Published in final edited form as:

Curr Pharm Des. 2013 ; 19(23): 4316–4322.

Rigorous Incorporation of Tautomers, Ionization Species, and Different Binding Modes into Ligand-Based and Receptor-Based 3D-QSAR Methods

Senthil Natesan and Stefan Balaz*

Albany College of Pharmacy and Health Sciences, Vermont Campus, Colchester, VT 05446

Abstract

Speciation of drug candidates and receptors caused by ionization, tautomerism, and/or covalent hydration complicates ligand- and receptor-based predictions of binding affinities by 3-dimensional structure-activity relationships (3D-QSAR). The speciation problem is exacerbated by tendency of tautomers to bind in multiple conformations or orientations (modes) in the same binding site. New forms of the 3D-QSAR correlation equations, capable of capturing this complexity, can be developed using the time hierarchy of all steps that lie behind the monitored biological process – binding, enzyme inhibition or receptor activity. In most cases, reversible interconversions of individual ligand and receptor species can be treated as quickly established equilibria because they are finished in a small fraction of the exposure time that is used to determine biological effects. The speciation equilibria are satisfactorily approximated by invariant fractions of individual ligand and receptor species for buffered experimental or in vivo conditions. For such situations, the observed drug-receptor association constant of a ligand is expressed as the sum of products, for each ligand and receptor species pair, of the association microconstant and the fractions of involved species. For multiple binding modes, each microconstant is expressed as the sum of microconstants of individual modes. This master equation leads to new 3D-QSAR correlation equations integrating the results of all molecular simulations or calculations, which are run for each ligand-receptor species pair separately. The multispecies, multimode 3D-QSAR approach is illustrated by a ligand-based correlation of transthyretin binding of thyroxine analogs and by a receptor-based correlation of inhibition of MK2 by benzothiophenes and pyrrolopyrimidines.

Keywords

ionization; tautomerism; covalent hydration; 3D-QSAR; prediction of binding affinity; CoMFA; Linear Response method; QM/MM

Introduction

Speciation of drug candidates in biological media complicates elucidation of structure-activity relationships [1]. Ionization, tautomerism, and/or covalent hydration lead to the replacement of one dominant ligand structure of known concentration by a mixture of interconverting species, which all have different binding substructures and affinities, and are present in concentrations, which need to be estimated. Ionization changes H-bond donors and acceptors to negatively or positively charged fragments for acids and bases,

*Corresponding author: Stefan Balaz, Albany College of Pharmacy and Health Sciences, Vermont Campus, 261 Mountain View Drive, Colchester, VT 05446; phone +1-802-735-2615, fax +1-802-654-0716, stefan.balaz@acphs.edu.

respectively. Tautomerism in most cases switches the positions of H-bond donors and acceptors. Covalent hydration introduces one or two H-bond donating and accepting hydroxyls bound to originally electron-deficient atoms of heteroaromatic rings or carbonyl/aldehyde groups, respectively. All these changes maintain, in most cases, connectivity of the ligand skeleton, although the character of some bonds is changing. These facts increase the probability of binding in multiple modes, which generally increases with growing flexibility of ligands and binding sites. Obvious candidates for multi-mode binding are the heterocycles having different tautomerizing heteroatoms located at symmetric positions with regard to either symmetry axis or rotation axis, which can easily flip by 180° to maintain the original bonding pattern.

Speciation is not limited to ligands, it also affects the receptors. Components of the binding sites, e.g., several amino acid residues [2,3], cofactors (porphyrin [4,5], NAD⁺ and biotin [6]), and nucleobases [7-9] can undergo ionization and tautomerism under physiological conditions. All species that exist in reasonable concentrations should be considered in binding studies, except the species of components that are too distant from the binding site to affect binding.

The majority of approved and investigational drugs contain tautomerism-prone heteroaromatic ring systems and heteroatom-rich substructures [10], as well as one or more ionizing groups [6,11]. We will focus on the impact of speciation on the receptor binding, although the consequences for pharmacokinetics are also difficult to predict and require more complex models than are those traditionally used for non-speciating ligands. Speciation has been studied for several decades but its consideration in drug candidate binding to the receptors is qualitative or semi-quantitative at best. Rigorous treatment schemes for 3D-QSAR techniques are scarce. Here we review the approaches that attempt to fill this gap in the armory of methods for prediction of binding affinity.

History of Speciation and Multi-Mode Binding in QSAR Studies

Chemical details about structure of acids and bases emerged in the 19th century, although the concept of acidity was known since the BC times in ancient Greece thanks to the use of taste to classify substances. First, Justus von Liebig (1803-1873) proposed the association of acidity with the presence of hydrogen. Later, Svante A. Arrhenius (1859-1927) defined acids and bases as the substances releasing protons and hydroxyls to the solution and described their mutual reaction as the basis of neutralization. Johannes Brønsted and Thomas Lowry independently modified the definition of a base as a substance capable of binding hydrogen [12]. For aqueous solutions, these definitions are still sufficient in most cases. The concentration of protons in the medium was initially estimated using colorimetry and pH-indicators. Søren P.L. Sørensen (1868-1939) introduced the pH scale in 1909, and was among the first to use a hydrogen ion electrode [12]. The effects of acidity of the medium on biological effects of ionizable substances were reported in the early years of the 20th century [13]. The elements of the pH-partition hypothesis, formulated much later [14], namely faster distribution of nonionized molecules in biological systems as compared to ionized species, were recognized at that time [15]. Quantitative relations of acidity and reactivity to compound structure appeared in the 1930s [16,17]. This knowledge led to semi-quantitative considerations of the influence of ionization on biological activity a few years later [18-23].

A straightforward approach to the speciation problem in QSAR studies is the recalculation of the binding characteristics to the concentration of the binding species, if the species is known. This approach was first used by Martin and Hansch in 1971 [24] to account for ionization in QSAR analysis of metabolic oxidation. In many situations, limited knowledge of the bound species prohibits the use of this simple measure. In addition, this approach

precludes examination of the possibility that some species of a ligand may bind simultaneously.

In 1976, Martin published model-based dependences for equilibrium distribution of ionizable molecules, which were combined with interactions of ionized or nonionized species with the receptors, if they were related in a simple way to structures of studied compounds [25]. Several application examples of this approach were published later [26].

The concept of tautomerism emerged in the second half of the 19th century [27], when the principles of chemical structure were formed. Thanks to the analytical ability to distinguish hydroxy and ketone groups, keto-enol tautomerism was recognized in 1880, when Emil Erlenmeyer formulated his rule stating that hydroxy group attached to a double bond becomes ketone [28]. Tautomerism was known to affect the drug fates since 1960s [29] and was intensively studied since then in both drugs [30] and potential receptors [5,31-34]. Quantitative analyses of partitioning of tautomers were made in 1990s [35,36]. Systematic studies of tautomer equilibria started at the turn of the century [37-39].

Covalent hydration seems to be a more contemporary phenomenon [40-42], affecting fewer compounds than ionization and tautomerism. Addition of water to electron-poor atoms of aromatic and heteroaromatic rings belongs to a broader category of Meisenheimer - complex formations, so quantitative analyses of covalent hydration rates and equilibria can be found under this term [43]. Although covalent hydration is less frequently emphasized in QSAR studies [44], its inclusion is equally significant for drug development as that of other speciation processes.

Speciation only slowly made its way into 3D-QSAR studies [45,46]. In numerous reports, several species were included on incidental basis: if the data were not explained by existing models for one species, binding of other species was examined as potential reason for explaining the behavior of outliers. Alternatively, the entire set of compounds was analyzed as one tautomer and then as other tautomer, neglecting the fractions of individual tautomers and the possibility of interconversions [47]. In sum, the species were not treated in a systematic way, i.e. not all species were examined and for those that were, the calculations were only done for outliers, not for all compounds.

An attempt to consider multiple binding modes and species systematically in 3D-QSAR was made in the QUASAR approach of Vedani et al. [48]. Unfortunately, the approach lacks a rigorous theoretical background, as described in detail previously [49]. There are significant issues with description of binding equilibria. Binding energy of a ligand is calculated as a linear combination of the contributions of individual modes and species, which are weighted by bound prevalences. This equation contradicts the thermodynamic master equation (eq 2 below, with each $K = \exp(-G/RT)$, where G is the free energy of binding, R is the universal gas constant and T is absolute temperature). Bound prevalences are calculated using "normalized Boltzmann distribution" but the used function is different from the Boltzmann function and does not even provide bound prevalences which would add to one for each ligand. In addition, eq 2 shows the used bound prevalences are not needed to calculate binding energy. On the other hand, fractions of individual species in the receptor surroundings, which are vital data for calculation of the overall binding energy of each ligand (see eq 2), are not used at all. The next approach from this laboratory, Raptor [50], is considered a descendant of the QUASAR approach. The Raptor's complete correlation function was not published, so it can be expected that the function contains the same errors as in the QUASAR approach. While QUASAR and Raptor approaches contain some interesting features, they do not have the ability to describe multi-species binding correctly.

We recently published a rigorous approach to inclusion of multiple species into prediction of binding affinities. The correlation equation is based on the thermodynamically correct eq 2. The approach was applied to both ligand- and receptor-based correlations. The results are briefly described below, with the focus on most important conclusions and some aspects that were not discussed in the original publications.

Time Hierarchy of Binding-Underlying Processes

Relative velocities of the steps partaking in binding process, including speciation, determine the model type that describes the overall outcome sufficiently well. If all involved processes are reversible, the time needed to establish the slowest equilibrium needs to be compared with the overall exposure time that is used to determine the biological effect. This criterion differs from the recommended use of the rate of processes for which the biological effect is measured for comparison with the speciation rates [51]. What are the rates of involved steps?

For ionization, the ability to establish the equilibria practically instantaneously has been known for long time [52]. The rate constants of ionization need only be considered if very fast processes are studied and the dissociation constants K_a are proper descriptors for the vast majority of processes in drug development.

The tautomerism rates depend on the nature of broken and created bonds. For prototropic tautomers, which are typical for drugs, a slow interchange with the half-lives in hours is only observed for tautomers using CH bond cleavage and formation. The CH-to-NH, -OH and -SH tautomer conversions have the half-lives in the range of a second [51]. The rates are even faster for the NH-to-OH and OH-to-OH tautomerism: some keto-enol tautomers convert on the picosecond time scale [53].

The rate of covalent hydration, a nucleophilic addition of a water molecule to double or triple bonds outside (e.g. carbonyl and aldehyde groups) or inside an aromatic or heteroaromatic ring, depends on the structure of the ligand but is often so fast that it needs to be monitored using stopped-flow techniques [54].

The fast species interconversions facilitate the description of the ligand binding to macromolecules because the binding process is fully characterized by the equilibrium constants and no kinetic models need to be used. A rule of thumb says that if the equilibria are established within 5-10% of the exposure time, after which overall binding of a ligand is measured [55], the equilibrium description is sufficient, eliminating the need to consider the kinetics of the species interconversion process.

The binding of ligands to receptors is mostly caused by weak, noncovalent interactions. In this situation, the equilibrium is established quickly [56] as compared to the exposure time and the association constants are the only parameters needed to fully characterize the event. Sometimes the primary complex is stabilized by subsequent steps, such as release or binding of water molecules [57-60], coordination bond formation [61], covalent interaction [62], larger conformational change of the receptor, e.g. closing the loops (flaps) on the binding site of HIV-1 aspartyl protease [63], or other processes denoted collectively as receptor isomerization [64]. The steps may be slower and necessitate the use the kinetic rate constants for a complete characterization of the ligand-receptor complex formation, if the exposure time is not sufficiently long. The ligand-receptor interactions with complex kinetics may exhibit slow dissociation, which can be of importance for overall efficacy [65]. We will focus on the primary complexes, for which the association constants sufficiently characterize the binding events.

Multispecies Multimode Binding

The binding equilibria for ligand (l species) and the receptor (r species) binding in m binding modes can be described by the scheme shown in Figure 1. The binding of the ligand species L_i with the receptor species R_j in the k-th binding mode is described by the association microconstant K_{ijk}

$$K_{ijk} = \frac{[LR_{ijk}]}{[L_i][R_j]} \quad (1)$$

Here the square brackets denote concentrations, which are, in most situations, sufficiently low to approximate chemical activities. The observed association constant K usually does not distinguish between individual ligand-receptor complexes, ligand species, and receptor species and uses the overall concentrations, as shown in the second term:

$$[L][R] = \frac{\sum_{i=1}^l \sum_{j=1}^r \sum_{k=1}^m [LR_{ijk}]}{[L][R]} \frac{[L_i][R_j]}{[L_i][R_j]} = \sum_{i=1}^l \sum_{j=1}^r \sum_{k=1}^m f_i f_j K_{ijk} \quad (2)$$

In the third term, the total complex concentration is expressed as the sum, over l ligand species, r receptor species, and m binding modes, of $l \times r \times m$ complexes. In the fourth term, each of the summands is multiplied by unity written as $[L_i][R_j]/[L_i][R_j]$, to introduce the fraction $f_i = [L_i]/[L]$ of ligand species L_i and the fraction $f_j = [R_j]/[R]$ of receptor species R_j . This is simply done by switching the denominators in the two ratios of the fourth term. The switch will also generate, in the fifth term, the microconstants K_{ijk} , which are defined in eq 1. The observed association constant K is given as the sum of all microconstants over m binding modes, r receptor species R and l ligand species L, whereby each microconstant is weighted by the product of the relevant ligand species fraction f_i and receptor species fraction f_j . Here f_i and f_j are the fractions of unbound ligand and receptor species, for which prediction methods are available (e.g. ionization/tautomerism [66-70] with benchmarking/additional references [71-73] and covalent hydration [74] for ligands and ionization for proteins [75,76] with benchmarking/additional references [73,77]). Thus, f_i and f_j are independent variables in the correlation equations, which are calculated before optimization assuming that the pH value of the medium is known. For cell-QSAR applications predicting binding affinities of the receptors inside the cells, the pH of the medium may differ from the plasma value and can become an optimized parameter for some simple speciation cases [78].

The speciation processes often form more complex networks than the catenary sequence shown in Figure 1. However, the form of eq 2 holds also in the more complex situations because it is not affected by the structure of the speciation network: only the sum of all ligand, receptor and complex species is used in the derivation of eq 2.

Multispecies Multimode QSAR Studies

The master equation 2 shows the association between the observed association constant (K), the fractions of participating species (f_i , f_j) and the microconstants (K_{ijk}). Individual 3D-QSAR techniques model binding of a single ligand species to a single receptor site, as described by a microconstant K_{ijk} . Therefore, the single-species 3D-QSAR expressions need to be adapted using eq 2 to describe multispecies, multimode binding, as illustrated below for selected ligand-based and receptor-based approaches.

The data characterizing interactions of individual ligand species with individual receptor species are obtained in the same way as for the single-mode situation. These data are then processed by any software capable of nonlinear regression analysis, according to the

correlation equation resulting from eq 2 and the QSAR equation characteristic for the given approach.

Once the multispecies, multimode models are optimized and the magnitude of each microconstant can be calculated, the prevalence of bound species LR_{ijk} can be calculated as K_{ijk}/K . This is an objective way to select one or more bound species from the set of all considered species and eliminate the species that do not contribute to binding. No subjective input about the binding prevalences is needed. The number of considered species, in principle, does not play a role: the association constants K_{ijk} in eq 2 of non-binding species will be practically equal to zero and will not contribute to the association constant. The number of considered species is thus only given by the ability of optimization techniques to extract appropriate signals from the data.

Applications in Ligand-Based QSAR

The LB-QSAR techniques use structures and binding data of a set of ligands to create a putative receptor site model explaining the data, in the situation when no detailed receptor structure is available. The most frequently used approach in this category is called Comparative Molecular Field Analysis (CoMFA) [79]. The independent input variables are the energies, X , of probes with energy types (subscript t) that are placed in grid points (subscript p) surrounding ligand superposition. They are correlated with the dependent variable, the binding characteristics proportional to the free binding energy (e.g., $\log K_{ijk}$) using regression analysis, usually partial least squares, according to the correlation equation:

$$\log K_{ijk} = \sum_p \sum_t C_{pt} X_{ptijk} \quad (3)$$

The ligands need to have the ligand species i , the receptor species j , and the binding mode k specified. The optimized regression coefficients C determine the strength and type of the interaction in individual grid points and represent the model of the receptor site.

The single-species, single-mode eq 3 is converted for description of the multispecies, multimode binding using eq 2:

$$\log K = \log \sum_{i=1}^l \sum_{j=1}^r \sum_{k=1}^m f_i f_j 10^{\sum_p \sum_t C_{pt} X_{ptijk}} \quad (4)$$

Interestingly, this extension only changes the form of the equation and does not increase the number of optimized coefficients C , which are associated with the grid points p and energy types t but do not depend on the species or modes. The energies X need to be calculated for each considered species and mode. Multiple receptor (model) species can, in principle, be considered in LB-QSAR models, although current methods do not use this opportunity and then $r = 1$ in eq 4.

Optimization of coefficients C in eq 4 depends on the number of grid points and probe types, which are chosen to characterize the binding site model. If the number of coefficients is larger than the number of compounds, partial least squares procedure can be used after linearization of eq 4, as described previously [80]. If a large number of combinations of a few points/types is examined [49], coefficients C are optimized by nonlinear regression analysis. In both cases, the best model is selected on the basis of quality of the fit for the training set, using the correlation coefficient and the coefficient errors, as well as the leave-one-out procedure to test the stability of the model. Predictivity of the model is then

evaluated using the test set, which was not used in any step of the coefficient optimization process.

Finding the right superposition is the key to a successful CoMFA model. The ability to consider several modes and/or species for each ligand in one optimization run has a significant impact on the way how CoMFA analyses are performed. There is much more freedom in examining different modes and species than in the classical approach. Let us assume that a series of 40 ligands is studied, each forming two species which can bind in two modes each, resulting in four species/modes for each ligand. To cover all possible combinations, 4^{40} classical one-mode/species CoMFA models need to be performed. If one CoMFA analysis only takes a second, the time needed to run all models would be 3.8×10^{16} years. So the CPU time in the order of thousands hours to find an acceptable, albeit not necessarily best, solution in the multispecies, multimode settings represents a significant gain in our ability to examine the complexity of binding events.

The first multimode CoMFA analysis [80] dealt with binding of 34 polychlorinated dibenzofurans (PCDF) to aryl hydrocarbon receptor. These compounds were ideal for a proof-of-the-concept study because of the lack of conformational flexibility and speciation (so $l = r = 1$ in eq 4). Yet the classical, one-mode CoMFA analysis resulted in poor predictions for the test set, which was not used in calibration. Inclusion of multiple binding modes, resulting from flipping the molecules around the symmetry axes and shifting the skeleton in the putative site model improved the predictions for the test set. Most compounds were predicted to bind in one or two binding modes, out of the 16 considered modes. The results confirmed feasibility of the multimode concept.

The next case [49] was more challenging: binding to transthyretin of 28 thyroxine analogs, each forming up to four ionization species under the conditions of experiment. For six compounds, the binding modes were determined by X-ray analysis. Two compounds exhibit unique (i.e. prevalence 100%), although completely reversed modes. For each of the remaining four compounds, two modes are observed with different prevalences. Obviously, classical one-mode analysis could not cope with this situation and the results showed very low predictivity. The multispecies, multimode CoMFA provided much better predictions and explained 75% of the variance of the test set that was not used in calibration. More importantly, the analysis correctly identified the binding modes for the six compounds, including the approximate prevalences of the dual modes for the four compounds. The receptor structure information that was used in building the model was limited to the definition of binding orientations. The results showed that the LB-QSAR models provide better predictions if they come closer to reality by considering multiple species and modes.

Applications in Receptor-Based QSAR

We prefer to use the Linear Response (LR) method in our quantum mechanics/molecular mechanics (QM/MM) modification [81-85] because of a favorable cost/performance ratio and the ability to cope with polarized H-bonds and electrostatic interactions, coordination bonds, and other interactions that cause problems when classical force fields are deployed. The procedure starts with best docked poses, which have the geometry and charges optimized by a QM/MM approach, and performs the conformational space mapping by a force-field based molecular dynamics (MD) of the hydrated complex. In the final step, the time-averaged geometries of the complex, ligand and unliganded receptor, all with appropriate hydration, are used to calculate the QM/MM binding energy $\langle E_{\text{QM/MM}} \rangle$ and the difference upon binding of the solvent-accessible surface area $\langle \text{SASA} \rangle$. The QM/MM-LR correlation equation for the single-species, single mode case with the ligand species i , the receptor species j , and the mode k specified, reads [86, 87]

$$\log K_{ijk} = \alpha - \Delta \langle E_{\text{QM/MM}} \rangle + \gamma - \Delta \langle \text{SASA} \rangle_{ijk} + \kappa \quad (5)$$

Here α , γ , and κ denote the optimized parameters. For the multispecies situation, eq 5 will apply to each microconstant K_{ijk} and the correlation equation for the observed association constant K is obtained from eq 2:

$$\log K = \log \sum_{i=1}^l \sum_{j=1}^r \sum_{k=1}^m f_i f_j f_k 10^{\alpha - \Delta \langle E_{\text{QM/MM}} \rangle_{ijk} + \gamma - \Delta \langle \text{SASA} \rangle_{ijk} + \kappa} \quad (6)$$

Multiple modes are usually covered by the time-averaged structures from MD simulations, which reach equilibrium characteristics. The $k = 1$ in eq 6 in these cases. Sometimes, MD simulations oscillate between two or more states and the correlations can be improved by considering multiple binding modes, as demonstrated for 28 inhibitors of matrix metalloproteinase 9 [84].

The multispecies QM/MM-LR approach was applied to 66 benzothiophene and pyrrolopyridine inhibitors of mitogen-activated protein kinase (MAPK)-activated protein kinase 2 (MK2) [85]. The compounds form a complex network of up to five tautomers and seven ionization species (233 species altogether), and the binding site is available in one ionization/tautomer species only. The procedure consisting of docking, QM/MM optimization of the best pose, 1-ns MD simulation, and calculation of the QM/MM energy for the time-averaged structures was performed for all 233 species, which took about 130,000 hours of the CPU time. The results were correlated with the inhibition potencies using eq 6. All steps were needed to achieve a meaningful correlation, as illustrated by the values of the square correlation coefficient (r^2) shown in the brackets: docking (0.002), QM/MM optimization of the best pose (0.202), MD simulation starting with the best pose and using the independently scaled time averages of the electrostatic and van der Waals energies and the SASA term in eq 6 (0.353), and the QM/MM energy of the time averaged structures (0.906). The use of all tautomers and ionization species was required to achieve the best correlation with $r^2 = 0.906$, because the single-tautomer, single species correlation resulted in $r^2 = 0.662$, and the use of all ionization species or all tautomers increased r^2 to 0.734 or 0.839, respectively. The results were extensively cross-validated using 70 different test sets of 9 or 10 randomly selected compounds, which were omitted from calibration and used for prediction. The deterioration of the statistical indices with the omission of ligands was minimal. There was an approximately linear trend between the fraction of species in water and bound prevalence. However, the correlation was rather loose and individual differences reached up to 40% either direction. The study illustrated how the predictions of observed binding affinities for multispecies situations can be made using the simulations and computations for individual pairs of ligands with receptors,

Conclusions

Speciation of drugs thanks to ionization, tautomerism, and covalent hydration is widely recognized as a factor complicating the 3D-QSAR studies at all levels. We have shown that the thermodynamic master equation 2 describes the relationship between the observed association constant, fractions of ligand and receptor species, and the microconstants characterizing individual binding complexes. The master equation provides a recipe for modifying the correlation equations of ligand- and receptor-based 3D-QSAR approaches to account for speciation of both ligands and receptors. For the application of the multispecies, multimode 3D-QSAR equation, the procedures that are typically performed for binding of each ligand to the receptor must be performed for each ligand-receptor species pair.

Optimization provides the estimates of binding energies and, in this way, also the binding prevalence of each ligand-species pair. Interestingly, the multispecies, multimode correlation does not require an increased number of optimized coefficients, and only needs a change in the form of the correlation equation. This approach also circumvents the combinatorial explosion that would be encountered with an exhaustive examination of multiple species and modes for each ligand by classical one-species, one-mode 3D-QSAR approaches. Although further studies are needed, the reviewed results indicate that the multispecies, multimode approach increases the realism of the description of drug-receptor interaction and has a potential to contribute to the drug development process.

Acknowledgments

This work was supported in part by the NIH NIGMS grant R01 GM80508 and the access to Teragrid computation resources for the projects MCB100078 and MCB110017.

References

1. Martin YC. Let's not forget tautomers. *J Comput Aid Mol Des.* 2009; 23:693–704.
2. Shimba N, Serber Z, Ledwidge R, Miller SM, Craik CS, Doetsch V. Quantitative identification of the protonation state of histidines in vitro and in vivo. *Biochemistry.* 2003; 42:9227–34. [PubMed: 12885258]
3. Vila JA, Arnautova YA, Vorobjev Y, Scheraga HA. Assessing the fractions of tautomeric forms of the imidazole ring of histidine in proteins as a function of pH. *P Natl Acad Sci USA.* 2011; 108:5602–7.
4. Clarke JA, Dawson PJ, Grigg R, Rochester CH. Spectroscopic study of the acid ionization of porphyrins. *J Chem Soc Perk T.* 1973; 2:414–6.
5. Braun J, Koecher M, Schlabach M, Wehrle B, Limbach HH, Vogel E. NMR study of the tautomerism of porphyrin including the kinetic HH/HD/DD isotope effects in the liquid and the solid state. *J Am Chem Soc.* 1994; 116:6593–604.
6. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 2008; 36:D901–D906. [PubMed: 18048412]
7. Sponer J, Leszczynski J, Hobza P. Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers.* 2001; 61:3–31. [PubMed: 11891626]
8. Sigel H. Acid-base properties of purine residues and the effect of metal ions: Quantification of rare nucleobase tautomers. *Pure Appl Chem.* 2004; 76:1869–86.
9. Lippert, B.; Gupta, D. Promotion of rare nucleobase tautomers by metal binding. *Dalton T*; 2009. p. 4619-34.
10. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem.* 1996; 39:2887–93. [PubMed: 8709122]
11. Lee PH, Ayyampalayam SN, Carreira LA, Shalaeva M, Bhattachar S, Coselmon R, Poole S, Gifford E, Lombardo F. In silico prediction of ionization constants of drugs. *Mol Pharmaceut.* 2007; 4:498–512.
12. Lesney MS. A basic history of acid - from Aristotle to Arnold. *Today's Chemist at Work.* 2003:47–8.
13. Browning CH, Gulbransen R, Kennaway EL. Hydrogen-ion concentration and antiseptic potency, with special reference to the action of acridine compounds. *J Pathol Bacteriol.* 1919; 23:106–8.
14. Shore PA, Brodie PP, Hogben CAM. The gastric secretion of drugs: A pH partition hypothesis. *J Pharmacol Exp Therap.* 1957; 119:361–9. [PubMed: 13417090]
15. Vermast PG. The theory of disinfection in the light of the Meyer-Overton lipid theory. *Biochem Z.* 1921; 125:106–48.
16. Hammett LP. Effect of structure upon the reactions of organic compounds. Benzene derivatives. *J Am Chem Soc.* 1937; 59:96–103.

17. Branch, GEK.; Calvin, M. The theory of organic chemistry: An advanced course. Prentice-Hall, Inc.; New York: 1941. p. 1-523.
18. Bell PH, Roblin RO Jr. Chemotherapy. VII. A theory of the relation of structure to activity of sulfanilamide-type compounds. *J Am Chem Soc.* 1942; 64:2905–17.
19. Cowles PB. Ionization and the bacteriostatic action of sulfonamides. *Yale J Biol Med.* 1942; 14:599–604. [PubMed: 21434043]
20. Eagle H. The spirocheticidal and trypanocidal action of acid-substituted phenylarsenoxides as a function of pH and dissociation constants. *J Pharmacol.* 1945; 85:265–82.
21. Cowles PB, Klotz IM. The effect of pH upon the bacteriostatic activity of certain nitrophenols. *J Bacteriol.* 1948; 56:277–82. [PubMed: 16561571]
22. Beevers H, Simon EW. Effect of pH on the activity of some respiratory inhibitors. *Nature.* 1949; 163:408–9. [PubMed: 18208222]
23. Simon EW. Effect of pH on the biological activity of weak acids and bases. *Nature.* 1950; 166:343–4. [PubMed: 14775656]
24. Martin YC, Hansch C. Influence of hydrophobic character on the relative rate of oxidation of drugs by rat liver microsomes. *J Med Chem.* 1971; 14:777–9. [PubMed: 5140003]
25. Martin YC, Hackbarth JJ. Theoretical model-based equations for the linear free energy relationships of the biological activity of ionizable substances. 1. Equilibrium-controlled potency. *J Med Chem.* 1976; 19:1033–9. [PubMed: 966250]
26. Martin, YC. The quantitative relationships between pKa, ionization, and drug potency: Utility of model-based equations. In: Yalkowsky, SH.; Sinkula, AA.; Valvani, SC., editors. *Physical Chemical Properties of Drugs.* Marcel Dekker Inc; New York: 1980. p. 49-110.
27. Lippmann. History of tautomerism. *Chem-Ztg.* 1910; 34:49.
28. Rabe P. Beiträge zur Aufklärung der Tautomerieerscheinungen. *Justus Liebigs Ann Chem.* 1900; 313:129–207.
29. Donald C, Cotzias GC. Interaction of trace metals with phenothiazine drug derivatives. II. Formation of free radicals. *P Natl Acad Sci USA.* 1962; 48:623–42.
30. Katritzky AR. Prototropic tautomerism of heteroaromatic compounds. *Chimia.* 1970; 24:134–46.
31. Grebow PE, Hooker TM Jr. Conformation of histidine model peptides. III. Chiroptical properties of cyclo-L-anyl-L-histidine and cyclo-L-histidinyll-L-histidine. *Biopolymers.* 1975; 14:1863–83.
32. Tanokura M. Proton NMR study on the tautomerism of the imidazole ring of histidine residues. II. Microenvironments of histidine-12 and histidine-119 of bovine pancreatic ribonuclease A. *BBA-Protein Struct M.* 1983; 742:586–96.
33. Franks F. Conformational equilibria and dynamics of small carbohydrates: Anomers, tautomers, and rotamers. *Spec Publ - R Soc Chem.* 1989; 74:213–25.
34. Farr-Jones S, Wong WYL, Gutheil WG, Bachovchin WW. Direct observation of the tautomeric forms of histidine in nitrogen-15 NMR spectra at low temperatures. Comments on intramolecular hydrogen bonding and on tautomeric equilibrium constants. *J Am Chem Soc.* 1993; 115:6813–9.
35. Abraham MH, Leo AJ. Partition between phases of a solute that exists as two interconverting species. *J Chem Soc Perk T 2: Phys Org Chem.* 1995:1839–42.
36. Leo AJ. Effect of tautomeric equilibria on hydrophobicity as measured by partition coefficients. *ACS Sym Ser.* 1995; 589:292–302.
37. Nagy PI. Theoretical calculations for the conformational/tautomeric equilibria of biologically important molecules in solution. *Recent Res Dev Phys Chem.* 1999; 3:1–21.
38. Remko M, Van Duijnen PT, Swart M. Theoretical study of molecular structure, tautomerism, and geometrical isomerism of N-methyl- and N-phenyl-substituted cyclic imidazolines, oxazolines, and thiazolines. *Struct Chem.* 2003; 14:271–8.
39. Kabelac M, Hobza P. Na⁺, Mg²⁺, and Zn²⁺ Binding to all tautomers of adenine, cytosine, and thymine and the eight most stable keto/enol tautomers of guanine: A correlated ab initio quantum chemical study. *J Phys Chem B.* 2006; 110:14515–23. [PubMed: 16854164]
40. Eberz WF, Welge HJ, Yost DM, Lucas HJ. Hydration of unsaturated compounds. IV. The rate of hydration of isobutene in the presence of silver ion. The nature of the isobutene-silver complex. *J Am Chem Soc.* 1937; 59:45–9.

41. Albert A, Armarego WL. Covalent hydration in nitrogen-containing heteroaromatic compounds. I. Qualitative aspects. *Adv Heterocycl Chem.* 1965; 4:1–42. [PubMed: 5331920]
42. Perrin DD. Covalent hydration in nitrogen heteroaromatic compounds. II. Quantitative aspects. *Adv Heterocycl Chem.* 1965; 4:43–73. [PubMed: 5331705]
43. Terrier F, Lakhdar S, Boubaker T, Vichard D, Goumont R, Buncel E. Superelectrophilicity in π -complexation processes. *Chem Sustain Develop.* 2008; 16:57–68.
44. Erion MD, Reddy MR. Calculation of relative hydration free energy differences for heteroaromatic compounds: Use in the design of adenosine deaminase and cytidine deaminase inhibitors. *J Am Chem Soc.* 1998; 120:3295–304.
45. Pospisil P, Ballmer P, Scapozza L, Folkers G. Tautomerism in computer-aided drug design. *J Recept Sig Transd.* 2003; 23:361–71.
46. Oprea TI, Matter H. Integrating virtual screening in lead discovery. *Curr Opin Chem Biol.* 2004; 8:349–58. [PubMed: 15288243]
47. Zou JW, Luo CC, Zhang HX, Liu HC, Jiang YJ, Yu QS. Three-dimensional QSAR of HPPD inhibitors, PSA inhibitors, and anxiolytic agents: Effect of tautomerism on the CoMFA models. *J Mol Graph Model.* 2007; 26:494–504. [PubMed: 17418602]
48. Vedani A, Briem K, Dobler M, Dollinger H, McMasters DR. Multiple-conformation and protonation-state representation in 4D-QSAR: The neurokinin-1 receptor system. *J Med Chem.* 2000; 43:4416–27. [PubMed: 11087566]
49. Natesan S, Wang T, Lukacova V, Bartus V, Khandelwal A, Balaz S. Rigorous treatment of multispecies multimode ligand-receptor interactions in 3D-QSAR: CoMFA analysis of thyroxine analogs binding to transthyretin. *J Chem Inf Model.* 2011; 51:1132–50. [PubMed: 21476521]
50. Lill MA, Vedani A, Dobler M. Raptor: Combining dual-shell representation, induced-fit simulation, and hydrophobicity scoring in receptor modeling: Application toward the simulation of structurally diverse ligand sets. *J Med Chem.* 2004; 47:6174–86. [PubMed: 15566288]
51. Katritzky A, Hall C, El Gendy B, Draghici B. Tautomerism in drug discovery. *J Comp Aid Mol Des.* 2010; 24:475–84.
52. Albert A. Ionization, pH and biological activity. *Pharmacol Rev.* 1952; 4:136–67. [PubMed: 14941673]
53. Chou PT, Chen YC, Yu WS, Chou YH, Wei CY, Cheng YM. Excited-state intramolecular proton transfer in 10-hydroxybenzo[h]quinoline. *J Phys Chem A.* 2001; 105:1731–40.
54. Tee OS, Pika J, Kornblatt MJ, Trani M. Rates and equilibrium constants for the covalent hydration of 5-bromo-2(1H)-pyrimidinone in aqueous solution. *Can J Chem.* 1986; 64:1267–72.
55. Zeigler, BP. *Theory of Modeling and Simulation.* John Wiley & Sons; New York: 1976. p. 1-435.
56. Hobza, P.; Zahradnik, R. *Weak Intermolecular Interactions in Chemistry and Biology.* Academia; Prague: 1980. p. 1-29.
57. Izquierdo-Martin M, Chapman KT, Hagmann WK, Stein RL. Studies on the kinetic and chemical mechanism of inhibition of stromelysin by an N-(carboxyalkyl)dipeptide. *Biochemistry.* 1994; 33:1356–65. [PubMed: 8312254]
58. Bartlett PA, Marlowe CK. Possible role for water dissociation in the slow binding of phosphorus-containing transition-state-analogue inhibitors of thermolysin. *Biochemistry.* 1987; 26:8553–61. [PubMed: 3442676]
59. Bernardo MM, Brown S, Li ZH, Fridman R, Mobashery S. Design, synthesis, and characterization of potent, slow-binding inhibitors that are selective for gelatinases. *J Biol Chem.* 2002; 277:11201–7. [PubMed: 11790786]
60. Rosenblum G, Meroueh SO, Kleifeld O, Brown S, Singson SP, Fridman R, Mobashery S, Sagi I. Structural basis for potent slow binding inhibition of human matrix metalloproteinase-2 (MMP-2). *J Biol Chem.* 2003; 278:27009–15. [PubMed: 12679334]
61. Brown S, Bernardo MM, Li ZH, Kotra LP, Tanaka Y, Fridman R, Mobashery S. Potent and selective mechanism-based inhibition of gelatinases. *J Am Chem Soc.* 2000; 122:6799–800.
62. Wang X, Minasov G, Shoichet BK. Noncovalent interaction energies in covalent complexes: TEM-1 β -lactamase and β -lactams. *Proteins.* 2002; 47:86–96. [PubMed: 11870868]

63. Dreyer GB, Lambert DM, Meek TD, Carr TJ, Tomaszek TAJ, Fernandez AV, Bartus H, Cacciavillani E, Hassell AM, Minnich M, Petteway SRJ, Metcalf BW, Lewis M. Hydroxyethylene isostere inhibitors of human immunodeficiency virus-1 protease: Structure-activity analysis using enzyme kinetics, X-ray crystallography, and infected T-cell assays. *Biochemistry*. 1992; 31:6646–59. [PubMed: 1637805]
64. Duggleby RG, Attwood PV, Wallace JC, Keech BD. Avidin is a slow-binding inhibitor of pyruvate carboxylase. *Biochemistry*. 1982; 21:3364–70. [PubMed: 7115676]
65. Copeland RA, Pompliano DL, Meek TD. Drug-target residence time and its implications for lead optimization. *Nat Rev Drug Discov*. 2006; 5:730–9. [PubMed: 16888652]
66. Hilal SH, Karickhoff SW, Carreira LA. Estimation of microscopic, zwitterionic ionization constants, isoelectric point and molecular speciation of organic compounds. *Talanta*. 1999; 50:827–40. [PubMed: 18967773]
67. Shelley J, Cholleti A, Frye L, Greenwood J, Timlin M, Uchimaya M. Epik: A software program for pKa prediction and protonation state generation for drug-like molecules. *J Comp Aid Mol Des*. 2007; 21:681–91.
68. PALLAS/pKalc. Compudrug International, Inc; Sedona, AZ: 2008.
69. Milletti F, Storchi L, Sforza G, Cross S, Cruciani G. Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *J Chem Inf Model*. 2009; 49:68–75. [PubMed: 19123923]
70. ten Brink T, Exner TE. pKa based protonation states and microspecies for protein-ligand docking. *J Comp Aid Mol Des*. 2010; 24:935–42.
71. Meloun M, Bordovska S. Benchmarking and validating algorithms that estimate pKa values of drugs based on their molecular structures. *Anal Bioanal Chem*. 2007; 389:1267–81. [PubMed: 17676314]
72. Dearden JC, Cronin MTD, Lappin DC. A comparison of commercially available software for the prediction of pKa. *J Pharm Pharmacol*. 2007; 59:16.
73. Lee AC, Crippen GM. Predicting pKa. *J Chem Inf Model*. 2009; 49:2013–33. [PubMed: 19702243]
74. Hilal SH, Bornander LL, Carreira LA. Hydration equilibrium constants of aldehydes, ketones and quinazolines. *QSAR Comb Sci*. 2005; 24:631–7.
75. Antosiewicz J, Briggs JW, Elcock AH, Gilson MK, McCammon JA. Computing ionization states of proteins with a detailed charge model. *J Comput Chem*. 1996; 17:1633–44.
76. Alexov E. Role of the protein side-chain fluctuations on the strength of pair-wise electrostatic interactions: comparing experimental with computed pK_as. *Proteins: Struct Funct Genetics*. 2002; 50:94–103.
77. Davies M, Toseland C, Moss D, Flower D. Benchmarking pKa prediction. *BMC Biochem*. 2006; 7:18. [PubMed: 16749919]
78. Natesan S, Wang T, Lukacova V, Bartus V, Khandelwal A, Subramaniam R, Balaz S. Cellular quantitative structure-activity relationship (cell-QSAR): Conceptual dissection of receptor binding and intracellular disposition in antifilarial activities of Selwood antimycins. *J Med Chem*. 2012; 55:3699–712. [PubMed: 22468611]
79. Cramer RD III, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc*. 1988; 110:5959–67. [PubMed: 22148765]
80. Lukacova V, Balaz S. Multimode ligand binding in receptor site modeling: Implementation in CoMFA. *J Chem Inf Comp Sci*. 2003; 43:2093–105.
81. Khandelwal A, Lukacova V, Comez D, Kroll DM, Raha S, Balaz S. A combination of docking, QM/MM methods, and MD simulation for binding affinity estimation of metalloprotein ligands. *J Med Chem*. 2005; 48:5437–47. [PubMed: 16107143]
82. Khandelwal A, Balaz S. Improved estimation of ligand/macromolecule binding affinities by linear response approach using a combination of multi-mode MD simulation and QM/MM methods. *J Comp Aid Mol Des*. 2007; 21:131–7.
83. Khandelwal A, Balaz S. QM/MM linear response method distinguishes ligand affinities for closely related metalloproteins. *Proteins*. 2007; 69:326–39. [PubMed: 17607744]

84. Khandelwal A, Lukacova V, Kroll DM, Raha S, Comez D, Balaz S. Processing multi-mode binding situations in simulation-based prediction of ligand-macromolecule affinities. *J Phys Chem A*. 2005; 109:6387–91. [PubMed: 16833982]
85. Natesan S, Subramaniam R, Bergeron C, Balaz S. Binding affinity prediction for ligands and receptors forming tautomers and ionization species: Inhibition of mitogen-activated protein kinase-activated protein kinase 2 (MK2). *J Med Chem*. 2012; 55:2035–47. [PubMed: 22280316]
86. Aqvist J, Medina C, Samuelsson JE. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng*. 1994; 7:385–91. [PubMed: 8177887]
87. Carlson HA, Jorgensen WL. An extended linear response method for determining free energies of hydration. *J Phys Chem*. 1995; 99:10667–73.

Figure 1.

Multispecies, multimode binding equilibria, illustrated only for one ligand species (L_i). Line-head arrows indicate speciation processes (ionization, tautomerism or covalent hydration) and full-head arrows indicate binding processes. Each binding process is characterized by the microconstant K_{ijk} (not shown), with the same subscript as that of the resulting complex LR_{ijk} .