. Since this state cannot be occupied by the residue, it is only relevant for the calculation of electrostatic energy terms. This reference state is forbidden in the actual titration of the protein, what is accomplished by adding a sufficient energy penalty.

This choice causes a complication for the special case when a titratable residue contains the chemical group that forms the N- or C-terminus of the protein. The electrostatic potential Φ used to obtain the Born, back and interaction energies in equations (24), (25) and (27) is generated by the atomic partial charges q_m. For the new procedure *KB2⁺MD* q_m includes all atoms of the titratable residue. Since the chemical group forming the N- or C-terminus is treated as an individual residue this would result in the unwanted effect that the same charge occurs in two electrostatic potentials obtained for two different residues. To avoid the resulting double counting of electrostatic interactions, the following rule is applied. If a residue contains the chemical group of a titratable terminus, the residue is split into two partial-residues. The atoms of the terminal chemical group and the backbone are forming one partial-residue while the atoms of the side chain from a second one. Each partial-residue can then be considered individually for the electrostatic energy calculations and the titration. The problem does not occur for the conventional procedure since it uses only those atomic partial charges that do not have the same value for all protonation states to generate Φ. For the standard amino acids this definition never includes backbone atoms.

The conformational energy is calculated with the titratable residues being in their new artificial reference state. Additionally, as was done for the *KB2⁺MD*(no<=1-4)* variant, the short ranged intramolecular interactions are excluded in the *KB2⁺MD* variant by using the software CHARMM to obtain the electrostatic self energy $G_{Coulomb}$ in the homogeneous environment in equation (32). Since the vdW radii of the atoms of titratable residues remain unchanged in the artificial reference state they still influence the special dependency of the dielectric continuum just as they did before. However, they do not contribute directly to the electrostatic conformation energy G_{conf} due to the lack of atomic partial charges. As a consequence the absolute self energy of titratable residues is completely excluded from the protonation energy ΔG_p.

It can be expected that the *KB2⁺MD* variant yields very similar results for the protonation energies ΔG_p compared to the *KB2⁺MD*(no<=1-4)* variant. The main difference is that the absolute electrostatic self energy of the titratable residues is neglected in case of *KB2⁺MD*. Also the *KB2⁺MD* variant has the huge advantage of producing directly interpretable energy terms as intermediate results. The conformational energy contains only the electrostatic self-interaction energy of non-titratable residues. The desolvation energy contains all interactions between a titratable residue in a certain state with the protein and the solvent. Finally, the interaction matrix contains all the information of the pairwise electrostatic interaction energies in between all titratable residues in all states.

Variation of the vdW radii

The program Karlsberg+ uses atomic radii from the CHARMM22[15] force field to define the solvent excluded surfaces (SES) of protein and titratable residues in solution needed for the electrostatic energy computations as discussed in chapter 2.2. In a recent work Nilsson et al.[67] used the atomic radii introduced by Rashin et al.[86] to compute the pK_A values of a leucine zipper. These atomic radii have originally been used to find protein cavities and are generally smaller than those employed in our previous studies.[13] This is especially the case for the oxygen atom, whose radius is 1.4 Å according to Rashin et al.,[86] instead of values around 1.7 Å used for the different oxygen atom types in the CHARMM22[15] force field. To test the effect of these differences, the results have been recalculated using the '*Rashin*' atomic radii for the SES determination.

The software Karlsberg2+

The new procedure and all variants discussed here are implemented in the new version of the software Karlsberg+: Karlsberg2+. The complete protocol, except for the initial modeling decisions described in step 1, is performed automatically by the software Karlsberg2+. This ensures that all proteins in the benchmark set have been treated in exactly the same way and that the procedure can easily be applied to new protein structures for future applications.

4.3. Results

Molecular dynamic simulations

The average coordinate RMSDs of protein backbone relative to the corresponding crystal structures over the MD simulations of the 13 proteins are given in Table 4. For lysozyme (PDB id: 2LZT), the time evolutions of the coordinate RMSDs for all four MD trajectories are plotted in Figure 11. Although the non-equilibrium protonation pattern *pH<4* and *pH>10* differ considerably from *pH5* and *pH7*, for lysozyme the average coordinate RMSDs of the corresponding trajectories are similar: 1.3 Å and 1.2 Å for *pH<4* and *pH>10*, respectively, compared to 1.1 Å for both *pH5* and *pH7*. The same can be observed for the remaining 12 proteins, with an exception being the MD for the *pH>10* protonation pattern for xylanase (PDB id: 1XNB). The xylanase structure shows a linear arranged row of Tyr residues (resid 5, 69, 80 and 166) within a large crevasse that does also contain the active site formed by the two residues Glu78 and Glu172. With the ionization of the Tyr residues with the *pH>10* protonation pattern a two stranded beta sheet that consists of the residues 109 to 127 bends in a way that the crevasse becomes wider and more solvent exposed to compensate for the six negatively charged residues (the five Tyr and two Glu residues) in the crevasse. Overall the coordinate RMSDs of the proteins in the benchmark set are not sensitive to the protonation state on the timescale simulated here.

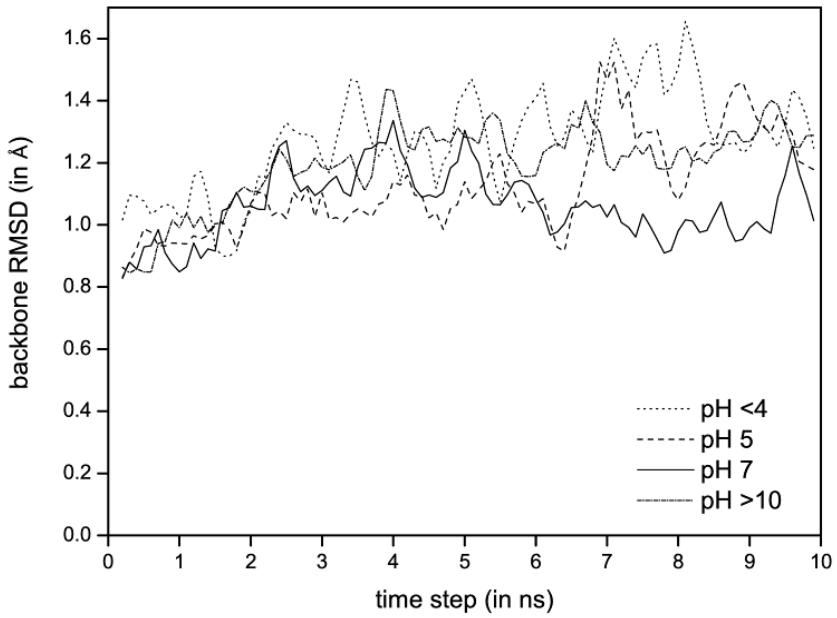


Figure 11: Backbone coordinate RMSDs in the MD simulations for lysozyme relative to the crystal structure (PDB id: 2LZT). In each of the MD simulations the protein has one of the four different protonation patterns, *pH<4*, *pH5*, *pH7* and *pH>10*, defined in Table 3. The curves are running averages over three trajectory frames with 50 ps between frames.

Table 4: Backbone RMSD values relative to the corresponding crystal structures for the four types of MD simulations averaged over the 10 ns trajectories. The types of MD simulation are defined in Table 3. The protein backbone is defined by C, CA and N atoms.

	average backbone RMSD values [in Å]			
PDB id:	**pH<4**	**pH5**	**pH7**	**pH>10**
4PTI	1.2	1.0	1.1	1.2
1PGA	1.1	0.8	0.7	1.0
1A2P	1.1	1.1	1.0	1.8
2LZT	1.3	1.1	1.1	1.2
3RN3	1.1	1.1	1.2	1.3
2RN2	1.8	1.2	1.3	2.3
1HNG	1.5	1.5	1.4	1.6
3ICB	1.7	2.4	2.7	2.3
1PPF	1.3	1.5	1.9	1.6
1ERT	0.7	0.7	1.1	1.7
1XNB	0.9	0.9	0.9	2.9
2ZTA	1.2	1.2	1.4	1.3
3BDC	0.9	0.9	0.8	1.2

Influence of variants from group I on pK_A calculation

The overall RMSD between calculated and measured pK_A values for all 194 residues of the benchmark set is 1.17 pH units for the standard procedure of Karlsberg+. The new *KB2+MD** procedure whose calculations are based on MD frames but still applies the same protocol as Karlsberg+ to obtain the required electrostatic energy terms yields a worse pK_A RMSD of 1.45. With the exclusion of the 1-2, 1-3 and 1-4 intramolecular electrostatic atom-pair interactions in the *KB2+MD*(no<=1-4)* procedure the pK_A RMSD improves significantly to 0.99, yielding a better result than Karlsberg+. Since the *KB2+MD* procedure does also exclude these interactions, it is not surprising that it yields a very similar pK_A RMSD of 0.96. Since the total pK_A RMSDs for the *KB2+MD*(no<=1-4)* and *KB2+MD* procedures are nearly the same, but the latter one has the huge advantage of providing clearly interpretable energy terms, the *KB2+MD* procedure will be analyzed in more detail in the remainder of this chapter and also used to explore the influence of variants from groups II and III.

Influence of variants from groups II and III on pK_A calculation

Table 5 lists the pK_A RMSDs between measured and computed values for the *KB2+MD* procedure with and without additionally applying the *'Rashin'*, *'ε=1'* or *'ε=4'* variants. The pK_A values of all 194 individual residues are listed in Table S4-16 of the Supporting Information of the publication. The results obtained with the new *KB2+MD* are compared with Karlsberg+ and PropKa 3.1[10, 79]. For the *KB2+MD* procedure, the structures from the MD simulations are used without post-processing. The results are listed in Table 5 under the entries *KB2+MD* and *KB2+MD-Rashin*, where the molecular surfaces are either based on the atomic radii from the CHARMM22 force field[15] or from Rashin et al.[86], respectively. In the two right columns of Table 5 the protein structures are energy-minimized before the computation of the electrostatic energies, which is performed in a homogeneous dielectric medium with dielectric constants of 1 or 4. The latter procedure yields the best results with a pK_A RMSD over all residues in the benchmark set of 0.79 as compared to 1.17 obtained with Karlsberg+.

Table 5: Root mean square deviations (RMSD) between measured and computed pK_A values (pK_A RMSD) of 194 titratable residues in 13 proteins in the benchmark set, given in pH units. The pK_A values obtained with two types of electrostatic energy computation (Karlsberg+ [KB+] and *KB2+MD*) and the empirical approach PropKa 3.1[10, 79] are compared. In addition, results for three variants applied to *KB2+MD* are shown: (i) a different set of vdW radii for electrostatic energy computations (*Rashin*) and (ii/iii) an energy minimization applied to the MD structures prior to electrostatic energy computations with dielectric constant of $\varepsilon_{min} = 1$ or $\varepsilon_{min} = 4$.

Protein					RMSD between measured and computed pK_A					
Nr.	**protein name**	**PDB id**	**Nb. of pK_A [a]**	**Nb. of atoms [b]**	**KB+**	**PropKa [c]**	***KB2+MD***	***KB2+MD Rashin***	***KB2+MD $\varepsilon_{min}=1$***	***KB2+MD $\varepsilon_{min}=4$***
1	pancrea trypsin inhibitor	4PTI	14	892	0.95	0.41	0.66	0.40	0.94	0.65
2	streptococcal protein G	1PGA	15	855	0.94	0.68	0.87	0.50	0.91	0.58
3	barnase	1A2P	12 [d]	1727	1.03	1.17	0.92	0.66	0.73	0.68
4	lysozyme	2LZT	20	1960	1.08	0.66	0.96	1.07	0.94	0.92
5	bovine ribonuclease A	3RN3	14 [e]	1856	0.77	0.79	0.79	0.58	0.77	0.61
6	ribonuclease H	2RN2	20 [f]	2455	1.67	0.78	1.04	0.60	1.11	0.84
7	rat T-lymphocyte adhesion glycoprotein	1HNG	14	1576	1.09	0.54	0.89	0.66	0.89	0.81
8	Ca binding protein	3ICB	19	1202	0.95	0.66	0.64	0.66	0.59	0.70
9	ovomucoid inhib OMTKY3	1PPF	11	418	0.84	0.54	0.65	0.60	0.63	0.65
10	thioredoxin	1ERT	17	821	1.22	0.93	1.60	1.39	1.84	1.16
11	xylanase [g]	1XNB	7	1448	1.47	0.89	1.08	1.04	0.98	1.13
12	GCN4 leucine zipper [h]	2ZTA	16/16	1120	1.22/1.13	0.46/0.51	1.12/1.19	0.71/0.78	1.07/1.12	0.80/0.85
13	staphylococ. nucl. (Δ+PHS)	3BDC	15	2101	1.65	0.83	0.61	0.93	0.60	0.50
	all residues		**194**		**1.17**	**0.74**	**0.96**	**0.81**	**0.99**	**0.79**

[a] number of pK_A values considered per protein
[b] number of protein atoms including hydrogens
[c] The proteins with the running numbers 2-6, 8, 9 and 11 were used to optimize the parameters of PropKa[10].
[d] Asp75 was not included, as in Ref.[13]. The computed pK_A value of this buried residue is very acidic, while the measured value of 3.1 is obtained under unfolding conditions and is therefore close to the value in aqueous solution.[87]
[e] His48 was not considered, as in Ref.[13], since together with Gln101 it undergoes a local but significant conformational change when titrated.
[f] Asp10 was not considered, as in Ref.[13], since it participates in a Mg^{2+} binding site.
[g] For xylanase the pK_A values are computed with the O and N atoms of Asn35 interchanged, as discussed in text.
[h] The PDB structure for the Leucine zipper is dimeric. The number of atoms refers to the whole dimer. The pK_A RMSDs of the two monomers are separated by "/". For the overall pK_A RMSD, the pK_A values from both chains are averaged.

The 'standard' *KB2+MD* procedure provides computed pK_A values whose total RSMD (0.96) from measured values is significantly smaller than that of those predicted with the Karlsberg+ (1.17). Usage of the atomic radii from Rashin et al.[86] decreases the pK_A RMSD further to 0.81. Energy minimizing the trajectory frames in the *KB2+MD* procedure with a dielectric constant of $\varepsilon_{min} = 1$ [denoted as *KB2+MD(ε_{min}=1)*] yields no decrease in the pK_A RMSD (0.99), while with $\varepsilon_{min} = 4$ [denoted as *KB2+MD(ε_{min}=4)*] the decrease (0.79) is significant. Combining the atomic radii of Rashin with energy minimization at $\varepsilon_{min} = 4$ resulted in no further improvement over using either procedure alone (0.79, not shown in Table 5). The so called Null model assumes that the pK_A values in the protein are equal to the values of the corresponding compound in aqueous solution; this model yields an overall pK_A RMSD of 0.97. Since the majority of titratable residues in proteins do in fact exhibit small pK_A shifts relative to aqueous solution, this result is expected. Even in the best procedure [*KB2+MD(ε_{min}=4)*] the pK_A RMSDs of different proteins can vary by a factor of two which is mainly influenced by outliers present in one protein but not in the other.

To investigate the influence of the length of MD simulations on the accuracy of pK_A computations, the results obtained with the *KB2⁺MD(ε_{min}=4)* method using either all 90 trajectory frames, the first (frames 10-54) or the second half (frames 55-99) of the MD simulations are compared. The resulting RMSDs between measured and computed pK_A values are 0.79, 0.88 and 0.84, respectively. This demonstrates that simulations of even less than 10 ns can be sufficient to obtain a reasonably good result for pK_A computations.

For comparison, the pK_A values of the benchmark set were also calculated with the widely used empirical pK_A prediction software PropKa 3.1[10, 79], yielding an overall pK_A RMSD of 0.74. The results obtained with PropKa are slightly better than with the *KB2⁺MD(ε_{min}=4)* procedure. It should be noted that the pK_A values from 8 of the 13 proteins in the current benchmark set are part of the training set as described in the publication of PropKa[10].

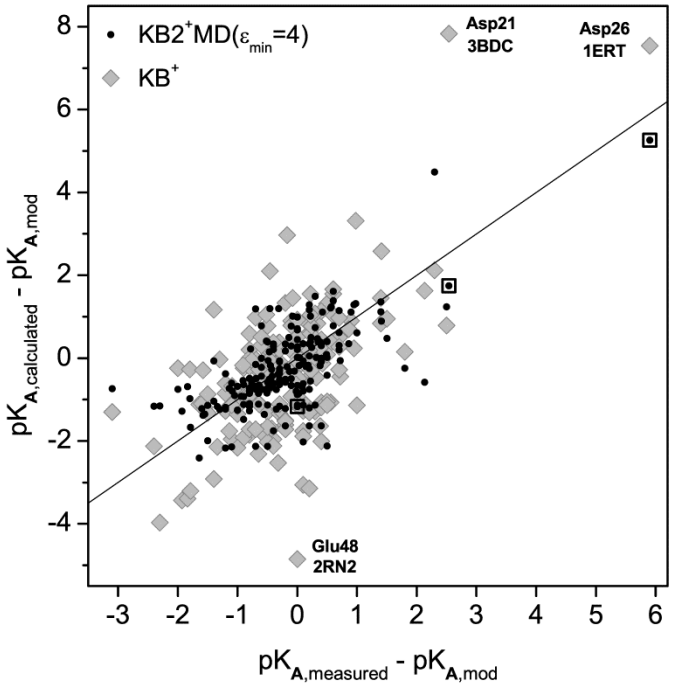


Figure 12: Correlation diagram of measured and computed pK_A shifts relative to the measured pK_A values in solution ($pK_{A,mod}$) for the 194 pK_A values of titratable residues in 13 proteins listed in Table 5. Results obtained with Karlsberg⁺ (KB^+) are shown as gray diamonds. Those computed with *KB2⁺MD(ε_{min}=4)*, (energy minimization with ε_{min} = 4, see text) are shown as black filled circles. The diagonal line has a slope of unity and marks perfect agreement between measured and computed results. The most obvious difference between the two computational results is that most of the outliers in the Karlsberg⁺ results are absent in the *KB2⁺MD(ε_{min}=4)* results. Three Karlsberg⁺ data points that are strong outliers are labeled. The corresponding values obtained with *KB2⁺MD(ε_{min}=4)* are highlighted with open squares.

Figure 12 shows a scatter plot of the differences between measured and computed shifts of pK_A values for the 194 titratable residues in the 13 different proteins obtained with either the Karlsberg$^+$ or the *KB2$^+$MD(ε_{min}=4)* procedure. A systematic improvement and removal of outliers can be observed for *KB2$^+$MD(ε_{min}=4)*. The maximum error in any residue is -2.71 (Glu41 in 1HNG) for the new procedure, while it was 5.3 (Asp21 in 3BDC) for Karlsberg$^+$. The performance of the *KB2$^+$MD* procedure for the individual residue types is shown in Table 6. The best results were obtained for lysine and the C-termini, with overall pK_A RSMDs of 0.54 and 0.55, respectively, while the worst results were for the N-termini, with an overall pK_A RMSD of 1.15. It should be noted here that only 4 pK_A values of N-termini are contained in the benchmark set.

Table 6: pK_A RMSDs between computed and measured pK_A values for specific residue types. Nter and Cter are the N- and C-terminus of the protein. The benchmark set does not contain cysteine and arginine residues, since no measured pK_A values are available for these residues. KB$^+$ is the abbreviation for Karlsberg$^+$.

		RMSD between measured and computed pK_A values					
residue type	**Nr. of residues**	**KB$^+$**	**PropKa**	***KB2$^+$MD***	***KB2$^+$MD Rashin***	***KB2$^+$MD ε_{min}=1***	***KB2$^+$MD ε_{min}=4***
Asp	55	1.36	0.74	0.98	1.07	1.08	0.75
Glu	73	1.20	0.80	1.13	0.73	1.12	0.91
Lys	32	0.86	0.53	0.59	0.53	0.54	0.54
His	11	0.76	0.98	0.90	0.77	0.87	0.84
Tyr	11	1.03	0.66	0.73	0.70	0.74	0.74
Nter	4	1.84	0.24	0.79	0.67	1.41	1.15
Cter	8	0.91	0.65	0.70	0.51	0.70	0.55

Table 7: Average differences between calculated ($pK_{A,calc}$) and model pK_A ($pK_{A,mod}$) by residue type. The model pK_A is the pK_A of the corresponding residue in aqueous solution. The table shows that the *KB2$^+$MD-Rashin* procedure tends to predict significantly smaller shifts for acids than *KB2$^+$MD(ε_{min}=4)*. On the other hand, the total RMSD values of experimental compared to calculated pK_A values (shown in Table 5) are nearly the same for the procedures *KB2$^+$MD-Rashin* and *KB2$^+$MD(ε_{min}=4)*, namely 0.81 and 0.79, respectively. KB$^+$ is the abbreviation for Karlsberg$^+$.

		$\langle pK_{A,calc} - pK_{A,mod} \rangle$						
residue type	**Nr. of residues**	**experiment**	**KB$^+$**	**PROPKA**	***KB2$^+$MD***	***KB2$^+$MD Rashin***	***KB2$^+$MD ε_{min}=1***	***KB2$^+$MD ε_{min}=4***
Asp	55	-0.56	-0.43	-0.38	-0.61	-0.29	-0.58	-0.42
Glu	73	-0.32	-0.63	-0.01	-0.81	-0.40	-0.81	-0.53
Lys	32	0.35	0.44	0.10	0.62	0.48	0.58	0.44
His	11	-0.38	-0.59	-0.65	-0.67	-0.79	-0.65	-0.79
Tyr	11	0.88	0.72	0.57	0.99	0.77	0.97	0.97
Nter	4	0.35	-1.44	0.31	-0.35	-0.24	-0.71	-0.58
Cter	8	-0.56	-0.67	-0.84	-0.74	-0.24	-0.66	-0.36

4.4. **Discussion**

The first important observation is that the straightforward use of the traditional protocol[21, 13] to calculate electrostatic energies, as implemented in the *KB2⁺MD** procedure, does not give accurate results for pK_A computations that are based on structures taken from molecular dynamic simulations. With a value of 1.45 the overall pK_A RMSD of the *KB2⁺MD** procedure performs worse than Karlsberg⁺ that yields 1.17. The main reason seems to be that the new procedure is sensitive to the contributions from 1-2, 1-3 and 1-4 electrostatic atom-pair interactions. These interactions are not considered in the CHARMM force-field, but included in the traditional approaches for electrostatic energy calculations. Their exclusion in the *KB2⁺MD*(no<=1-4)* protocol improved the pK_A RMSD to 0.99; an even better result than the one obtained with Karlsberg⁺. The PAC structures generated by Karlsberg⁺ are very similar to each other, varying only in the coordinates of hydrogen atoms and side-chain atoms of residues in salt bridges. Therefore the majority of the false energy contributions coming from the short range atom-pair interactions may cancel out. In contrast, the new procedure does not apply any constrains to atom coordinates and the false contributions may therefore not cancel out.

The *KB2⁺MD* procedure is derived from the *KB2⁺MD*(no<=1-4)* variant and yields a similar overall pK_A RMSD of 0.96. The difference is the abandonment of the concept that all energy terms (desolvation energy, interaction matrix and conformational energy) are relative values referring to a certain protonation pattern called reference protonation. In the reference protonation all residues are in a protonation state of neutral total charge; a configuration that usually does not occur in protein. Instead all energy terms do contain absolute values in the *KB2⁺MD* procedure and are therefore easily interpretable. This variation should influence the intermediate energy terms but not the final results of the pK_A calculation. But as a side effect of the way this procedure is implemented, additionally the electrostatic self energies of titratable residues are completely neglected. Since the overall pK_A RMSD does not change significantly and is actually even slightly better, this side effect seems to be acceptable. Due to the enormous advantage of better interpretable energy terms the *KB2⁺MD* procedure is chosen as the standard and the remaining variants are studied by combing them with this procedure.

Using different atomic radii

Using the set of atomic radii from Rashin at al.[86] to define the protein volume for electrostatic energy computations instead of the CHARMM22 radii improved the accuracy of the computed pK_A values lowering the pK_A RMSD from 0.96 with *KB2⁺MD* to 0.81 with *KB2⁺MD-Rashin*. This improved result is comparable to the pK_A RMSD of 0.79 obtained with the *KB2⁺MD(ε_{min}=4)* procedure. A difference can be found in the maximum error for an individual residue. The *KB2⁺MD(ε_{min}=4)* procedure yields a smaller maximum error of +2.71 pH units (for Glu41 in 1HNG) as compared to -4.61 pH units (for Asp21 in 3BDC) using the *KB2⁺MD-Rashin* procedure. The latter tends to underestimate the pK_A shifts of acids, yielding pK_A values that are closer to the pK_A value in aqueous solution as compared to the *KB2⁺MD(ε_{min}=4)* procedure as shown in

Table 7. The most significant difference in atomic radii between the CHARMM22 force field and those used by Rashin at al. is for oxygen, whose radius is, depending on the atom type, at least 0.3 Å smaller in the latter case. Hence, the atomic charges of oxygen atoms at the protein surface are closer to the solvent dielectric medium. As a consequence the pK_A values of the acidic residues (Glu, Asp and Cter) are closer to the pK_A value in aqueous solution.

In terms of the average error in pK_A values there are no significant differences between the two procedures *KB2⁺MD(ε_{min}=4)* and *KB2⁺MD-Rashin* as shown in Table 5. The CHARMM22 radii that are used by Karlsberg⁺ are optimized for MD simulations. The radii taken from Rashin et al. have been optimized to find cavities in protein structures. Neither has been optimized for electrostatic energy computations, but their use and comparison demonstrates the strong influence that differences in atomic radii can have. This suggests a re-parameterization of the atomic radii for pK_A calculations may be worthwhile. By combining both procedures [*KB2⁺MD-Rashin(ε_{min}=4)*] an overall RMSD of 0.79 pH units was obtained, which is the same value obtained for *KB2⁺MD(ε_{min}=4)*, demonstrating that the positive effect of the two variations in the procedure is not additive.

Effect of energy minimization

As shown in Table 5, minimizing the energy of the whole protein water box with a dielectric constant of $\varepsilon_{min} = 4$ before electrostatic energy computations were performed improved the agreement with the measured pK_A values significantly. In contrast, the same energy minimization procedure yielded no improvement if performed with $\varepsilon_{min} = 1$. An analysis of this dependence can be summarized as follows. Energy minimization removes primarily the influence of kinetic energy on protein structures obtained by MD simulation. Thus, it regularizes the structures. But, in the absence of kinetic energy attractive electrostatic interactions are emphasized too much. As a consequence H-bonds and salt-bridges that stabilize specific protonation patterns are strengthened. For atom pairs, where the electrostatic interactions are attractive, the Lennard-Jones (LJ) potentials that model the vdW interactions are generally repulsive. If the energy minimization is performed with weaker electrostatic interactions ($\varepsilon_{min} = 4$), the relative influence of the LJ interactions is enhanced. As a consequence H-bonds and salt-bridges are weakened, leading to a more balanced regularization of the structures.

Role of MD simulations with different protonation pattern

The results showed that the accuracy of computed pK_A values can be enhanced if MD simulations are used to include conformational variability. However, conventional MD simulations need to use a specific fixed choice of protonation pattern for all residues. An MD trajectory relaxes around this protonation pattern, and as a consequence, this pattern appears to be more stable than others not used for the MD simulation. Therefore, MD simulations with different sets of protonation patterns are needed to compensate for this bias. For the same reason different PACs are used in Karlsberg⁺. Good results could be obtained with the set of four protonation patterns

listed in Table 3 which are a combination of assigning protonation states according to a fixed scheme and according to the results of pK_A computations based on the crystal structures.

The acidic residues Asp and Glu are all protonated in MD simulations with *pH<4* and deprotonated in MD simulations with *pH7* and *pH>10*, while their protonation states are determined by pK_A computations using the crystal structures in MD simulations with *pH5*. The basic residue Lys is protonated in MD simulations with *pH<4* and *pH5* and deprotonated with *pH>10*, while the protonation state is determined by pK_A computations with *pH7*. All other residues considered to be titratable (His, Tyr, Cys, Nter, Cter) are protonated for MD simulations with *pH<4* and deprotonated for those with *pH>10*, while their protonation is determined by pK_A computations for moderate pH values of *pH5* and *pH7*. Hence, the accuracy of the final pK_A results in the current study depends not only on the appropriateness of the protonation pattern used for the MD simulations but also indirectly on the accuracy of the initial pK_A computations using the crystal structures.

This scheme for the protonation pattern was motivated by the observation that pK_A calculations with Karlsberg+, especially those that are based only on a single structure (as is the case for the sc_{ph7} protocol), tend to overestimate pK_A shifts. Especially for the residues Asp, Glu and Lys that are usually charged in the crystal structure. For example, sc_{ph7} calculations often yield a pK_A far below 4 for Glu residues in salt bridges, although their measured pK_A values are nearly not shifted at all. If on the other hand a sc_{ph7} calculation yields an elevated pK_A for a Glu residue, this is mostly caused by the residue being either located in an unpolar environment of a protein crevasse or cavity close to other negatively charged residues or in an environment that involves both. As a result the measured pK_A is usually also shifted to higher values, but not as much as the calculation suggests.

Empirical versus electrostatic energy based pK_A value computation

PropKa[10, 79] is an empirical method for predicting pK_A values in proteins using a physicochemically motivated parameterization. The method is fast, widely used and has been improved considerably over the years. In contrast to earlier versions[13], the present version of PropKa 3.1 also yields good results for protein residue pK_A values that are shifted considerably compared to aqueous solution. It was demonstrated that the new PropKa 3.1 yields significantly better agreement with measured pK_A values in proteins than methods based on electrostatic energy computations[10, 79].

Among the 194 titratable residues with measured pK_A values there are 38 residues whose side chains are buried to more than 80%, where the degree of being buried is evaluated by computing the solvent exclusion surface of the residue side chains. Among these 38 buried residues are 17 Glu, 15 Asp 5 Tyr and 1 His. Measured pK_A shifts between protein and solvent environments are for buried residues generally larger than for solvent exposed residues. The 38 buried residues exhibit an average measured pK_A shift of $<|\Delta pK_A|> = 1.13$ as compared to $<|\Delta pK_A|> = 0.54$ for the 156 less or not buried titratable residues. The pK_A RMSD is for the

KB2⁺MD(ε_{min} = 4) approach 1.08 and 0.70 for the buried and less buried residues, respectively. The corresponding values for PropKa 3.1 are 1.15 and 0.59. These numbers differ more for PropKa than with the present approach. PropKa is an empirical method with physically motivated terms that are parameterized by using known experimental data, which works obviously better, if more data are available.

With the optimized procedure of the approach presented here the quality of agreement with measured values is nearly equal to PropKa as shown in Table 5. However, the present electrostatic approach is quite expensive. Hence, for "*routine*" pK_A computations in uncomplicated or straightforward cases, the use of PropKa[10, 79] by itself is generally adequate and often preferred.

However, for "*non-routine*" cases—i.e., those that are more complex, unusual and not covered by the learning set for the empirical approach—a more detailed, physics-based approach like the current one becomes necessary. "*Non-routine"* cases involving pK_A computations include titratable residues that are in contact with redox active centers such as the two propionates bound to heme[88], or reaction centers like the Mn-cluster in photosystem II (PSII)[89]. Other *non-routine* cases are non-equilibrium scenarios, which occur for proton conduction, creation and depletion processes. Prominent examples are processes of proton creation and conduction at the Mn-cluster of PSII[90] or proton conduction and depletion in cytochrome *c* oxidase[8]. In the latter case it may be necessary to introduce an additional *FPP* with a protonation pattern that differs from the four listed in (17), for which a corresponding MD simulation is necessary to account for unconventional protonation states. Another advantage of the electrostatic energy approach over empirical methods is the option to analyze the problem in terms of energy contributions and individual protein conformations to determine which ones contribute preferentially to the computed pK_A value. Also, a particular advantage of the approach described in the current work is the possibility for *ab initio* computation of pK_A values for arbitrary organic compounds. Here, high level quantum chemistry calculations would need to be combined with the electrostatic energy computations.[91]

Future developments

The method presented here is limited to structural changes in proteins that can be generated with classical MD simulations. The results indicate that standard protein MD simulations over a 10 ns time span are sufficient for computing reliable pK_A values. If required by the particular problem, the protocol can easily be changed using longer or even alternative MD simulations with different protonation states and even different MD simulation techniques. However, a requirement is that the MD simulations for different protonation patterns of the protein generate thermodynamically correct ensembles.

The proteins used to test the new method do not form complexes with ligands, ions or other proteins. Some of the structures have binding pockets for ions (e.g. staphylococcal nuclease binds calcium[51]) but the related pK_A values have been measured in their absence. However, with

a straightforward extension of the present procedure it would also be possible to compute pK_A values in proteins involving bound ligands, ions or other cofactors, although computing pH-dependent binding affinities may introduce additional complications.[92]

The difficulty of calculating pK_A values with electrostatic energy calculations in staphylococcal nuclease (SNase) mutants has been discussed extensively in the first project of this thesis in chapter 3. In the study presented in this chapter these proteins have not been considered. The KB+2MD procedure is substantially different from the procedure applied by Karlsberg+ and it turned out that several aspects of the traditional approach to electrostatic energy calculations had to be changed and optimized. To study and optimize the new procedure it seemed to be adequate to start with a benchmark set that contains more "standard" proteins. The optimized new procedure was applied to the SNase benchmark set separately and the results will be discussed in the next chapter "*Karlsberg2+ MD and the SNase Benchmark Set*".

4.5. **Summary & Conclusions**

For the study presented in this chapter, structures from MD simulations with different specific protonation patterns are used to compute pK_A values of titratable groups in proteins using electrostatic energies obtained by solving the Poisson Boltzmann equation. The new procedure is extending the concept of the software Karlsberg+, which essentially uses protein crystal structure information combined only with local structural variation of the side chains of titratable residues. The performance of the new approach is evaluated for a set of 13 standard proteins, most of which had been used in the past already to optimize Karlsberg+. The results reveal that it is not sufficient to simply apply the traditional protocol of Karlsberg+ to MD structures to obtain a good accuracy for pK_A calculations. Several variants of the protocol are explored, of which the most fundamental one is the exclusion of the 1-2, 1-3 and 1-4 electrostatic atom-pair interactions. With the most optimized approach the agreement between computed and measured pK_A values show to be significantly improved compared to Karlsberg+. Performing energy minimization of the individual MD structures including the whole water box with an increased dielectric constant of $\varepsilon_{min} = 4$ prior to electrostatic energy evaluation yields the best results, with an pK_A RMSD of 0.79 pH units over the entire set of 194 measured pK_A values in the 13 proteins. The reliability in terms of the maximum deviation between computed and measured pK_A values is improved, since the new method produces less outliers. It is also demonstrated that with an alternative set of atomic radii, which differs from the vdW radii in the CHARMM22[15] force field used by Karlsberg+, the pK_A RMSD value can be lowered to a value that is comparable to the value obtained by energy minimizing the structures prior to electrostatic energy computation. However, combining the use of the different atomic radii with energy minimization did not yield additional improvement. This indicates that a careful re-parameterization of atomic partial charges and radii may improve the agreement between computed and measured pK_A values in proteins, which will be the subject of future work.

The present procedure is considerably more CPU time-intensive, since four MD simulations with a combined length of 40 ns and 720 electrostatic energy computations (4x90 TAPBS[13] + 4x90 APBS[18]) are required for each protein. Sampling protein conformations is performed with the well established MD simulation technique, which is a huge advantage over elaborate modeling methods. To compute thermodynamically averaged pK_A values from the different MD simulations, the Boltzmann factors corresponding to the total electrostatic energies of protein conformations and protonation patterns of the individual MD trajectories are used. These Boltzmann factors can also be used to identify and analyze the protein conformations that prevail at specific pH values.

The current work also introduces an alternative protocol for electrostatic energy calculations that does not influence the overall accuracy of the procedure but offers the huge advantage that the obtained energies for the pair interactions of titratable residues, the desolvation energies and the conformational energy of the protein structure become easily interpretable.

5. **Karlsberg2⁺ MD and the SNase Benchmark Set**

5.1. **Introduction**

In the previous chapter a new procedure was introduced that determines pK_A values with electrostatic energy calculations based on structures obtain with MD simulation. The procedure was named *KB2⁺MD*, an abbreviation for *Karlsberg2⁺ MD*. The set of 13 proteins used to optimize and benchmark the procedure is very similar to the one that had been used in the past[13] to evaluate the performance of Karlsberg⁺. Most of the 194 titratable residues in the benchmark set are "standard" residues, in that their pK_A values are not shifted much compared to their pK_A values in aqueous solution. With the naive assumption that the pK_A values of these residues are not shifted at all (the so called Null model), a quite good pK_A RMSD of 0.97 can be obtained. Additionally, proteins in the benchmark set are mostly reacting with only minor structural changes to the protonation or deprotonation of titratable residues. Already the rather conservative modeling approach of Karlsberg⁺, that modifies only hydrogen atoms and side-chain atoms of residues in salt bridges, yields a good pK_A RMSD of 1.17.

In this regard the 'standard' benchmark set used in the previous chapter differs significantly from the staphylococcal nuclease (SNase) benchmark set[38] introduced and discussed in chapter 3. Proteins in the SNase benchmark set are variants of the SNase protein, where charged residues had been introduced by point mutation in the hydrophobic interior of the protein, resulting in titratable residues whose measured pK_A values are shifted strongly. Karlsberg⁺ failed to predict these pK_A shifts correctly. Two main problems for this failure have been identified[30]: (i) the presence of water molecules buried in cavities that are not considered correctly in the electrostatic energy calculations and (ii) significant structural changes as a consequence of a titratable residue becoming charged. For the first problem a solution has been proposed and evaluated in chapter 3. For the second problem it became evident that a novel approach is needed, that would be capable of modeling larger structural changes.

In this chapter the new *KB2⁺MD* procedure is applied to the SNase benchmark set. By employing unconstrained MD simulations, the procedure has the potential to account for the extensive structural changes that are expected for the SNase variants. Furthermore, the cavity-algorithm discussed in chapter 3 is combined with *KB2⁺MD* to correctly account for water molecules in internal protein cavities. The cavity-algorithm has the drawback that it cannot distinguish between water filled and empty cavities. This distinction, however, can be made by analyzing the MDs used by *KB2⁺MD* to sample protein conformations, since they contain explicit water. The final variant tested in this study takes advantage of this information and discards the empty cavities found by the cavity-algorithm considering only the cavities filled with water in the electrostatic energy calculations.

5.2. **Materials and Methods**

The SNase variants

The staphylococcal nuclease variants have been introduced and discussed in chapter 3.2. In this study a smaller subset of SNase variants is used, since the new procedures employ MD simulations for the pK_A calculations and are therefore significantly more CPU-time demanding. Of the 20 SNase variants used in chapter 3, 12 variants have been selected. These 12 variants cover point mutations at 9 different positions in the SNase protein (see Table 9 and Figure 6).

The SNase mutants are based on the two variants PHS and Δ+PHS.[44, 35] The PHS structures do contain six residues (44–49) in a flexible loop and two point mutations (F50G and N51V) that are not present in the Δ+PHS variant.[38] All SNase variant structures are prepared in the same way, except that for PHS residues 44-49 had to be added by modeling, since they were disordered in the crystal structure and therefore their coordinates were missing. Missing residues at the N- and C-terminus of the protein sequence were not added by modeling; instead the termini of the crystal structures were neutralized with an acetylated N-terminus and a methylated C-terminus to avoid the artifactual introduction of charged residues.

Karlsberg$^+$ combined with the cavity-algorithm

The performance of Karlsberg$^+$ with and without the cavity-finder modification has been discussed in detail in chapter 3. To allow a direct comparison some of these results (taken from Table 2) are also shown here. It should be noted that the protocol used for Karlsberg+ in chapter 3 (sc_{pH7}) differs from the standard protocol (sb, H_{flex}). This choice was made, since none of the titratable residues that are considered here is part of a salt bridge. Details of the protocols can be found in chapter 2.4.

Choice of the KB2$^+$MD procedure

The *KB2$^+$MD* procedure and its variants have been discussed in detail in chapter 4.2 and are applied in the same way here, with one exception. The protocol for the MD simulations is adjusted as follows. Since larger structural changes are expected for the SNase variants the total simulation time is increased from 10 to 20 ns. Structures are written out every 200 ps instead of every 100 ps, and the first 2 ns are ignored instead of just 1 ns. As a result, there are still 90 structures per MD that are used for the electrostatic energy calculations, but the time step in between them is doubled.

Since the calculations are computationally very expensive it is not reasonable to study all variants again with the new benchmark set. Instead, following the conclusion from chapter 4, the *KB2$^+$MD(ε_{min}=4)* procedure was chosen for this study. This procedure turned out to be the most successful one for a standard benchmark set, yielding the highest accuracy in terms of pK_A RMSDs.

KB2⁺MD combined with the cavity-algorithm

To consider the cavity volumes found by the cavity-finder (see chapter 3.2) two types of electrostatic energy calculations have to be modified in the *KB2⁺MD(ε_{min}=4)* procedure. First, desolvation energies of titratable residues and the residue-pair interaction energies have to be calculated with the software TAPBS. This step needs no further adjustment and is done exactly as described in chapter 3.2. The second type of energy calculation, the conformational energy, was not required in chapter 3. The conformational energy consists of the two parts defined in equation (32): (i) the desolvation energy ΔG_{PB} and (ii) the absolute electrostatic self energy $G_{Coulomb}$ of the protein. The latter one is not influenced by cavities, since the calculation is performed in a homogeneous dielectric continuum. The desolvation energy ΔG_{PB} is calculated with the software APBS and cavity volumes are incorporated in a similar way as done for TAPBS. As a first step, the cavity-algorithm searches for cavity volumes. The obtained cavity volumes are then used to update the ε-grid of APBS, in that every ε-grid point that is located in a cavity volume is set to $\varepsilon_{cav} = 80$. This combination of *KB2⁺MD(ε_{min}=4)* and cavity-algorithm is named *KB2⁺MD(ε_{min}=4, cav)*.

As it was done in the study of chapter 3, for the *KB2⁺MD(ε_{min}=4,cav)* procedure described so far all cavities found by the cavity-algorithm are considered to be filled with water. Since the *KB2⁺MD* procedures use MDs with explicit waters to sample protein conformations, this coarse assumption is not required any more. Instead the water molecules found in the protein structures obtained from the MDs can be used to decide whether a cavity contains water or not. This possibility is explored with a further variant named *KB2⁺MD(ε_{min}=4,cav*)*. Usually, all water molecules are removed from the structure prior to any electrostatic energy calculation. For the *KB2⁺MD(ε_{min}=4,cav*)* variant, water molecules (MD-waters) and protein structure are split and stored separately. The protein structure (without any water) is used for the actual energy calculations and the MD-waters are provided to the cavity-algorithm. Since the MD simulations were performed with periodic boundary conditions, in some cases the protein leaves the water box and therefore is only partially surrounded by water. As a result, the water molecules have to be translated in a way that the geometric center of the protein becomes the center of the water box before MD-waters are extracted. This was done with the PBCTools plugin of the software VMD.[29] The protocol of the cavity-algorithm itself was extended as follows for the *KB2⁺MD(ε_{min}=4,cav*)* procedure. First, as it was done before, the cavity volumes are determined and the result is stored on a discrete high resolution grid (cavity grid) were every grid point is marked as either belonging to a cavity (*cavity-grid-point*) or not (*not-cavity-grid-point*). In a second step the cavity grid is filtered using the coordinates of the MD-water atoms. Only *cavity-grid-points* that are located within 2 Å of any MD-water oxygen atom are kept. All other grid points become *not-cavity-grid-points*. As a result only cavity volumes that are close to an MD-water remain in the calculations. The generous cutoff of 2 Å was chosen to account for the high mobility of water molecules. As a result, cavity volumes that do not exactly overlap with the

vdW-volume of a water molecule but are located close to water are still considered to be water filled.

5.3. **Results**

The average backbone RMSDs for all MD simulation data are shown in Table 8. The maximum RMSD of an individual frame is given in brackets. Three trajectories marked with a star in Table 8 show an diverging RMSD, also all of them remain stable for at least some time. For the MD simulations of 3EVQ/*pH>10,* 3D6C/*pH<4* and 2OXP/*pH>10* the corresponding RMSD starts to increases continuously over time till the end of the simulation, after remaining stable for 10, 15 and 16 ns respectively.

The detailed results of the pK_A calculations are shown in Table 9. The overall pK_A RMSD is 8.94 for Karlsberg+ and decreases dramatically to 4.85, if cavities are considered with a cavity parameter of c = 0.7. The *KB2+MD(ε_{min}=4)* procedure that employs MD structures for the electrostatic energy calculations yields an even better pK_A RMSD of 3.84, although it does not account for cavities directly.A further improvement was achieved by combining the cavity-algorithm and the *KB2+MD(ε_{min}=4)* procedure. Astonishingly, the variants of this procedure that use different cavity parameters do perform very similar, although their accuracy differs for individual variants. The decrease of the cavity parameter for the *KB2+MD(ε_{min}=4,cav)* procedure from c = 0.9 to 0.7 improved the overall pK_A RMSD slightly from 2.53 to 2.33.

Filtering of the cavity volumes found with a cavity parameter of c = 0.7 with MD-waters in the *KB2+MD(ε_{min}=4,cav*)* procedure yields a slightly larger pK_A RMSD of 2.58 compared to the one obtained with the *KB2+MD(ε_{min}=4,cav)* procedure (2.33). The pK_A values of most SNase variants are not influenced by the filtering procedure, except for the two residues at position 25 and the Lys62 variant. The latter residue is the reason for the increased overall pK_A RMSD of *KB2+MD(ε_{min}=4,cav*)*, yielding a pK_A that deviates from the experimental value by 5.15 pH units. The total cavity volumes found by the cavity-algorithm with and without filtering based on MD-waters is compared in Figure 13 for the example of the *pH5* MD of the Lys62 variant. On average half of the cavity volumes are removed by the filtering process. The most accurate result was obtained with the empirical prediction scheme PROPKA 3.1, with a total pK_A RMSD of 1.24.

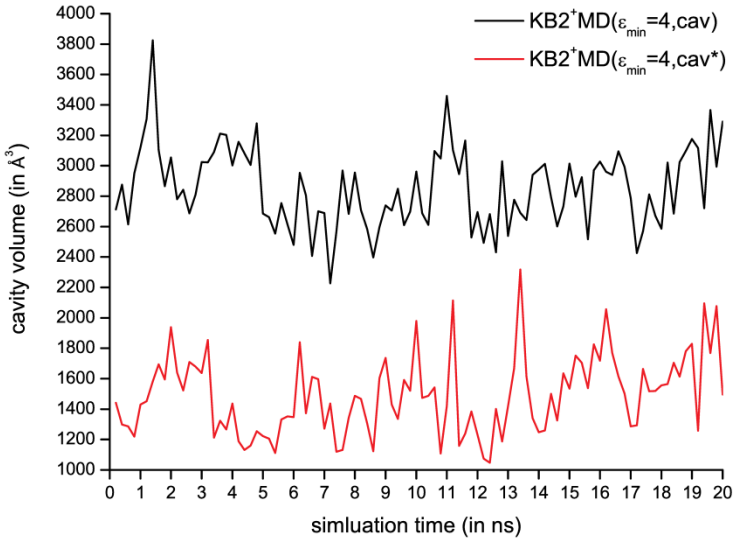


Figure 13: Total cavity volumes found by the cavity-algorithm for the *pH5* MD of the Lys62 variant with a cavity parameter of c = 0.7. The step size in between two frames is 200 ps and residue Lys62 was protonated in the MD. The cavity volumes for two procedures are compared: The *KB2+MD(εmin=4,cav)* procedure (black line) selects cavities purely based on geometric criteria, while the *KB2+MD(εmin=4,cav*)* procedure (red line) filters the cavity volumes using water molecules found in the MD structures.

Table 8: Backbone RMSD values relative to the corresponding crystal structures for the four types of MD simulations averaged over the 20 ns trajectories. In brackets the maximum RMSD for an individual frames is given. The types of MD simulation are defined in Table 3. The protein backbone is defined by C, CA and N atoms. Trajectories that diverged, i.e. where the RMSD increased over time, are marked with a star.

	average backbone RMSD values [Å]			
PDB id:	**pH<4**	**pH5**	**pH7**	**pH>10**
3EVQ	1.0 (1.4)	1.0 (1.4)	1.0 (1.3)	1.9 (2.9)*
3ERQ	1.2 (1,6)	1.1 (1.4)	1.0 (1.3)	1.5 (2.2)
3EJI	1.1 (1.6)	0.9 (1.1)	1.0 (1.3)	1.4 (1.8)
3D6C	1.1 (1.8)*	1.7 (2.0)	1.4 (1.9)	1.6 (2.0)
3DMU	2.0 (2.7)	2.8 (4.0)	1.7 (2.8)	2.2 (3.0)
2OXP	1.0 (1.7)	1.6 (2.4)	1.5 (2.3)	2.1 (3.5)*
3ERO	0.9 (1.4)	0.9 (1.2)	0.9 (1.1)	1.6 (2.1)
2RBM	1.0 (1.3)	1.2 (1.5)	1.2 (1.6)	1.9 (2.5)
3C1F	1.1 (1.4)	1.3 (1.6)	0.9 (1.1)	1.4 (1.6)
3C1E	1.2 (1.4)	1.2 (1.6)	1.1 (1.6)	1.5 (1.9)

Table 9: Deviations between measured ($pK_{A,exp}$) and computed ($pK_{A,calc}$) pK_A values of 12 SNase variants, given in pH units. The pK_A values obtained with the original Karlsberg$^+$ are compared with the values obtained with the modified version *Karlsberg$^+$(cav)* that includes the cavity-algorithm, the *KB2$^+$MD(ε_{min}=4)*, *KB2$^+$MD(ε_{min}=4,cav)* and *KB2$^+$MD(ε_{min}=4,cav*)* procedures as well as with the empirical approach PropKa 3.1. For the *KB2$^+$MD(ε_{min}=4,cav)* procedure the results for two values of the cavity parameter (c = 0.7 and 0.9) are shown, for the *KB2$^+$MD(ε_{min}=4,cav*)* procedure a cavity parameter of c = 0.9 was used. The last row shows the overall root mean square deviations between measured and computed pK_A values (pK_A RMSD) for all 12 residues.

SNase variant				**$pK_{A,calc} - pK_{A,exp}$**						
PDB id	**mutation / SNase type**	**residue**	**exp. pK_A**	**Karlsberg$^+$**	**PropKa**	**Karlsberg$^+$ cav c = 0.7**	***KB2$^+$MD ε_{min}=4***	***KB2$^+$MD ε_{min}=4,cav c = 0.9***	***KB2$^+$MD ε_{min}=4,cav c = 0.7***	***KB2$^+$MD ε_{min}=4,cav* c = 0.7***
3EVQ	L25E/Δ+PHS	Glu25[c]	7.50	15.40	0.36	1.70	-1.72	-1.15	-2.95	-1.55
3ERQ	L25K/Δ+PHS	Lys25[c]	6.30	-4.90	0.38	-3.60	-6.18	-2.49	-1.05	-2.53
3EJI	L36K/Δ+PHS	Lys36	7.20	-1.30	0.32	1.10	-1.42	1.87	1.62	1.05
3D6C	L38E/PHS	Glu38	7.20	12.00	-1.12	-4.10	7.05	3.55	4.08	3.55
3DMU	T62K/PHS	Lys62[c]	8.10	-2.60	-0.97	-0.40	6.65	5.16	2.65[#]	5.15[#]
2OXP	V66D/PHS	Asp66[c]	8.70	14.80	-1.83	-1.80	-0.39	-2.22	-2.80	-2.83
3ERO	I72E/Δ+PHS	Glu72	7.30	5.90	-2.41	-1.50	-1.53	-1.53	-1.89	-1.53
2RBM	I72K/Δ+PHS	Lys72	8.60	13.70	1.38	5.30	2.05	1.56	1.59	1.53
3C1F	V104K/Δ+PHS	Lys104	7.70	-1.10	0.76	1.00	0.09	0.05	-0.94	-0.90
3C1E	L125K/PHS	Lys125	6.20	-9.90	1.20	-12.90	2.00	2.05	1.82	2.07
total pK_A RMSD				**8.94**	**1.25**	**4.85**	**3.84[+]**	**2.53[+]**	**2.33[+]**	**2.58[+]**

[c]: These residues are located near the large central cavity of the SNase protein (cluster of cavities 1-3 in Figure 9A).

[#]: These two pK_A values for Lys62 change significantly if a longer equilibration time of 6 ns is used, yielding values of 0.19 and 2.65 for *KB2+MD(ε_{min}=4,cav)*/c=0.7 and *KB2+MD(ε_{min}=4,cav*)*/c=0.7, respectively.

[+]: With an increased equilibration time of 6 ns, the total pK_A RMSDs change to 3.66, 2.56, 2.24 and 2.27 respectively.

5.4. **Discussion**

Molecular dynamic simulations

Unlike for the structures used in chapter 4, all SNase variants have one titratable residue located in their hydrophobic interior. It is therefore not surprising that, as shown in Table 8, the average structural RMSDs of most MDs are larger than the RMSDs of the Δ+PHS protein without any point mutation (protein with PDB id 3BDC in Table 4). Overall, the RMSDs are still moderate and 7 of the 10 MDs remain stable (i.e. the RMSD reaches a plateau value), within the 20 ns simulation time. The MD simulations for 3EVQ/*pH>10,* 3D6C/*pH<4* and 2OXP/*pH>10* start to diverge after a certain time, but all of them remain stable for at least half of the simulation time.

Influence of cavities and conformational changes

In chapter 3 and 4 of this work, two very different approaches have been discussed, each addressing a specific weakness of the Karlsberg$^+$ protocol. The first approach, the *Karlsberg$^+$(cav)* method, extends Karlsberg$^+$ by accounting more carefully for cavities in the electrostatic energy calculations. The second approach, the *KB2$^+$MD(ε_{min}=4)* method, employs MD simulations to sample protein structures to overcome the limitations of the modeling protocol of Karlsberg$^+$. The performance of both procedures is directly compared with one benchmark set. As it was shown, both procedures, *Karlsberg$^+$(cav)* with c = 0.7 and *KB2$^+$MD(ε_{min}=4)*, do improve the accuracy of the pK_A calculation enormously yielding a pK_A RMSDs of 4.85 and 3.84 respectively compared to the value of 8.94 obtained with the standard Karlsberg$^+$ protocol. For the five variants Glu25, Lys36 Asp66, Glu72 and Lys104 both procedures yield similar results with less than 2 pK_A units of difference. For the three variants Lys25, Lys38 and Lys62 *Karlsberg$^+$(cav)* performs better, while for the two residues Lys72 and Lys125 *KB2$^+$MD(ε_{min}=4)* yields more accurate pK_A values. *KB2$^+$MD(ε_{min}=4)* is overall more accurate and gives also a lower maximum deviation between measured and calculated pK_A values, which is 7.05 for Glu38 compared to 12.09 for Lys125 with *Karlsberg$^+$(cav)*. On the other hand, the latter procedure has the advantage that it is significantly less expensive in terms of CPU time. The result does clearly show that both aspects, the correct consideration of cavities and an extensive sampling of the conformational changes, are important for pK_A calculations based on electrostatic energy calculations.

Combining the cavity-algorithm with MD simulations

As both approaches are helpful in specific aspects, they were combined in the *KB2$^+$MD(ε_{min}=4,cav)* procedure. Again, the results improve significantly yielding an overall pK_A RMSD of 2.53 and 2.33 for cavity parameters of c = 0.9 and 0.7, respectively. Interestingly, unlike for *Karlsberg$^+$(cav)*, the choice of the cavity parameter turns out to have less effect on the results. Even a more conservative parameter of c = 0.9 yields good results, which improve only slightly by choosing a smaller cavity parameter of c = 0.7. Only the two residues Lys25 and Lys62 benefit from a smaller cavity parameter, in that their pK_A improves by more that 0.5 pH units. For the four residues Glu25, Glu38, Asp66 and Lys125 the larger cavity parameter yields a better result.

This result is consistent with the observation made in chapter 3, that a cavity parameter of c = 0.9 is enough to reliably detect all crystal waters. The insensitivity of the *KB2⁺MD(ε_{min}=4,cav)* procedure to the cavity parameter is on the other hand surprising, since its choice had a significant and mostly positive effect on *Karlsberg⁺(cav)* as discussed in chapter 3. This behavior can be explained by looking at the effect that the presence of a cavity nearby a titratable residue has on the pK_A calculations. *Karlsberg⁺(cav)* uses only a single protein structure, namely one PAC for pH 7. If no other titratable residue is located nearby, only the desolvation energy of the titratable residue is influenced by the high dielectric constant of the cavity volume. For the *KB2⁺MD(ε_{min}=4,cav)* procedure, there is an additional effect, since the procedure combines the results from several structures. Beside the desolvation energy of the individual titratable residues, also the conformational energies of the whole structure are influenced by the cavities. A smaller cavity parameter may have a positive influence on the desolvation energies, but the opposite effect on the conformational energy. On should keep in mind that Karlsberg⁺ generally tends to overestimate pK_A shifts and, since cavity volumes are modeled with a dielectric continuum with ε = 80, the presence of a cavity usually reduces the pK_A shift. This effect can be seen very clearly in Table 2 and Figure 10A. For *Karlsberg⁺(cav)* a smaller cavity parameter may be required to compensate for the inability of the procedure to sample conformational changes of the protein. The *KB2⁺MD(ε_{min}=4,cav)* procedure is capable of sampling even larger structural changes to a certain degree and this kind of compensation may therefore not be required any more.

The main weakness of the cavity-algorithm is its inability to discriminate between empty and water filled cavities. All found cavities are considered to be filled with water. The *KB2⁺MD(ε_{min}=4,cav*)* procedure overcomes this drawback by using MD-waters to filter out empty cavities. Astonishingly, the pK_A values calculated with *KB2⁺MD(ε_{min}=4,cav*)* do not differ very much from the values obtained with *KB2⁺MD(ε_{min}=4)* and both procedures yield similar overall pK_A RMSDs of 2.33 and 2.58 respectively. The residues that react most sensitive to the changed cavity volumes, yielding pK_A values that differ by more than one pH unit, are those at position 25 and 62. While the pK_A of Glu25 improves, the pK_A of Lys25 behaves in the opposite way. The main contribution to the larger pK_A RMSD of the *KB2⁺MD(ε_{min}=4,cav*)* procedure comes from the Lys62 variant whose calculated pK_A deviates from the experimental value by 5.15 pH units. Interestingly, this shift is cut to half if the equilibration time of the MDs is increased from 2 ns to 6 ns, yielding a pK_A of 10.75 that deviates by 2.65 pH units from the experimental value. The effect of an increased equilibration time for the MDs is discussed in detail the next section. Overall it can be concluded, that the filtering of the cavity volumes with MD waters has little influence on the accuracy of the procedure.

Influence of the equilibration time on the pK_A calculations

As discussed already, for the MDs of those SNase variants where charged residues are located in the hydrophobic interior of the protein, larger structural changes can occur. Therefore the protein may need more time to equilibrate than the 2 ns used here. To test this influence all

KB⁺MD calculations shown in Table 9 were repeated with an increased equilibration time of 6 ns. For all MDs the first 30 structures (instead of 10) extracted from the MDs were neglected and only the last 70 structures were used for the calculation. Significant changes could be observed only for the two SNase variants Lys62 and Glu25. For Glu25 the pK_A improved by about 0.5 pH units for all procedures (0.49, 0.54, 0.41 and 0.64 for the KB2⁺MD procedures as they appear in Table 9). For Lys62 the pK_A does improve by 0.46 for *KB2⁺MD(ε_{min}=4)*, remains unchanged for *KB2⁺MD(ε_{min}=4,cav)* with c = 0.9, but improves by 2.46 and 2.5 pH units for *KB2⁺MD(ε_{min}=4,cav)* and *KB2⁺MD(ε_{min}=4,cav*)* with c = 0.7, respectively. As a result, the overall pK_A RMSD for the two *KB2⁺MD* procedures that use a cavity parameter of 0.7 becomes nearly identical with values of 2.24 (cav) and 2.27 (cav*), respectively.

The Lys62 variant

Of all SNase variants in the benchmark set Lys62 stands out for two reasons. It is the variant with the highest structural RMSDs as shown in Table 8 and the variant that reacts most sensitive to the choice of the cavity-parameter and the cavity filtering based on MD-waters used in the cav* procedure. Therefore, it is discussed in detail.

The two important MD simulations for the pK_A of Lys62 are *pH5* and *pH7*, since they consider the protonated and deprotonated Lys62 respectively. For the *pH7* MD the charge neutral Lys62 remains in a conformation that is similar to the one found in the crystal structure for the whole time of 20 ns. It is completely surrounded by unpolar residues and does not form any hydrogen bonds. No water molecules are interacting with the lysine. The charged Lys62 in the *pH5* bends in a way that it can form hydrogen bonds with three oxygen atoms of the backbone amides of the residues Ile18, Gly20 and Thr22. Astonishingly, Lys62 is again not interacting with any water molecule. Some bulk water molecules are located close to Lys62 in a distance of about 3-4 Å, but in none of the MD structures a Lys-water hydrogen bond could be observed. The three hydrogen bonds of Lys62 are already present in the first structure taken after 200 ps, nevertheless the local environment may require some time to completely adopt to the new protonation of Lys62, explaining the need for a longer equilibration time.

Empirical based pK_A computation

The best results could be obtained with the empirical prediction scheme PropKa 3.1, yielding a very good overall pK_A RMSD of 1.25. It should be noted here that the calculations performed here are no blind predictions and the software PropKa has been optimized for the SNase benchmark set as indicated by the authors.[10] The same is valid for the cavity-algorithm implemented in the *Karlsberg⁺(cav)* procedure. For a fair comparison of the prediction performance of both procedures, a new benchmark set offering again the chance for a real blind prediction would be useful.

5.5. **Summary & Conclusions**

The accuracy of the two new procedures *KB2⁺MD(ε_{min}=4)* and *Karlsberg⁺(cav)* introduced in chapter 3 and 4 are compared directly on a challenging benchmark set that consists of 10 SNase variants. Both procedures improve the accuracy of pK_A calculation enormously compared to Karlsberg⁺. *KB2⁺MD(ε_{min}=4)* yields a better overall pK_A RMSD, but the performance for individual SNase variants differs vastly. This result clearly shows that each of the new procedures addresses a different drawback of the Karlsberg⁺ procedure. Consequently the best result could be obtained by combining both procedures into one [*KB2⁺MD(ε_{min}=4,cav)]*, yielding a more accurate result for all individual SNase variants.

The *KB2⁺MD(ε_{min}=4,cav)* procedure turns out to be very robust and further optimizations like tuning the cavity parameter or filtering cavities with MD-waters did not change the overall performance. The simulation scheme used here is exactly the same as the one used in chapter 4, except for the longer simulation time. The equilibration time of 2 ns was sufficient for all but one SNase variant.

The *KB2⁺MD(ε_{min}=4,cav)* procedure did not reach the accuracy of the empirical prediction scheme PropKa 3.1 or the accuracy obtained for the standard benchmark set. It should be noted, however, that the procedures presented here have not been optimized specifically for this benchmark set, with the only exception being the cavity parameter. The two procedures that are newly introduced in this chapter are straightforward approaches to combine the already existing procedures. The case of Lys62 shows an example where further optimization may be needed. It may be beneficial to find criteria to decide for each protein and protonation pattern individually how long the structure needs to be equilibrated and simulated. Further optimization could also be achieved by adjusting the force field parameters, as demonstrated for the vdW radii in chapter 4.

6. A Histidine Residue of the Influenza Virus Hemagglutinin Controls the pH Dependence of the Conformational Change Mediating Membrane Fusion

In this chapter a brief summary of a collaboration work is given that lead to the following publication:

Mair, C. M.; Meyer, T.; Schneider, K.; Huang, Q.; Veit, M.; Herrmann, A., A Histidine Residue of the Influenza Virus Hemagglutinin Controls the pH Dependence of the Conformational Change Mediating Membrane Fusion. *J. Virol.* **2014,** 88, 13189-13200.
DOI: 10.1128/JVI.01704-14

The major part of the project has been performed by Caroline Mair. She also prepared the manuscript. My contribution to the project was the following:

- Assistance in planning and interpretation of experiments.
- Computational modeling of mutated proteins.
- Analysis of the crystallographic and modeled protein structures.
- Conduction of pK_A calculations.

Summary

The subject of this study was the influenza virus that causes an infectious disease commonly known as the flu. In particular the H5N1 strain of the virus was studied, that caused massive bird die-offs associated with sporadic spillover infections to humans and other mammals in the 2003-2004 outbreak.[93, 94] The mechanism employed by the influenza virus to release their genome into the cell interior can be summarized as follows.[95, 7, 96] First the virus envelope protein hemagglutinin (HA) binds to sialic acid sugars on the surface of an epithelial cells. Subsequently the target cell takes up the virus by endocytosis. After uptake the virus fuses with the endosomal membrane, mediated by a conformational change of the HA protein. The fusion is triggered in the acidic milieu of late endosomes.

The HA protein (see Figure 14) is a homo trimer with each monomer being composed of the two covalently linked subunits HA1 and HA2.[97, 98] The majority of the HA1 subunit forms the globular domain in the top part of Figure 14, while the main part of the HA2 subunit consists of the two parallel alpha helices in the lower 2/3 of the image. The lower part of the protein is attached to the virus envelope and sialic acid binding pockets are located on the top part of the HA1 domain. To fuse with the host cell the complex formed by the three HA1 domains has to dissociate.[99, 100] Subsequently the shorter of the two alpha helices of the HA2 subunits can move upwards to

prolongate the longer helix. The loop connecting the two helices is called B-loop and marked in red in the middle part of Figure 14. This conformational change allows the fusion peptide, also marked in red, to attach to the endosomal membrane.[98, 101, 102, 7, 96] The next step is an extensive refolding of the whole protein facilitating the fusion of the virus envelope with the endosomal membrane.

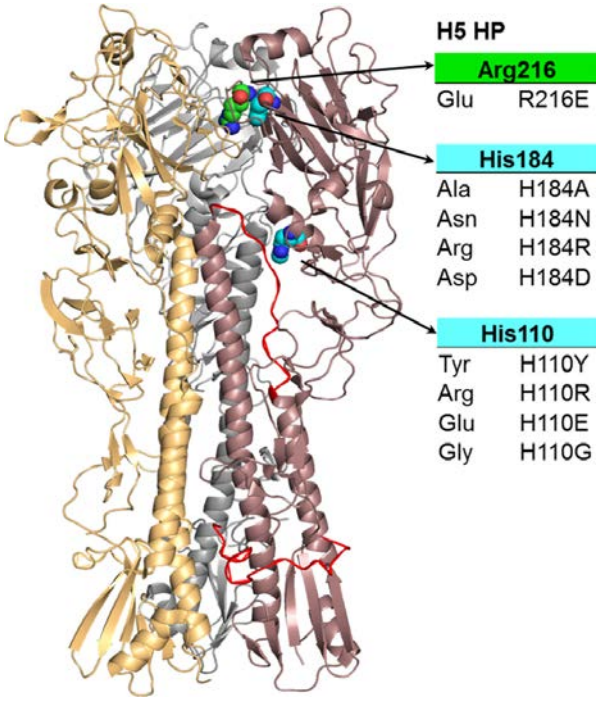


Figure 14: Crystal structure of the highly pathogenic H5 HA in cartoon representation (PDB id 2IBX). Monomers are marked in brown, orange, and gray. Histidines at positions 184 and 110 (cyan) as well as arginine at position 216 (green) are depicted as spheres in the brown monomer. The tables list the selected substitutions for these residues. The two red lines mark the fusion peptide (lower part of the image) and the B-loop (middle part of the image).

While the essential steps of the fusion mediating conformational change of the HA protein are well understood, the mechanism that initially triggers this process is essentially unknown and the subject of this study. The fusion is triggered by the decrease of the pH value in the endosome to a value of about pH 6. Since histidines have pK_A that match the pH required for activation, it has been proposed that they play a key role in the pH-sensing mechanism. Especially four highly conserved histidines were suggested[6] to play such a role. Two of them were analyzed in this study. These are the histidines His184 in subunit HA1 and His110 in subunit HA2 (see Figure 14). While His110 is located in the HA1-HA2 interface, His184 is close to the HA1-HA1 interface. Both are located in positions were they could potentially facilitate the dissociation of the HA1 monomers. To test their potential role as pH-dependent triggers of the conformational change, both have been replaced with two sets of different amino acids, listed in the Figure 14. Consequences of the mutations were studied with two experimental methods. The pH-dependence of the conformational change was analyzed with conformation specific antibodies, while the pH-dependence of fusion was studied with human red blood cells. The structural

consequences of mutations were rationalized by computational modeling. Furthermore, the residue 216 in the HA1 subunit was mutated, since it is located in close proximity to His184 and is one of the few residues that differ in the sequences of a highly pathogenic and a low pathogenic H5 stem. A residue at position 216 should influence the pK_A of His184 and therefore the pH-dependency of HA activation. To proof this idea residue 216 was exchanged with a residue of opposite charge in both the highly and low pathogenic H5 type.

The experiments showed that the protonation of His110 is not important for pH-sensing of the protein. Only one replacement (H110Y) shifted the pH of conformational change and fusion, and this effect has already been reported.[103] The shift is caused by the Tyr residue forming a hydrogen bond to Asn413 of the adjacent monomer thus stabilizing the complex. In contrast to the findings for His110, it could be shown that His184 has a significant impact on the pH-dependence of conformational change and fusion. Interestingly, the role of His184 seems to be more complex than simply destabilizing the complex due to protonation. Instead the results suggest that it has a double role of both, stabilizing the protein complex in its neutral state and destabilizing it when charged. This behavior can be rationalized by looking at its close environment shown in Figure 15. Two arginines (Arg220 and Arg229) are located in a very unusual location. Being completely shielded from the bulk water the two arginines have to be stabilized by an extensive hydrogen bond network mainly formed with backbone oxygens of residues of the same monomer. The only other type of hydrogen bond is formed with Asn210 of the adjacent monomer that is at the same time the only intermonomeric (HA1-HA1) hydrogen bond in the nearby region. Like His184 the two Arg residues are highly conserved among most subtypes. Asn210 is partially conserved as it is an asparagine, glutamine, threonine or serine in all subtypes. All these residues are structurally similar, polar and can form a hydrogen bond with Arg220. Any perturbation of this interaction network is very likely destabilizing the HA1-HA1 complex. Computational modeling showed indeed that only the protonated His184 interacts with Asn210 and thereby competes with Arg220 for hydrogen bonding. pK_A calculation with Karlsberg$^+$ showed that the His184 is deprotonated at pH 7. It therefore seems reasonable to surmise that the protonation of His814 destabilizes the structure. Support for the hypothesis that the two Arg residues are crucial for the protein stability comes from the experimental result of the point mutations. The three mutations H184A, H184N and H184D increased the pH for conformational change and fusion. Modeling of the variants showed that H184N and H184D destabilize Arg220 by interacting directly with Asn210, while the alanine residue is small enough to allow bulk water to reach Asn210. Replacing His184 with an arginine on the other hand prevents any conformational change and fusion. While this seems to be surprising at first glance, it becomes reasonable by analyzing the corresponding modeled structure. Arg184 forms a very stable double salt bridge with Glu231 on the same monomer, has contact to bulk water and due to its long side chain it does not interact with Asn210. Therefore Arg184 has no destabilizing influence on the hydrogen bond network around Arg220 and Arg229 resulting in the high stability of the protein complex as observed in the experiments. The latter mutation

shows that the residue at position 184 has not only the potential for destabilizing the complex but can also have a stabilizing effect. The role of His184 in pH-sensing was further studied with mutations at position 216. It would be expected that the exchange of a charged residue with one of opposing charge should shift the pK_A of His184 and therefore also the pH of fusion and conformational change. The experiments showed indeed that the R216E mutation in the high pathogenic H5 type increases the pH for fusion and conformational change while the E216R mutation has the opposite effect.

Overall it is supposed in this study that His184 is a crucial molecular switch at the HA1-HA1 interface that regulates the pH dependence of the conformational change of the Hemagglutinin protein. A detailed model is presented that explains how the deprotonated and protonated His184 stabilizes and destabilizes the HA1-HA1 complex respectively. It is further supposed that the threshold of the conformation change is fine-tuned by charged residues in the close proximity of His184 in particular at position 216.

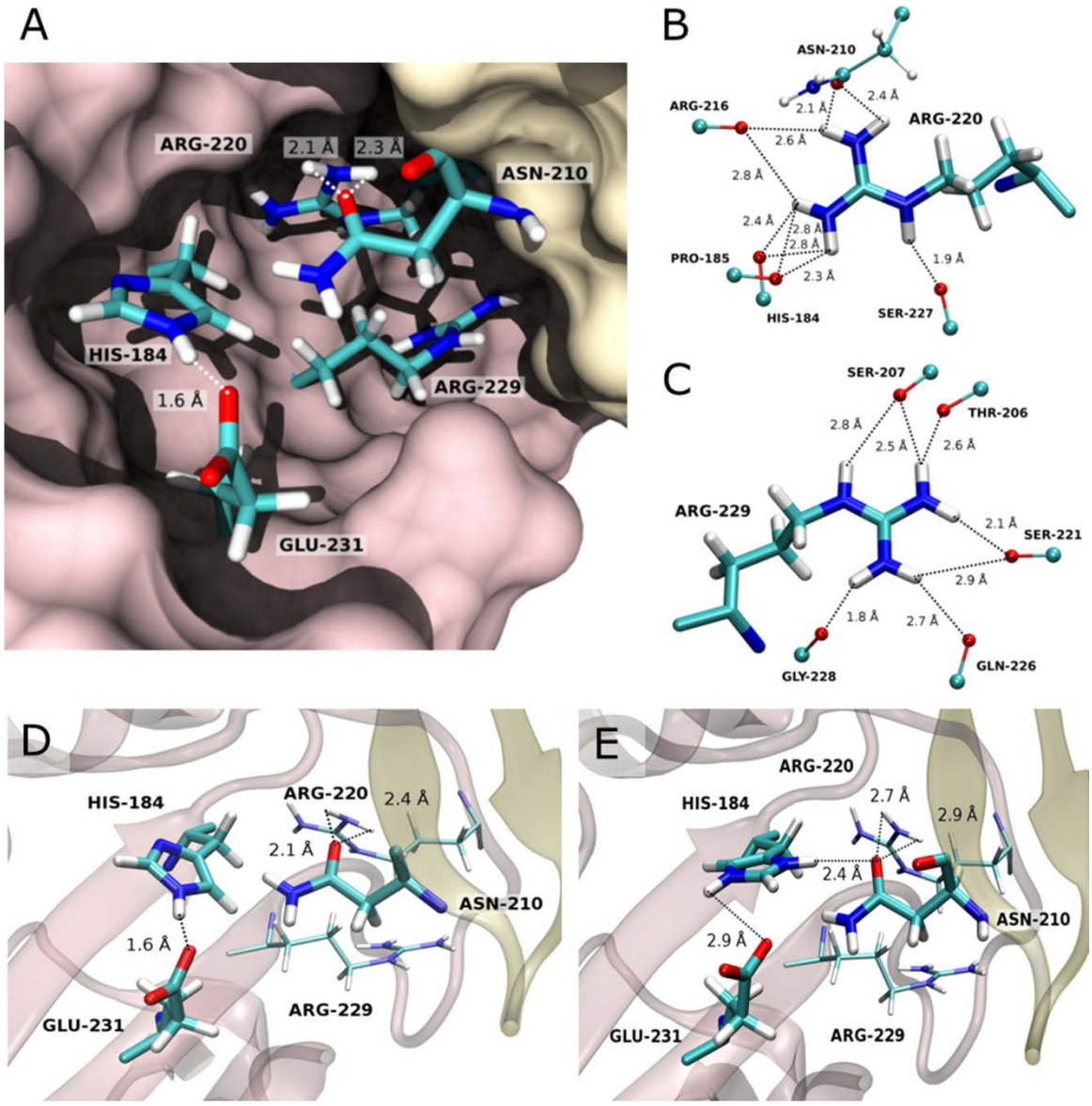


Figure 15: Interactions of His184 and neighboring residues at the HA1-HA1 interface. All images except for (E) show coordinates found in the crystal structure with PDB id 2IBX. Hydrogen atoms were added and energy was minimized with the software CHARMM. **(A)** Interactions at the HA1-HA1 interface. Monomers (chain A in brown and chain E in yellow) are depicted in surface representation, and residues which might be crucial for the regulation of HA1 monomer dissociation are shown with stick models. **(B and C)** Hydrogen bond network of residues Arg220 and Arg229. Interactions are formed between the polar atoms of residues Arg220 (B) and Arg229 (C) with the oxygen of the backbone amides, except for Asn210, where the hydrogen bond is formed with the side chain carbonyl-oxygen. **(D and E)** Crystal structure of the HA1-HA1 interface at neutral pH (D) and its modeled conformation upon protonation of His184 at a pH below 5 (E). Secondary structures of chains A (brown) and E (yellow) are displayed in cartoon representation with residues His184, Arg216, Glu231, and Asn210 in stick model. For modeling, the side chain of His184 has been rotated by 180 degrees, and the structure has been subsequently energy minimized. It is suggested here that a strong hydrogen bond is formed between His184 and Asn210, while the interaction between Asn210 and Arg220 is significantly weakened.

7. **Conclusions & Outlook**

For this thesis I developed two new procedures that extend the concept of the pK_A computation software Karlsberg+ with the aim to improve its accuracy. The two extensions address two of the main weaknesses of Karlsberg+ that had been proposed to be the main reasons for the surprisingly low performance in the SNase blind prediction competition in 2009. The challenging SNase benchmark set includes pK_A values of residues buried in the hydrophobic interior of the protein and thus being strongly shifted. The protonation or deprotonation of these residues is partially accompanied by local rearrangements of the protein structure and penetration of water into the protein interior.

The first new procedure named *Karlsberg+(cav)* was developed to investigate the influence of water filled cavities on pK_A calculations. The development was motivated by the observation, that the ability of the conventional SES-algorithm to describe deep crevasses and cavities inside the protein is insufficient. A novel algorithm named "cavity-algorithm" has been developed, that reliably detects protein cavities. These cavities have then been used to correct the spatial dependency of the dielectric constant for the electrostatic energy calculations. The procedure was extensively benchmarked with pK_A values from SNase variants. With *Karlsberg+(cav)* the accuracy of the calculated pK_A values could be enormously improved for some residues compared to the original version of Karlsberg+, especially for those residues located nearby the central cavity of the SNase protein. For some of the residues, however, the new procedure yielded no improvement at all. Overall the results showed that the correct consideration of water filled cavities in the electrostatic energy calculations is an important factor for an accurate computation of pK_A values in proteins.

While the first procedure simply applied a correction to the conventional Karlsberg+ approach, the second procedure, named *KB2+MD*, marks a step into a fundamental new direction. The underlying concept of Karlsberg+ to generate so called PACs, protein structures adapted to a certain pH interval, was kept, but the approach employed to create these PAC structures was radically reworked. A PAC is now represented by structures taken from a molecular dynamic (MD) simulation. This approach vastly generalizes the modeling capability and it becomes possible to account for even larger conformational changes. A key element of the procedure is the appropriate choice for a set of protonation pattern used for the individual MDs, each representing the protein at a different pH interval. This choice is a combination of assigning protonation states according to a fixed scheme and according to the results of pK_A computations based on the crystal structures. The usage of structures from molecular dynamic simulations, instead of subtle modified crystal structures made it necessary to adjust the traditional protocol used to calculate the electrostatic protonation energies. The effect of several adjustments was explored with variants of the *KB2+MD* procedure. The accuracy of the new procedure and its variants was evaluated with extensive benchmark calculations using a set of standard proteins, similar to the one that had been used in previous studies. With the most optimized variant of

KB2⁺MD, the accuracy of the pK_A calculations could be significantly improved compared to Karlsberg⁺ and is now on a par with the latest version of the widely used empirical prediction scheme PropKa.

The optimized *KB2⁺MD* procedure was furthermore applied to a part of a challenging SNase benchmark set and directly compared to the first new procedure *Karlsberg⁺(cav)*. Both procedures were able to calculate pK_A values with an enormously increased accuracy compared to Karlsberg⁺. For some of the individual SNase variants the two procedures perform very differently, indicating that the procedures indeed address complementary weak spots of Karlsberg⁺. Since the two new procedures modify completely different aspects of the pK_A prediction protocol, both could be straightforwardly combined into the *KB2⁺MD(cav)* procedure. The *KB2⁺MD(cav)* procedure further improved the accuracy of the pK_A prediction for the SNase variants significantly. This result shows clearly that both aspects are important to accurately predict pK_A values, the correct consideration of internal cavities and the extensive sampling of protein conformations.

Elucidating the function of the protein hemagglutinin of the influenza virus reveals the need for accurate pK_A computation methods that can account for larger structural changes of a protein. Hemagglutinin is located on the virus envelope and one of its functions is to trigger the fusion of virus and host cell membrane in response to acidification of the late endosomes. Also the pH value where fusion occurs is an essential property that has to be fine tuned by the virus to adapt to different hosts and transmission modes. The detailed pH-sensing mechanism employed by the hemagglutinin of a virus is still not well understood. In a collaboration project it was proposed that His184 is a crucial trigger residue in this mechanism and by influencing the pK_A of this residue with mutations in its nearby environment, the virus is able to adjust the pH of fusion. This hypothesis was supported with point mutations of hemagglutinin studied experimentally with measurements of conformational change and fusion, as well as with computational modeling to rationalize the experimental results. A detailed model was suggested, explaining how the deprotonated and protonated His184 is stabilizing and destabilizing the complex of the HA1 subunits of hemagglutinin, respectively. At the time of this project, the development and testing of the *KB2⁺MD* procedure was not yet finished. Due to its ability to sample even larger structural changes the new procedure should be an adequate tool to study the pH dependent stability of the HA1 complex. With experimental data of the pH-dependent conformational change being available now, hemagglutinin may serve as a benchmark system for the new procedure in future studies. The procedure itself has the potential to provide further insight into the structural details of this mechanism.

Although the accuracy of pK_A computation using electrostatic energy calculations could be vastly improved with the new procedures developed in this work, the SNase benchmark set shows that there is still a lot of space for improvement. The pK_A RMSD of about 2.3 pH units obtained with *KB2⁺MD(cav)* for the 10 variants is still far from the value of about 0.8 that could be achieved for

standard proteins. The experimental results of the hemagglutinin project showed that it would be beneficial to improve the accuracy of *KB2⁺MD* for standard proteins even further, since the measured shifts in the pH triggering the conformational change of the hemagglutinin protein are in the range of 0.2 to 0.5 pH units only.

The most obvious aspects to improve the results for the SNase variants are to play with the conditions of the MD simulation. Besides varying the length of equilibration and simulation it might be interesting to investigate the influence of explicit ions being used in the simulation. In contrast, the proteins in the standard benchmark set turned out to be rather insensitive to these details of the simulation protocol. Further improvement for both benchmark sets could be achieved by fine-tuning the force field parameters, e.g. the atomic partial charges and the vdW radii used to define the protein surface. The positive influence that an alternative set of vdW radii can have on the results was already shown in this work. However, in this study neither charges nor vdW radii were specifically optimized for electrostatic energy computations.

Besides a higher accuracy it would be desirable to extend the field of application of the new *KB2⁺MD* procedure, i.e. to account for complexes with ligands, ions or other proteins that dissociate with the protonation or deprotonation of titratable residues. While the technical implementation of this feature would be straightforward e.g. by simply running a MD for the bound and unbound state, the additional challenge to reliably estimate binding energies is introduced.

8. **Summary**

The pH dependent protonation of amino acids is important for many functions and properties of proteins. Titratable residues determine the pH stability of a protein, the efficiency of enzymatic reactions and they play an important role in proton and ion transport. To understand the function of many proteins on a molecular level in detail it is therefore crucial to have computational tools available that allow a reliable computation of the protonation behavior of these residues; a property usually described in terms of pK_A values. The pK_A of a residue is mainly influenced by electrostatic interactions, which in turn are sensitive to structural details of the protein in the environment of the residue. Within this thesis two new procedures have been developed that address the challenge of accurate pK_A computation in proteins.

Usually the interior of a protein is hydrophobic. Nevertheless in some proteins there are deep pockets or cavities that are filled with water molecules. These water-filled cavities offer two difficulties for pK_A computation. First, crystal structures of proteins are an unreliable source of information about these buried waters, since waters are often not resolved in crystal structures since they are disordered. Second, waters that were found in the protein structures have not been considered correctly in electrostatic energy calculations. The first new procedure *Karlsberg⁺(cav)* solves these problems by introducing an algorithm that is capable of reliably locating protein cavities and correcting the electrostatic energy calculations performed by the software Karlsberg⁺ based on the detected cavities. This procedure significantly improved the accuracy of the pK_A computations for a set of SNase variants, of which many do contain buried water molecules in the corresponding crystal structures.

While the first procedure was a correction for Karlsberg⁺, the second new procedure *KB2⁺MD* is a complete rework of its underlying concept. To account for the protonation dependent conformational variability of a protein, Karlsberg⁺ generates a set of structures that are subtle modeled crystal structures, each representing the protein at a different pH interval. The new *KB2⁺MD* procedure generalizes this idea by performing short molecular dynamic simulations for a set of different protonation patterns. Structures taken from these simulations are then analyzed with electrostatic energy calculations to obtain pK_A values. Extensive benchmark calculations on 194 residues in 13 standard proteins have been performed to optimize the procedure. It was found that the exclusion of intramolecular 1-2, 1-3 and 1-4 interactions, the so called "*non-bonded exclusion*", is essential to obtain good agreement between calculated and measured pK_A values. Furthermore the use of an alternative set of vdW radii to define the molecular surface of a protein, as well as energy minimization of the whole water box with a dielectric constant of $\varepsilon = 4$ prior to electrostatic energy calculations increased the accuracy of the pK_A computations. With the optimized *KB2⁺MD* procedure the accuracy could be significantly improved compared to Karlsberg⁺ and the result is now on a pair with the latest version of the widely used empirical prediction scheme PropKa, yielding a pK_A RMSD of 0.79.

Since both new procedures address different aspects of the pK_A prediction protocol, they could be combined directly. A reduced set of SNase variants has been used to compare the new procedures individually to the combination of both. While each of the new procedures vastly improved the accuracy of pK_A calculations, the combination of them yielded an even better agreement with the experimental values. It turned out that for the combined procedure the details of the cavity determination are less important than they are for *Karlsberg⁺(cav)*. While the resulting pK_A RMSD of 2.3 is still far from the value obtained for the standard benchmark set, the results mark an important step forward in the effort to compute precise pK_A values for the challenging SNase variants.

In a collaboration project the influenza virus protein hemagglutinin was studied. Hemagglutinin triggers the fusion of virus and host cell membranes in response to acidification of the late endosomes. Also being thought to play an important role in the strategy of the virus to adapt to different hosts and transmission modes, the pH-sensing mechanism of the protein is still not understood. Based on experimental studies and computational modeling a specific histidine was identified to be one of the key elements of this mechanism. Additionally, a detailed model on how this histidine regulates the pH dependent fusion was developed.

9. **Zusammenfassung**

Die pH abhängige Protonierung von Aminosäuren ist für zahlreiche Eigenschaften und Funktionen von Proteinen verantwortlich. Titrierbare Residuen bestimmen die pH Stabilität eines Proteins, die Effizienz von enzymatischen Reaktionen und spielen eine wichtige Rolle beim Transport von Protonen und Ionen. Um die Funktion von vielen Proteinen auf molekularer Ebene im Detail zu verstehen ist es daher essenziell, Werkzeuge zur Verfügung zu haben, die es erlauben das Titrationsverhalten dieser Residuen zuverlässig vorherzusagen. Der pK_S-Wert eines Moleküls wird hauptsächlich durch elektrostatische Interaktionen beeinflusst, die wiederum sensibel auf die strukturellen Details des Proteins in ihrer Umgebung reagieren. Im Rahmen dieser Dissertation wurden zwei neue Verfahren entwickelt, die sich den Herausforderungen der pK_S Bestimmung stellen.

Gewöhnlich ist das Innere eines Proteins hydrophob. Es gibt jedoch Proteine, die tiefe Taschen oder Hohlräume aufweisen, welche mit Wassermolekülen gefüllt sind. Diese wassergefüllten Hohlräume stellen in zweierlei Hinsicht eine Schwierigkeit für die pK_S Bestimmung dar. Einerseits sind Kristallstrukturen von Proteinen keine zuverlässige Informationsquelle über Positionen von Wassermolekülen, da diese häufig ungeordnet und deshalb experimentell nicht aufgelöst werden können. Desweiteren wurden Wassermoleküle, obwohl sie in den entsprechenden Proteinstrukturen sichtbar waren, in der Vergangenheit in elektrostatischen Energieberechnungen teilweise nicht richtig berücksichtigt. Das erste neue Verfahren *Karlsberg+(cav)* löst diese Probleme, indem es einen Algorithmus einsetzt, der zuverlässig Hohlräume im Inneren von Proteinen findet und diese Information nutzt, um die elektrostatischen Energieberechnungen von Karlsberg$^+$ entsprechend zu korrigieren. Dieses neue Verfahren erlaubte es die pK_S Werte von SNase Varianten mit einer deutlich höheren Genauigkeit zu berechnen. Die Kristallstrukturen vieler SNase Varianten weisen tatsächlich interne Wassermoleküle auf.

Das erste neue Verfahren ist eine Korrektur für das Programm Karlsberg$^+$, wohingegen das zweite Verfahren *KB2$^+$MD* eine grundlegende Überarbeitung des zugrundeliegenden Konzepts darstellt. Um die protonierungsabhängige strukturelle Variabilität von Proteinen zu berücksichtigen, erstellt Karlsberg$^+$ einen Satz von Proteinstrukturen durch vorsichtige Modellierung an Kristallstrukturen. Jede dieser erstellen Strukturen repräsentiert das Protein in einem bestimmten pH Intervall. Das neue *KB2$^+$MD* Verfahren dagegen verallgemeinert diese Idee, indem es für einen bestimmten Satz von Protonierungsmustern jeweils eine kurze Molekulardynamiksimulationen durchführt. Strukturen, die diesen Simulationen entnommen sind, dienen als Grundlage für die elektrostatischen Energieberechnungen mit denen die pK_S Werte bestimmt werden. Es wurden ausführliche Testrechungen an 194 titrierbaren Gruppen in 13 Standardproteinen durchgeführt um das Verfahren zu optimieren. Es stellte sich heraus, dass das Ausschließen von intramolekularen 1-2, 1-3 und 1-4 Interaktionen, die sogenannte "*non-bonded exclusion*", einen essentiellen Schritt darstellt um eine gute Übereinstimmung zwischen

experimentellen und berechneten Werten zu erzielen. Des Weiteren konnte die Genauigkeit der pK_S Berechnung durch die Verwendung eines alternativen Satzes von vdW Radien zur Definition der molekularen Oberfläche des Proteins sowie durch eine vor der Energieberechnung durchgeführte Energieminimierung der Wasserbox mit einer Dielektrizitätskonstante von $\varepsilon = 4$ verbessert werden. Mit dem neuen Verfahren konnte die Genauigkeit gegenüber Karlsberg+ deutlich verbessert werden. Mit einem pK_s RMSD von 0.79 ist das Ergebnis nun auf Augenhöhe mit der neusten Version der empirischen pK_S Vorhersagesoftware PropKa.

Da die beiden neuen Verfahren unterschiedliche Aspekte des verwendeten Protokolls zur pK_s Bestimmung betreffen, konnten beide ohne größere technische Probleme kombiniert werden. Anhand einer reduzierten Auswahl von SNase Varianten wurden die beiden Verfahren einzeln mit der Kombination aus beiden verglichen. Jedes der Verfahren im Einzelnen verbesserte die Genauigkeit der pK_S Berechnung enorm. Eine noch bessere Übereinstimmung mit den gemessenen Werten konnte jedoch mit der Kombination aus beiden Verfahren erzielt werden. Es stellte sich heraus, dass mit dem kombinierten Verfahren die Details der Hohlraumbestimmung viel weniger ins Gewicht fallen als dies für *Karlsberg+(cav)* der Fall war. Obwohl der erzielte pK_S RMSD von 2.3 deutlich hinter dem Wert zurückbleibt, der für den Standardtestsatz erreicht wurde, ist das Ergebnis dennoch ein großer Schritt auf dem Weg die herausfordernden pK_S Werte der SNase Varianten richtig vorherzusagen.

In einem Kooperationsprojekt wurde das Protein Hemagglutinin des Influenza Virus untersucht. Hemagglutinin induziert die Fusion des Virus mit der Membran der Wirtszelle als Reaktion auf die Ansäuerung später Endosomen. Obwohl vermutet wird, dass der Mechanismus zum „erfühlen“ des pH Wertes eine wichtige Rolle in der Strategie des Virus spielt, sich an bestimmte Wirtsorganismen und Übertragungswege anzupassen, ist dieser nach wie vor unverstanden. Auf der Grundlage von experimentellen Untersuchungen und computergestützten Modellierungen wurde ein einzelnes Histidin identifiziert, welches ein Schlüsselelement dieses Mechanismus darstellt. Desweiteren wurde ein Modell erarbeitet, das erklärt wie dieses Histidin den pH Wert der Fusion reguliert.

10. **References**

(1) Warshel, A., Calculations of Enzymatic Reactions: Calculations of pKa, Proton Transfer Reactions, and General Acid Catalysis Reactions in Enzymes. *Biochemistry* **1981,** 20, 3167-77.

(2) Warshel, A.; Åqvist, J., Electrostatic Energy and Macromolecular Function. *Annu. Rev. Biophys. and Biophys. Chem.* **1991,** 20, 267-298.

(3) McDonald, I. K.; Thornton, J. M., Satisfying Hydrogen Bonding Potential in Proteins. *J. Mol. Biol.* **1994,** 238, 777-793.

(4) Honig, B.; Nicholls, A., Classical Electrostatics in Biology and Chemistry. *Science* **1995,** 268, 1144-1149.

(5) Pace, C. N.; Grimsley, G. R.; Scholtz, J. M., Protein Ionizable Groups: pKa Values and Their Contribution to Protein Stability and Solubility. *J. Biol. Chem.* **2009,** 284, 13285-13289.

(6) Kampmann, T.; Mueller, D. S.; Mark, A. E.; Young, P. R.; Kobe, B., The Role of Histidine Residues in Low-pH-Mediated Viral Membrane Fusion. *Structure* **2006,** 14, 1481-1487.

(7) White, J. M.; Delos, S. E.; Brecher, M.; Schornberg, K., Structures and Mechanisms of Viral Membrane Fusion Proteins: Multiple Variations on a Common Theme. *Crit. Rev. Biochem. Mol. Biol.* **2008,** 43, 189-219.

(8) Woelke, A. L.; Galstyan, G.; Galstyan, A.; Meyer, T.; Heberle, J.; Knapp, E. W., Exploring the Possible Role of Glu286 in CcO by Electrostatic Energy Computations Combined with Molecular Dynamics. *J. Phys. Chem. B* **2013,** 117, 12432-12441.

(9) Alexov, E.; Mehler, E. L.; Baker, N.; M. Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M., Progress in the Prediction of pKa Values in Proteins. *Proteins: Struct., Funct., Bioinf.* **2011,** 79, 3260-3275.

(10) Olsson, M. H. M.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H., PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011,** 7, 525-537.

(11) Baptista, A. M.; Teixeira, V. H.; Soares, C. M., Constant-pH Molecular Dynamics Using Stochastic Titration. *J. Chem. Phys.* **2002,** 117, 4184-4200.

(12) Lee, M. S.; Freddie R Salsbury, J.; Brooks III, C. L., Constant-pH Molecular Dynamics Using Continuous Titration Coordinates. *Proteins: Struct., Funct., Bioinf.* **2004,** 56, 738-752.

(13) Kieseritzky, G.; Knapp, E. W., Optimizing pK_A Computation in Proteins with pH Adapted Conformations. *Proteins: Struct., Funct., Bioinf.* **2008,** 71, 1335-1348.

(14) Rabenstein, B.; Knapp, E. W., Calculated pH-Dependent Population and Protonation of Carbon-Monoxy-Myoglobin Conformers. *Biophys. J.* **2001,** 80, 1141-1150.

(15) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-hydrogen Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins using the CHARMM22 Force Field. *J. Phys. Chem. B* **1998,** 102, 3586-3616.

(16) Schutz, C. N.; Warshel, A., What Are the Dielectric "Constants" of Proteins and How To Validate Electrostatic Models? *Proteins: Struct., Funct., Genet.* **2001,** 44, 400-417.

(17) Simonson, T.; Brooks III, C. L., Charge Screening and the Dielectric Constant of Proteins: Insights from Molecular Dynamics. *J. Am. Chem. Soc.* **1996,** 118, 8452-8458.

(18) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A., Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **2001,** 98, 10037-10041.

(19) Bashford, D.; Karplus, M., pKa's of Ionizable Groups in Proteins: Atomic Detail from a Continuum Electrostatic Model. *Biochemistry* **1990,** 29, 10219-10225.

(20) You, T. J.; Bashford, D., Conformation and Hydrogen Ion Titration of Proteins: A Continuum Electrostatic Model with Conformational Flexibility. *Biophys. J.* **1995,** 69, 1721-1733.

(21) Ullmann, G. M.; Knapp, E. W., Electrostatic Models for Computing Protonation and Redox Equilibria in Proteins. *European Biophysics Journal* **1999,** 28, 533-551.
(22) Tanford, C.; Kirkwood, J. G., Theory of Protein Titration Curves. *J. Am. Chem. Soc.* **1957,** 79, 5333-5339.
(23) The Protein Data Bank: http://www.rcsb.org/pdb/.
(24) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000,** 28, 235-242.
(25) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K., Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005,** 26, 1781-1802.
(26) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: A Program for Macromolecular Energy Minimization and Dynamics Calculations. *J. Comput. Chem.* **1983,** 4, 187-217.
(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983,** 79, 926.
(28) Katayanagi, K.; Miyagawa, M.; Matsushima, M.; Ishikawa, M.; Kanaya, S.; Nakamura, H.; Ikehara, M.; Matsuzaki, T.; Morikawa, K., Structural details of ribonuclease H from Escherichia coli as refined to an atomic resolution. . *J. Mol. Biol.* **1992,** 223, 1029-1052.
(29) Humphrey, W.; Dalke, A.; Schulten, K., VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics and Modelling* **1996,** 14, 33-38.
(30) Kieseritzky, G. Shaping Electrostatic Energy Computations in Proteins: The ClC-Type Proton-Chloride Antiporter Function Dissertation. Freie Universität Berlin, **2011**.
(31) García-Moreno, B. E.; Dwyer, J. J.; Gittis, A. G.; Lattman, E. E.; Spencer, D. S.; Stites, W. E., Experimental Measurement of the Effective Dielectric in the Hydrophobic Core of a Protein. *Biophys. Chem.* **1997,** 64, 211-224.
(32) Dwyer, J. J.; Gittis, A. G.; Karp, D. A.; Lattman, E. E.; Spencer, D. S.; Stites, W. E.; Garcia-Moreno, B., High Apparent Dielectric Constants in the Interior of a Protein Reflect Water Penetration. *Biophys. J.* **2000,** 79, 1610-1620.
(33) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; Garcia-Moreno, B., Experimental pKa Values of Buried Residues: Analysis with Continuum Methods and Role of Water Penetration. *Biophys. J.* **2002,** 82, 3289-3304.
(34) Nguyen, D. M.; Reynald, R. L.; Gittis, A. G.; Lattman, E. E., X-ray and Thermodynamic Studies of Staphylococcal Nuclease Variants 192E and 192K: Insights Into Polarity of the Protein Interior. *J. Mol. Biol.* **2004,** 341, 565-574.
(35) Karp, D. A.; Gittis, A. G.; Stahley, M. R.; Fitch, C. A.; Stites, W. E.; Garcia-Moreno, B., High Apparent Dielectric Constant Inside a Protein Reflects Structural Reorganization Coupled to the Ionization of an Internal Asp. *Biophys. J.* **2007,** 92, 2041-2053.
(36) Harms, M. J.; Schlessman, J. L.; Chimenti, M. S.; Sue, G. R.; Damjanovic, A.; Garcia-Moreno, B., A Buried Lysine That Titrates With a Normal pKa: Role of Conformational Flexibility at the Protein-Water Interface as a Determinant of pKa Values. *Protein Sci.* **2008,** 17, 833-845.
(37) Isom, D. G.; Cannon, B. R.; Castaneda, C. A.; Robinson, A.; Bertrand, G. M. E., High Tolerance for Ionizable Residues in the Hydrophobic Interior of Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2008,** 105, 17784-17788.
(38) Harms, M. J.; Castaneda, C. A.; Schlessman, J. L.; Sue, G. R.; Isom, D. G.; Cannon, B. R.; Garcia-Moreno, B., The pKa Values of Acidic and Basic Residues Buried at the Same Internal Location in a Protein Are Governed by Different Factors. *J. Mol. Biol.* **2009,** 389, 34-47.
(39) Nielsen, J. E.; Gunner, M. R.; García-Moreno E, B., The pKa Cooperative: A Collaborative Effort to Advance Structure-Based Calculations of pKa Values and Electrostatic Effects in Proteins. *Proteins: Struct., Funct., Bioinf.* **2011,** 79, 3249-3259.
(40) Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; Velu, P. D.; García-Moreno, B., Charges in the hydrophobic interior of proteins. *Proceedings of the National Acadamy of Sciences of the United States of America* **2010,** 107, 16096-16100.

(41) Isom, D. G.; Castaneda, C. A.; Cannon, B. R.; Garcia-Moreno, B. E., Large Shifts in pKa Values of Lysine Residues Buried Inside a Protein. *Proc. Natl. Acad. Sci. U. S. A.* **2011,** 108, 5260-5265.
(42) Harms, M. J.; Schlessman, J. L.; Sue, G. R.; Garcia-Moreno, E. B., Arginine Residues at Internal Positions in a Protein are Always Charged. *Proceedings of the National Academy of Sciences* **2011,** 108, 18954-18959.
(43) Chimenti, Michael S.; Khangulov, Victor S.; Robinson, Aaron C.; Heroux, A.; Majumdar, A.; Schlessman, Jamie L.; Garcia-Moreno, B., Structural Reorganization Triggered by Charging of Lys Residues in the Hydrophobic Interior of a Protein. *Structure* **2012,** 20, 1071-1085.
(44) Chen, J. M.; Lu, Z. Q.; Sakon, J.; Stites, W. E., Increasing the Thermostability of Staphylococcal Nuclease: Implications for the Origin of Protein Thermostability. *J. Mol. Biol.* **2000,** 303, 125-130.
(45) Karp, D. A.; Stahley, M. R.; Garcia-Moreno, E. B., Conformational Consequences of Ionization of Lys, Asp, and Glu Buried at Position 66 in Staphylococcal Nuclease. *Biochemistry* **2010,** 49, 4138-4146.
(46) Ernst, J. A.; Clubb, R. T.; Zhou, H. X.; Gronenborn, A. M.; Clore, G. M., Demonstration of Positionally Disordered Water within a Protein Hydrophobic Cavity by NMR. *Science* **1995,** 267, 1813-1817.
(47) Yu, B.; Blaber, M.; Gronenborn, A. M.; Clore, G. M.; Caspar, D. L. D., Disordered Water Within a Hydrophobic Protein Cavity Visualized by X-ray Crystallography. *Proc. Natl. Acad. Sci. U. S. A.* **1999,** 96, 103-108.
(48) Adamek, D. H.; Guerrero, L.; Blaber, M.; Caspar, D. L. D., Structural and Energetic Consequences of Mutations in a Solvated Hydrophobic Cavity. *J. Mol. Biol.* **2005,** 346, 307-318.
(49) Otting, G.; Liepinsh, E.; Halle, B.; Frey, U., NMR Identification of Hydrophobic Cavities with Low Water Occupancies in Protein Structures Using Small Gas Molecules. *Nature Structural Biology* **1997,** 4, 396-404.
(50) Schlessman, J. L.; Abe, C.; Gittis, A.; Karp, D. A.; Dolan, M. A.; Garcia-Moreno, B. E., Crystallographic Study of Hydration of an Internal Cavity in Engineered Proteins with Buried Polar or Ionizable Groups. *Biophys. J.* **2008,** 94, 3208-3216.
(51) Castaneda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; Garcia-Moreno, B. E., Molecular Determinants of the pKa Values of Asp and Glu Residues in Staphylococcal Nuclease. *Proteins: Struct., Funct., Bioinf.* **2009,** 77, 570-588.
(52) Muegge I, K. E. W., Protein Water Interactions, Model Studies on BPTI. *Conference Proceedings* **1993,** 43.
(53) Weisel, M.; Proschak, E.; Schneider, G., PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors. *Chemistry Central Journal* **2007,** 1.
(54) Damjanovic, A.; Schlessman, J. L.; Fitch, C. A.; Garcia, A. E.; Garcia-Moreno, B., Role of Flexibility and Polarity as Determinants of the Hydration of Internal Cavities and Pockets in Proteins. *Biophys. J.* **2007,** 93, 2791-2804.
(55) Damjanovic, A.; Wu, X. W.; Garcia-Moreno, B.; Brooks, B. R., Backbone Relaxation Coupled to the Ionization of Internal Groups in Proteins: A Self-Guided Langevin Dynamics Study. *Biophys. J.* **2008,** 95, 4091-4101.
(56) Sandberg, L.; Edholm, O., pKa Calculations Along a Bacteriorhodopsin Molecular Dynamics Trajectory. *Biophys. Chem.* **1997,** 65, 189-204.
(57) Zhou, H.-X.; Vijayakumar, M., Modeling of Protein Conformational Fluctuations in pKa Predictions. *J. Mol. Biol.* **1997,** 267, 1002-1011.
(58) Wlodek, S. T.; Antosiewicz, J.; McCammon, J. A., Prediction of Titration Properties of Structures of a Protein Derived from Molecular Dynamics Trajectories. *Protein Sci.* **1997,** 6, 373-382.
(59) van Vlijmen, H. W. T.; Schaefer, M.; Karplus, M., Improving the Accuracy of Protein pKa Calculations: Conformational Averaging Versus the Average Structure. *Proteins: Struct., Funct., Genet.* **1998,** 33, 145-158.

(60) Koumanov, A.; Karshikoff, A.; Friis, E. P.; Borchert, T. V., Conformational Averaging in pK Calculations: Improvement and Limitations in Prediction of Ionization Properties of Proteins. *J. Phys. Chem. B* **2001,** 105, 9339-9344.
(61) Alexov, E., Role of the Protein Side-Chain Fluctuations on the Strength of Pair-Wise Electrostatic Interactions: Comparing Experimental with Computed pKas. *Proteins: Struct., Funct., Genet.* **2003,** 50, 94-103.
(62) Nielsen, J. E.; McCammon, J. A., On the Evaluation and Optimization of Protein X-Ray Structures for pKa Calculations. *Protein Sci.* **2003,** 12, 313-326.
(63) Eberini, I.; Baptista, A. M.; Gianazza, E.; Fraternali, F.; Beringhelli, T., Reorganization in Apo- and Holo-β-Lactoglobulin Upon Protonation of Glu89: Molecular Dynamics and pKa Calculations. *Proteins: Struct., Funct., Bioinf.* **2004,** 54, 744-758.
(64) Kuhn, B.; Kollman, P. A.; Stahl, M., Prediction of pKa Shifts in Proteins Using a Combination of Molecular Mechanical and Continuum Solvent Calculations. *J. Comput. Chem.* **2004,** 25, 1865-1872.
(65) Archontis, G.; Simonson, T., Proton Binding to Proteins: A Free-Energy Component Analysis Using a Dielectric Continuum Model. *Biophys. J.* **2005,** 88, 3888-3904.
(66) Makowska, J.; Baginska, K.; Makowski, M.; Jagielska, A.; Liwo, A.; Kasprzykowski, F.; Chmurzynski, L.; Scheraga, H. A., Assessment of Two Theoretical Methods to Estimate Potentiometric Titration Curves of Peptides: Comparison with Experiment. *The Journal of Physical Chemistry B* **2006,** 110, 4451-4458.
(67) Nilsson, L.; Karshikoff, A., Multiple pH Regime Molecular Dynamics Simulation for pK Calculations. *PLoS One* **2011,** 6, e20116.
(68) Mertz, J. E.; Pettitt, B. M., Molecular-Dynamics at a Constant pH. *International Journal of Supercomputer Applications and High Performance Computing* **1994,** 8, 47-53.
(69) Baptista, A. M.; Martel, P. J.; Petersen, S. B., Simulation of Protein Conformational Freedom as a Function of pH: Constant-pH Molecular Dynamics Using Implicit Titration. *Proteins: Struct., Funct., Genet.* **1997,** 27, 523-544.
(70) Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F., Simulating Proteins at Constant pH: An Approach Combining Molecular Dynamics and Monte Carlo Simulation. *Proteins: Struct., Funct., Bioinf.* **2002,** 47, 469-480.
(71) Mongan, J.; Case, D. A.; McCammon, J. A., Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comput. Chem.* **2004,** 25, 2038-2048.
(72) Machuqueiro, M.; Baptista, A. M., Constant-pH Molecular Dynamics with Ionic Strength Effects: Protonation-Conformation Coupling in Decalysine. *J. Phys. Chem. B* **2006,** 110, 2927-2933.
(73) Stern, H. A., Molecular Simulation with Variable Protonation States at Constant pH. *J. Chem. Phys.* **2007,** 126.
(74) Machuqueiro, M.; Baptista, A. M., Acidic Range Titration of HEWL Using a Constant-pH Molecular Dynamics Method. *Proteins-Structure Function and Bioinformatics* **2008,** 72, 289-298.
(75) Wallace, J. A.; Shen, J. K., Predicting pKa Values with Continuous Constant pH Molecular Dynamics. In *Methods in Enzymology, Vol 466: Biothermodynamics, Pt B*, 2009; Vol. 466, pp 455-475.
(76) Foit, L.; George, J. S.; Zhang, B. W.; Brooks III, C. L.; Bardwell, J. C. A., Chaperone Activation by Unfolding. *Proc. Natl. Acad. Sci. U. S. A.* **2013,** 110, E1254-E1262.
(77) Zeng, X.; Mukhopadhyay, S.; Brooks III, C. L., Residue-Level Resolution of Alphavirus Envelope Protein Interactions in pH-Dependent Fusion. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** 112, 2034-2039.
(78) Chen, J. H.; Brooks III, C. L.; Khandogin, J., Recent Advances in Implicit Solvent-Based Methods for Biomolecular Simulations. *Curr. Opin. Struct. Biol.* **2008,** 18, 140-148.
(79) Sondergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H., Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011,** 7, 2284-2295.

(80) Oshea, E. K.; Klemm, J. D.; Kim, P. S.; Alber, T., X-ray Structure of the GCN4 Leucine Zipper, a 2-Stranded, Parallel Coiled Coil. *Science* **1991,** 254, 539-544.
(81) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R., Combining Conformational Flexibility and Continuum Electrostatics for Calculation pKas in Proteins. *Biophys. J.* **2002,** 83, 1731-1748.
(82) Joshi, M. D.; Hedberg, A.; McIntosh, L. P., Complete Measurement of the pKa Values of the Carboxyl and Imidazole Groups in Bacillus Circulans Xylanase. *Protein Sci.* **1997,** 6, 2667-2670.
(83) Matousek, W. M.; Ciani, B.; Fitch, C. A.; Garcia-Moreno E, B.; Kammerer, R. A.; Alexandrescu, A. T., Electrostatic Contributions to the Stability of the GCN4 Leucine Zipper Structure. *J. Mol. Biol.* **2007,** 374, 206-219.
(84) Im, W.; Lee, M. S.; Brooks III, C. L., Generalized Born Model with a Simple Smoothing Function. *J. Comput. Chem.* **2003,** 24, 1691-1702.
(85) Nose, S., A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.* **1984,** 81, 511-519.
(86) Rashin, A. A.; Iofin, M.; Honig, B., Internal Cavities and Buried Waters in Globular Proteins. *Biochemistry* **1986,** 25, 3619-3625.
(87) Oliveberg, M.; Arcus, V. L.; Fersht, A. R., pKa Values of Carboxyl Groups in the Native and Denatured States of Barnase: The pKa Values of the Denatured State Are on Average 0.4 Units Lower Than Those of Model Compounds. *Biochemistry* **1995,** 34, 9424-9433.
(88) Gamiz-Hernandez, A. P.; Kieseritzky, G.; Galstyan, A. S.; Demir-Kavuk, O.; Knapp, E. W., Understanding Properties of Cofactors in Proteins: Redox Potentials of Synthetic Cytochromes b. *ChemPhysChem* **2010,** 11, 1196-1206.
(89) Robertazzi, A.; Galstyan, A.; Knapp, E. W., PSII Manganese Cluster: Protonation of W2, 05, 04 and His337 in the Si State Explored by Combined Quantum Chemical and Electrostatic Energy Computations. *Biochim. Biophys. Acta, Bioenerg.* **2014,** 1837, 1316-1321.
(90) Ishikita, H.; Saenger, W.; Loll, B.; Biesiadka, J.; Knapp, E. W., Energetics of a Possible Proton Exit Pathway for Water Oxidation in Photosystem II. *Biochemistry* **2006,** 45, 2063-2071.
(91) Schmidt am Busch, M.; Knapp, E. W., Accurate pKa Determination for a Heterogeneous Group of Organic Molecules. *ChemPhysChem* **2004,** 5, 1513-1522.
(92) Woelke, A. L.; Kuehne, C.; Meyer, T.; Galstyan, G.; Dernedde, J.; Knapp, E. W., Understanding Selectin Counter-Receptor Binding from Electrostatic Energy Computations and Experimental Binding Studies. *J. Phys. Chem. B* **2013,** 117, 16443-16454.
(93) Guan, Y.; Poon, L. L. M.; Cheung, C. Y.; Ellis, T. M.; Lim, W.; Lipatov, A. S.; Chan, K. H.; Sturm-Ramirez, K. M.; Cheung, C. L.; Leung, Y. H. C.; Yuen, K. Y.; Webster, R. G.; Peiris, J. S. M., H5N1 Influenza: A Protean Pandemic Threat. *Proc. Natl. Acad. Sci. U. S. A.* **2004,** 101, 8156-8161.
(94) Qin, Z.-l.; Zheng, Y.; Kielian, M., Role of Conserved Histidine Residues in the Low-pH Dependence of the Semliki Forest Virus Fusion Protein. *J. Virol.* **2009,** 83, 4670-4677.
(95) Kielian, M.; Rey, F. A., Virus Membrane-Fusion Proteins: More Than One Way to Make a Hairpin. *Nat. Rev. Microbiol.* **2006,** 4, 67-76.
(96) Harrison, S. C., Viral Membrane Fusion. *Nat. Struct. Mol. Biol.* **2008,** 15, 690-698.
(97) Wilson, I. A.; Skehel, J. J.; Wiley, D. C., Structure of the Hemagglutinin Membrane Glycoprotein of Influenca-Virus at 3-A Resolution. *Nature* **1981,** 289, 366-373.
(98) Bullough, P. A.; Hughson, F. M.; Skehel, J. J.; Wiley, D. C., Structure of Influenca Hemagglutinin at the pH of Membrane Fusion. *Nature* **1994,** 371, 37-43.
(99) Kemble, G. W.; Bodian, D. L.; Rose, J.; Wilson, I. A.; White, J. M., Intermonomer Disulfide Bonds Impair the Fusion Activity of Influenza-Virus Hemagglutinin. *J. Virol.* **1992,** 66, 4940-4950.
(100) Godley, L.; Pfeifer, J.; Steinhauer, D.; Ely, B.; Shaw, G.; Kaufmann, R.; Suchanek, E.; Pabo, C.; Skehel, J. J.; Wiley, D. C.; Wharton, S., Introduction of Intersubunit Disulfide Bonds in the Membrane-Distal Region of the Influenza Hemagglutinin Abolishes Membrane-Fusion Activity. *Cell* **1992,** 68, 635-645.
(101) Bottcher, C.; Ludwig, K.; Herrmann, A.; van Heel, M.; Stark, H., Structure of Influenza Haemagglutinin at Neutral and at Fusogenic pH by Electron Cryo-Microscopy. *FEBS Lett.* **1999,** 463, 255-259.

(102) Huang, Q.; Opitz, R.; Knapp, E. W.; Herrmann, A., Protonation and Stability of the Globular Domain of Influenza Virus Hemagglutinin. *Biophys. J.* **2002,** 82, 1050-1058.
(103) Herfst, S.; Schrauwen, E. J. A.; Linster, M.; Chutinimitkul, S.; de Wit, E.; Munster, V. J.; Sorrell, E. M.; Bestebroer, T. M.; Burke, D. F.; Smith, D. J.; Rimmelzwaan, G. F.; Osterhaus, A. D. M. E.; Fouchier, R. A. M., Airborne Transmission of Influenza A/H5N1 Virus Between Ferrets. *Science* **2012,** 336, 1534-1541.